Π Α Ν Ε Π Ι Σ Τ Η Μ Ι Ο Π Ε Ι Ρ Α Ι Ω Σ

Σχολή Χρηματοοικονομικής και Στατιστικής

Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ

ΣΤΗΝ **ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ**

ΠΡΟΒΛΕΨΗ ΧΡΗΜΑΤΙΣΤΗΡΙΑΚΩΝ ΔΕΙΚΤΩΝ ΜΕ ΤΕΧΝΙΚΕΣ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ ΣΕ ΥΒΡΙΔΙΚΕΣ ΠΗΓΕΣ ΔΕΔΟΜΕΝΩΝ.

Χρυσαυγή Χ. Μήτση

Διπλωματική Εργασία

που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην Εφαρμοσμένη Στατιστική

Πειραιάς

Ιούνιος 2016

Π Α Ν Ε Π Ι Σ Τ Η Μ Ι Ο Π Ε Ι Ρ Α Ι Ω Σ

Σχολή Χρηματοοικονομικής και Στατιστικής

Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ

ΣΤΗΝ **ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ**

ΠΡΟΒΛΕΨΗ ΧΡΗΜΑΤΙΣΤΗΡΙΑΚΩΝ ΔΕΙΚΤΩΝ ΜΕ ΤΕΧΝΙΚΕΣ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ ΣΕ ΥΒΡΙΔΙΚΕΣ ΠΗΓΕΣ ΔΕΔΟΜΕΝΩΝ.

Χρυσαυγή Χ. Μήτση

Διπλωματική Εργασία

που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην Εφαρμοσμένη Στατιστική

Πειραιάς

Ιούνιος 2016

Η παρούσα Διπλωματική Εργασία εγκρίθηκε ομόφωνα από την Τριμελή Εξεταστική Επιτροπή που ορίσθηκε από τη ΓΣΕΣ του Τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς στην υπ' αριθμ. ……. συνεδρίασή του σύμφωνα με τον Εσωτερικό Κανονισμό Λειτουργίας του Προγράμματος Μεταπτυχιακών Σπουδών στην Εφαρμοσμένη Στατιστική

Τα μέλη της Επιτροπής ήταν:

- Επίκουρος Καθηγητής Ν. Πελέκης (Επιβλέπων)
- Επίκουρος Καθηγητής Ε. Κοφίδης
- Καθηγητής Μ.Κούτρας

Η έγκριση της Διπλωματική Εργασίας από το Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς δεν υποδηλώνει αποδοχή των γνωμών του συγγραφέα.

UNIVERSITY OF PIRAEUS

School of Finance and Statistics

Department of Statistics and Insurance Science

POSTGRADUATE PROGRAM IN

APPLIED STATISTICS

PREDICTION OF STOCK MARKET INDICES WITH DATA MINING TECHNIQUES IN HYBRID DATA SOURCES

By

Chrysavgi C. Mitsi

Msc Dissertation

submitted to the Department of Statistics and Insurance Science of the University of Piraeus in partial fulfillment of the requirements for the degree of Master of Science in Applied Statistics

Piraeus, Greece
June 2016

Στους γονείς μου

Χρήστο και Πηνελόπη

ΕΥΧΑΡΙΣΤΙΕΣ

Θα ήθελα να ευχαριστήσω θερμά τον επιβλέποντα καθηγητή μου κ. Νικόλαο Πελέκη, του οποίου η καθοδήγηση και η στήριξη υπήρξαν πολύτιμες καθ' όλη τη διάρκεια της έρευνας για την ολοκλήρωση της παρούσας διπλωματικής εργασίας. Η πολύτιμη καθοδήγησή του, ο χρόνος που αφιέρωσε και η άριστη συνεργασία που είχαμε συνέβαλαν σημαντικά στην εκπόνηση της διπλωματικής.

Παράλληλα, θα ήθελα να εκφράσω τις ευχαριστίες μου στους κκ. Ελευθέριο Κοφίδη και κκ. Μάρκο Κούτρα για τη συμμετοχή τους στη συμβουλευτική επιτροπή. Τέλος, δε θα μπορούσα να παραλείψω την αναφορά στους γονείς και τον αδερφό μου, τους οποίους ευχαριστώ για την αμέριστη συμπαράσταση και την εμπιστοσύνη που μου δείχνουν σε κάθε βήμα της ζωής μου.

Χρυσαυγή Μήτση
Πειραιάς
Ιούνιος 2016

ΠΕΡΙΛΗΨΗ

Στις μέρες μας η πρόβλεψη χρονοσειρών και μάλιστα αποτελούμενων από χρηματοοικονομικά δεδομένα αποτελεί αναμφισβήτητα αντικείμενο εκτεταμένης ερευνητικής δραστηριότητας. Πράγματι, στη χρηματοοικονομική επιστήμη η ανάλυση χρονοσειρών εφαρμόζεται ευρέως για την πρόβλεψη των τιμών των διεθνών και εθνικών χρηματαγορών αλλά και σε εφαρμογές που σχετίζονται με τη διαδικασία πρόβλεψης είτε χρηματοοικονομικών κρίσεων είτε επενδυτικών στρατηγικών.

Είναι γεγονός, ακόμη, ότι τα σύγχρονα συστήματα υποστήριξης λήψης αποφάσεων βασίζονται σε πληροφόρηση που προέρχεται κυρίως από δομημένα δεδομένα αγνοώντας τα αδόμητα, τα οποία δύναται να προσφέρουν σημαντική πληροφόρηση. Αυτό, σε συνδυασμό με τη ραγδαία ανάπτυξη της τεχνολογίας συντέλεσε στη δημιουργία νέων δυναμικών εργαλείων μετατροπής των αδόμητων δεδομένων σε δομημένη πληροφορία, η οποία σε συνδυασμό με τη πληροφόρηση από τα δομημένα θα ωφελήσει τους επενδυτές, προκειμένου για λήψη βέλτιστης απόφασης. Την ανάγκη αυτή καλείται να καλύψει ένας νέος κλάδος της επιστήμης, η εξόρυξη δεδομένων, που αποτελεί ένας συνδυασμό ετερόκλητων επιστημονικών πεδίων όπως της στατιστικής, της μηχανικής μάθησης, της θεωρίας της πληροφορίας και των υπολογιστικών διαδικασιών.

Στη παρούσα εργασία παρουσιάζουμε ένα καινοτόμο σύστημα επεξεργασίας υβριδικών δεδομένων προκειμένου για πρόβλεψη της τάσης τραπεζικών μετοχών του Χρηματιστηρίου Αθηνών για το έτος 2014. Για το λόγο αυτό εξετάζονται τρεις εφαρμογές του συστήματος σε διαφορετικά σύνολα δεδομένων (αριθμητικά, κειμενικά και υβριδικά) και αξιολογούνται προκειμένου να εντοπιστεί η αποτελεσματικότερη εφαρμογή. Στο σύστημα εφαρμόζονται τεχνικές κατηγοριοποίησης και συσταδοποίησης σε σύνολα κατηγορικών δεδομένων ενώ συγχρόνως παρουσιάζονται οι βασικότερες έννοιες και μέθοδοι που χρησιμοποιούνται κατά τη διάρκεια της κατηγοριοποίησης, ομαδοποίησης και πρόβλεψης.

Σκοπός της εργασίας είναι εξηγώντας τις μεθόδους της εξόρυξης δεδομένων, να αναδειχθεί η χρησιμότητά της στα χρηματοοικονομικά δεδομένα καθώς και η σημασία της για την εξαγωγή σημαντικών συμπερασμάτων από αδόμητα και δύσκολα στη χρήση τους δεδομένα.

ABSTRACT

Nowadays, time series prediction, especially in the case of financial time series, is undoubtedly a matter of widespread research activity. Indeed, in finance, time series analysis is applied widely not only for the purposes of predicting prices of international and national markets but also it is used for the prediction of financial crises or / and investment strategies.

It is also a fact that modern decision support systems are based on information extracted from structured data. So, by neglecting the unstructured data may provoke the loss of significant information. This, in combination with the rapid development of technology has led to the need for new dynamic tools that will help transform unstructured data into structured information, which when combined with information extracted from structured data will help investments to make the best decision. A new field of science, called data mining, is going to supply that need. It is about a combination of different sciences such us statistics, machine learning, information theory and computational procedures.

In this thesis we introduce an innovative data processing system in order to predict the trend of banking shares of Athens Stock Exchange for the year 2014. Therefore, are examined three applications of the system in different data sets (arithmetic, text and hybrid) and then are evaluated so as to identify the more effective one. In this system are applied classification and clustering techniques in categorical data while presented the most basic concepts and methods used by classification, clustering and prediction techniques in data mining.

The aim of this thesis is by explaining data mining methods, to demonstrate not only its usefulness in finance but also its significance for drawing meaningful conclusions from unstructured and difficult in use data.

ΠΕΡΙΕΧΟΜΕΝΑ

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

ΚΑΤΑΛΟΓΟΣ ΣΧΗΜΑΤΩΝ

ΚΕΦΑΛΑΙΟ 1
ΕΙΣΑΓΩΓΗ

Ένα ερώτημα που συχνά απασχολεί τους αναλυτές χρηματιστηριακών δεδομένων είναι το κατά πόσο είναι δυνατή η πρόβλεψη χρηματιστηριακών δεικτών επιχειρήσεων που είναι εισηγμένες στις χρηματιστηριακές αγορές προκειμένου να εξαχθούν χρήσιμες πληροφορίες που θα επιδείξουν τη κίνηση της αγοράς και οπότε θα συμβάλλουν στη λήψη σωστής απόφασης από τους ενδιαφερόμενους. Αναμφισβήτητα, ένας μεγάλος αριθμός ερευνητών εστιάζει στην εύρεση ενός αποτελεσματικού μέσου ανάλυσης χρηματιστηριακών δεδομένων προκειμένου να αποφανθεί για τον τρόπο κίνησης της αγοράς και να βοηθήσουν στη λήψη αποφάσεων των όλο και αυξανόμενων επενδυτών. Η εξαγωγή χρήσιμων πληροφοριών από τον αυξανόμενο όγκο διαθέσιμων πληροφοριών που είναι διαθέσιμα δυσχεραίνει τη δράση των ερευνητών, ενώ καθιστά αναγκαία την εύρεση μιας καλύτερης προσέγγισης.

Πράγματι, ο ολοένα αυξανόμενος όγκος χρηματιστηριακών δεδομένων αποθηκεύεται καθημερινά σε βάσεις δεδομένων με αποτέλεσμα αυτές να γιγαντώνονται από στοιχεία που στη πλειοψηφία τους δεν οδηγούν σε κάποια απόφαση. Το μεγαλύτερο ποσοστό των διαθέσιμων δεδομένων είναι σε μορφή ελεύθερου κειμένου, γεγονός που δυσκολεύει την ανάλυσή τους. Η αδυναμία ανάλυσης μεγάλου ποσοστού των ανωτέρω δεδομένων οδήγησε στην ανάπτυξη της επιστήμης εξόρυξης κειμένου, η οποία ορίζεται ως η ανάλυση την ημι-δομημένων ή αδόμητων δεδομένων κειμένων προκειμένου οι πληροφορίες αυτών να μετατραπούν σε αριθμούς και οπότε να εφαρμοστούν οι γνωστοί αλγόριθμοι εξόρυξης δεδομένων και να ληφθούν χρήσιμα συμπεράσματα. Η εξόρυξη δεδομένων, γενικότερα, είναι η επιστήμη που σχεδιάστηκε για να βοηθήσει τους επενδυτές να ανακαλύψουν πρότυπα που εμφανίζονται σε ιστορικά δεδομένα, ούτως ώστε να βοηθηθούν ως προς τις μελλοντικές αποφάσεις. Η επιστήμη αυτή, ακόμη, δύναται να ανακαλύψει πρότυπα και σχέσεις μεταξύ των δεδομένων, να ομαδοποιήσει και να ταξινομήσει δεδομένα, να μοντελοποιήσει και, τέλος, να προβλέψει τα αποτελέσματα τους.

Η ραγδαία ανάπτυξη της τεχνολογίας επέφερε τη δημιουργία κατάλληλων λογισμικών και πακέτων τα οποία εισήχθηκαν στα χρηματιστηριακά κέντρα (ή και σε επιχειρήσεις) και τα οποία ισχυρίζονται ότι είναι σε θέση να προβλέψουν τη μελλοντική τιμή των χρηματιστηριακών δεικτών έτσι ώστε να μπορέσουν οι χρήστες να αποκομίσουν υπερβάλλοντα κέρδη. Στο φαινόμενο αυτό βασίζεται η παρούσα εργασία.

1.1. ΣΚΟΠΟΣ

Σε αυτή τη μεταπτυχιακή διατριβή θα υλοποιηθεί ένα σύστημα πρόβλεψης που βασίστηκε στο σύστημα NewsCATS (Mittermayer, 2004) και το οποίο θα δίνει τη δυνατότητα στους επενδυτές τραπεζικών μετοχών του Χρηματιστηρίου Αθηνών να λάβουν τη βέλτιστη απόφαση όσον αφορά τις επόμενες κινήσεις τους. Το αντικείμενο της παρούσας διπλωματικής, ουσιαστικά, είναι η ανάλυση των χρηματιστηριακών δεδομένων με τη μελέτη και την εφαρμογή τεχνικών εξόρυξης γνώσης από υβριδικές πηγές δεδομένων (δηλαδή δεδομένα κειμένου αλλά και αριθμητικά) βάσει των οποίων δύναται η πρόβλεψη χρηματιστηριακών δεικτών.

1.2. ΠΕΡΙΓΡΑΦΗ

Το θεωρητικό υπόβαθρο της εργασίας αναφέρεται στο γνωστικό αντικείμενο των Χρηματοοικονομικών και ειδικότερα αναφέρεται στο τρόπο διενέργειας πρόβλεψης των χρηματιστηριακών δεικτών. Αυτό επιτυγχάνεται ορίζοντας αρχικά την έννοια της

χρονοσειράς καθώς και τη σημασία της ανάλυσής της προκειμένου να προβλέψουμε τη μελλοντική τιμή του υπό μελέτη φαινομένου. Επιπλέον, ορίζουμε και την έννοια της εξόρυξης δεδομένων, περιλαμβάνοντας, ακόμη, εννοιολογικά θεμέλια καθώς και τη διαδικασία της νέας αυτής επιστήμης. Στη συνέχεια μελετώνται αντίστοιχες έρευνες και υλοποιήσεις εστιάζοντας εκτενώς στο τρόπο λειτουργίας του συστήματος NewsCATS (*News Categorization and Trading System*), που μέσω εξόρυξης κειμένου είναι σε θέση να προβλέψει τις τάσεις των τιμών των χρηματιστηριακών δεικτών. Η πρόβλεψη διενεργείται κατά το χρονικό διάστημα έως και 60 λεπτών μετά τη δημοσίευση ενός δελτίου τύπου. το οποίο δίνει αποτέλεσμα εφαρμόζοντας τρία στάδια: Το πρώτο στάδιο αφορά την ανάκτηση σχετικών πληροφοριών (δηλαδή πληροφοριών που αφορούν τους εκάστοτε χρηματιστηριακούς δείκτες που μας ενδιαφέρουν) από τα δελτία τύπου εφαρμόζοντας τεχνικές προ-επεξεργασίας κειμένου. Στο δεύτερο στάδιο, το σύστημα ταξινομεί (σορτάρει) τα δελτία τύπου σε προκαθορισμένες κατηγορίες Το τρίτο στάδιο, το οποίο είναι και το σημαντικότερο, είναι το στάδιο από το οποίο εξάγονται οι κατάλληλες στρατηγικές χρηματιστηριακών συναλλαγών και βασίζεται στη κατηγοριοποίηση των δελτίων τύπου που πραγματοποιήθηκε στο δεύτερο στάδιο.

Το εμπειρικό πλαίσιο της παρούσας εργασίας αφορά το σχεδιασμό και την υλοποίηση ενός διευρυμένου συστήματος πρόβλεψης, που αναφέρεται στη κατηγοριοποίηση και στην ομαδοποίηση υβριδικών χρηματιστηριακών δεδομένων του Χρηματιστηρίου Αθηνών. Συγκεκριμένα, διακρίνουμε τρεις κατηγορίες ανάλυσης προκειμένου να αποφανθούμε για το ποια δεδομένα οδηγούν στη βέλτιστη λύση. Η πρώτη κατηγορία εμπεριέχει αριθμητικά δεδομένα των τραπεζικών μετοχών, δηλαδή τις τιμές κλεισίματος και τις τιμές του όγκου για κάθε ημέρα λειτουργίας του Χρηματιστηρίου για το έτος 2014. Η δεύτερη κατηγορία περιλαμβάνει κειμενικά δεδομένα, δηλαδή δελτία τύπου/ άρθρα που εκδόθηκαν εντός των ωρών λειτουργίας του Χρηματιστηρίου για τις μετοχές που πραγματευόμαστε. Η τρίτη κατηγορία συμπεριλαμβάνει τα κειμενικά και τα αριθμητικά δεδομένα (υβριδικά) για τις επικείμενες μετοχές. Θα χρησιμοποιηθεί το στατιστικό πακέτο R και θα δοθεί βαρύτητα στην ελληνική γλώσσα, ώστε ο εκάστοτε αλγόριθμος που θα εξάγει τα αποτελέσματα να έχει τη μέγιστη δυνατή απόδοση. Το σύστημα που θα δημιουργήσουμε θα αξιολογηθεί για κάθε συνδυασμό δεδομένων που θα χρησιμοποιήσουμε προκειμένου για υπολογισμό της ακρίβειας των αποτελεσμάτων που θα παρέχει στον χρήστη. Αναφέρουμε ότι θα χρησιμοποιηθούν οι πιο γνωστοί αλγόριθμοι κατηγοριοποίησης, δηλαδή QDA, KNN και CART και οι πιο γνωστοί της συσταδοποίησης, δηλαδή ιεραρχική συσταδοποίηση και k-means. Τέλος, θα παρουσιαστούν προτάσεις βελτίωσης του νεοδημιουργηθέντος συστήματος.

1.3. ΟΡΓΑΝΩΣΗ ΤΗΣ ΕΡΓΑΣΙΑΣ

Η παρούσα εργασία διαρθρώνεται ως εξής: Στο δεύτερο κεφάλαιο γίνεται λόγος για τις χρονοσειρές, δηλαδή για ακολουθίες σημείων σχετιζόμενων με τη περιγραφή ενός φαινομένου, ανά τακτά χρονικά διαστήματα και για την αναγκαιότητα πρόβλεψης αυτών προκειμένου για ασφαλέστερες επενδυτικές αποφάσεις. Επίσης, αναφέρεται η έννοια εξόρυξη κειμένου, δηλαδή μιας διαδικασίας ανεύρεσης χρήσιμης πληροφορίας από όγκο αδόμητων δεδομένων. Εν συνεχεία, παρουσιάζονται οι τεχνικές εξόρυξης κειμένου και ο τρόπος αναπαράστασής τους σε αριθμούς προκειμένου να εφαρμοστούν οι κατάλληλοι αλγόριθμοι.

Στο τρίτο κεφάλαιο γίνεται η βιβλιογραφική ανασκόπηση. Ουσιαστικά, παρουσιάζεται εκτενώς η εργασία του Mittermayer και των συναδέλφων του, οι οποίοι υλοποίησαν και εφάρμοσαν σε πραγματικά δεδομένα, κειμενικά και αριθμητικά, το σύστημα NewsCATS προκειμένου να προβλέψουν τις τάσεις των μετοχών στο Χρηματιστήριο της Αμερικής (*National Market System, NMS*) από το έτος 2002 έως το 2004. Ειδικότερα, περιγράφεται ο

τρόπος σχεδιασμού του συστήματος, οι τεχνολογίες που εφαρμόστηκαν κατά την υλοποίηση αλλά, ο τρόπος λειτουργίας του και τα συμπεράσματα των αναλυτών από την εφαρμογή του.

Στο τέταρτο κεφάλαιο, ορίζονται η κατηγοριοποίηση και η συσταδοποίηση, δηλαδή οι τεχνικές που θα εφαρμοστούν προκειμένου να εφαρμόσουμε το δικό μας σύστημα πρόβλεψης. Χρησιμοποιώντας το λογισμικό της R, παραθέτουμε τα αποτελέσματα των αλγορίθμων κατηγοριοποίησης και συσταδοποίησης για τις τρεις ομάδες δεδομένων που δημιουργήσαμε (αριθμητικά, κειμενικά, υβριδικά) και παρουσιάζουμε μια διαδικασία αξιολόγησης. Τελικό βήμα, είναι η επισκόπηση των συμπερασμάτων και των στόχων που επετεύχθησαν από την εργασία αυτή αλλά και η πρόταση ιδεών για περεταίρω βελτίωση του συστήματος.

1.4. ΣΥΜΠΕΡΑΣΜΑΤΑ

Στην παρούσα εργασία προσπαθήσαμε να δείξουμε τη δυναμική των μεθόδων που σχετίζονται με την εξόρυξη γνώσης, δηλαδή εστιάσαμε στην εύρεση κρυμμένων τυποποιημένων μορφών (*patterns*) από μεγάλες βάσεις δεδομένων έτσι, ώστε να είναι δυνατή η πρόγνωση μελλοντικών συμπεριφορών. Ειδικότερα, χρησιμοποιήσαμε τεχνικές εξόρυξης γνώσης σε χρηματοοικονομικά (αριθμητικά και κειμενικά) δεδομένα για την λήψη της αποδοτικότερης, όσο το δυνατόν, απόφασης όσον αφορά τη κίνηση των τραπεζικών μετοχών του Χρηματιστηρίου Αθηνών. Η επιλογή της καταλληλότερης μεθόδου για την επίτευξη του σκοπού μας έγινε αφού αξιολογήσαμε μερικούς από τους γνωστότερους αλγορίθμους εξόρυξης γνώσης των σημαντικότερων τεχνικών κατηγοριοποίησης και συσταδοποίησης. Από την ανάλυση των δεδομένων και τη σύγκριση των αποτελεσμάτων που πραγματοποιήθηκε, καταλήξαμε στο ότι η προτεινόμενη μέθοδος κατηγοριοποίησης KNN είναι η πιο ακριβής.

Χρησιμοποιώντας τη μέθοδο KNN για το προγνωστικό μοντέλο (είτε περιλαμβάνει κειμενικά δεδομένα είτε υβριδικά), υλοποιήσαμε ένα πρωτότυπο λογισμικό εργαλείο που προβλέπει την επόμενη κίνηση της εκάστοτε μετοχής (‘Good’, ‘Bad’) και επομένως υποδεικνύει στους ενδιαφερόμενους τη καταλληλότερη απόφαση. Σημειώνουμε, δε, ότι τα δεδομένα που χρησιμοποιήθηκαν ήταν ελεύθερα διαθέσιμα από το διαδίκτυο. Μετά από το σύνολο πειραμάτων που προηγήθηκαν, καταλήγουμε στο συμπέρασμα ότι οι μέθοδοι της Μηχανικής Μάθησης με την συμβολή της R είναι σε θέση να αναλύσουν με αρκετά μεγάλη ακρίβεια τα χρηματοοικονομικά δεδομένα είτε είναι αριθμητικά είτε πρόκειται για δελτία τύπου.

Στο δικό μας πρωτότυπο σύστημα, ο εκπαιδευόμενος ταξινομητής έχει εκ των προτέρων όλα τα διαθέσιμα δεδομένα εκπαίδευσης και τα χρησιμοποιεί για να κατασκευάσει μια υπόθεση, η οποία χρησιμοποιείται έκτοτε για την ταξινόμηση. Όμοια, στη περίπτωση συσταδοποίησης, ορίζουμε το πλήθος συστάδων που θέλουμε να δημιουργηθούν και ο αντίστοιχος αλγόριθμος μας εμφανίζει το επιθυμητό αποτέλεσμα.

ΚΕΦΑΛΑΙΟ 2
ΘΕΩΡΗΤΙΚΟ ΥΠΟΒΑΘΡΟ

2.1 ΟΡΙΣΜΟΣ ΤΟΥ ΠΡΟΒΛΗΜΑΤΟΣ

Είναι γενικά αποδεκτό ότι η ανάπτυξη ενός συστήματος πρόβλεψης της τιμής ή της τάσης των μετοχών ή γενικά των δεικτών στα Χρηματιστήρια αποτελεί αντικείμενο εκτεταμένων ερευνητικών προσπαθειών. Πράγματι, έχουν αναπτυχθεί διάφορες στρατηγικές, μοντέλα και μεθοδολογίες για την πρόβλεψη της τιμής ή της τάσης μιας μετοχής ή ενός χρηματιστηριακού δείκτη.

Συνήθως, είναι περισσότερο βολικό να προβλέπεται η τιμή ή η τάση ενός δείκτη παρά μιας μετοχής και αυτό διότι όχι μόνο απαιτούνται λιγότερα δεδομένα για να προβλεφθεί ένας δείκτης αλλά επίσης ένας δείκτης παρουσιάζει μικρότερες αυξομειώσεις (υπολογίζεται ως άθροισμα πολλών επιμέρους μετοχών). Είναι γνωστό από τη Στατιστική ότι ο μέσος όρος έχει μικρότερη διασπορά σε σχέση με τα πρωτογενή δεδομένα. Από την άλλη μεριά, όμως, οι δείκτες αυτοί καθώς αποτελούν μέσο όρο δεδομένων για την κίνηση της αγοράς, δεν περιγράφουν την κίνηση της μετοχής και συνεπώς δίνουν λανθασμένες εντυπώσεις για την απόδοση μιας μεμονωμένης μετοχής. Η αποτελεσματικότητα των δεικτών αυτών εξαρτάται από το ποσοστό της αγοράς που αντιπροσωπεύουν. Συμβαίνει πολύ συχνά, για παράδειγμα, ο γενικός δείκτης του χρηματιστηρίου να είναι ανοδικός και κάποιες μετοχές να είναι καθοδικές και αντιστρόφως (Ατσαλάκης, 2006).

Για το λόγο αυτό η πρόβλεψη της τιμής ή της τάσης μιας μετοχής είναι η πλέον χρήσιμη πρόβλεψη από πρακτικής πλευράς για τους εμπλεκόμενους στις χρηματιστηριακές αγορές. Οι εκτός δείγματος θετικές προβλέψεις αποτελούν σήμα αγοράς και οι αρνητικές σήμα πώλησης της μετοχής.

Στη παρούσα εργασία η πρόβλεψη των χρηματιστηριακών δεικτών θα γίνει χρησιμοποιώντας, κατά κύριο λόγο, εξόρυξη δεδομένων και πιο συγκεκριμένα εξόρυξη κειμένου. Θα χρησιμοποιηθούν υβριδικά δεδομένα προκειμένου να αποδοθεί πρόβλεψη των δεικτών. Με τον όρο υβριδικά, εννοούμε ότι η πρόβλεψη θα γίνει τόσο χρησιμοποιώντας κειμενικά δεδομένα, τα οποία μέσω ενός αλγορίθμου θα μετατραπούν σε αριθμητικά, όσο και με αριθμητικά δεδομένα. Ουσιαστικά, λοιπόν, η εξόρυξη γνώσης στη προκειμένη θα γίνεται είτε από κάποιο άρθρο αναφερόμενο σε τραπεζικές μετοχές είτε από τον ίδιο το δείκτη τραπεζικών μετοχών.

2.2 Η ΕΝΝΟΙΑ ΤΗΣ ΧΡΟΝΟΣΕΙΡΑΣ

Με τον όρο χρονοσειρά (Λισγάρα, 2011) εννοούμε την ακολουθία σημείων σχετιζόμενων με τη περιγραφή ενός φαινομένου, ανά τακτά χρονικά διαστήματα. Η χρονοσειρά μπορεί να αντιπροσωπεύει σεισμικές καταγραφές, ημερήσιες τιμές μετοχών, τη μέγιστη ή την ελάχιστη τιμή μέτρων όπως η θερμοκρασία καθώς και αποδόσεις. Κύριο γνώρισμά μιας χρονοσειράς είναι ότι χαρακτηρίζουν επακριβώς το υπό περιγραφή μοντέλο και για το λόγο αυτό χρησιμοποιείται σε αρκετούς κλάδους.

Στη περίπτωση των χρηματοοικονομικών χρονοσειρών, η χρήση και η εξαγωγή συμπερασμάτων, χρησιμοποιώντας τις κατάλληλες τεχνικές, θεωρείται μια από τις βασικότερες και πιο αναγκαίες προκειμένου να ληφθεί η σωστή απόφαση. Η ανάλυση των χρονοσειρών δίνει τη δυνατότητα ελέγχου των προγενέστερων τιμών σε ποσοτικό επίπεδο (διαδικασία παραγωγής) αλλά και τη δυνατότητα παροχής πληροφόρησης για μελλοντικά χρηματοοικονομικά δεδομένα που επηρεάζουν τη διαδικασία λήψης αποφάσεων.

Ένα εναλλακτικός ορισμός της χρονοσειράς χρησιμοποιώντας μαθηματικούς όρους δίνεται ως εξής:

Ως χρονοσειρά θεωρείται μια ακολουθία σημείων τοποθετημένα αναλόγως στο χρόνο, τα οποία περιγράφουν μια μεταβλητή η οποία συμβολίζεται ως *Y*. Θεωρώντας, τώρα, t τα χρονικά διαστήματα χρόνου *T* και *Yt* τη χρονοσειρά για *t=1,2, ... ,T*, η οποία περιλαμβάνει δύο σημαντικά στοιχεία:

1. Το χρονικό σημείο αναφοράς.
2. Το μέγεθος της χρονοσειράς τη χρονική στιγμή *t.*

Τότε η Y δίνεται ακολούθως:

$$Y = \{Y_t : t \in T\}$$

2.1.1 ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ ΧΡΟΝΟΣΕΙΡΩΝ

Η συστηματική μελέτη των χρονοσειρών αναφέρεται στο πεδίο του χρόνου της παρατήρησης του φαινομένου καθώς και στο μέγεθος που αυτό ερμηνεύει. Τρία είναι τα βασικότερα ποιοτικά χαρακτηριστικά που προκύπτουν από τη μελέτη μιας χρονοσειράς (Λισγάρα, 2011):

- Η τάση (*trend*). Αναφέρεται στη μακροχρόνια μεταβολή (αύξηση ή μείωση) της μέσης τιμής της παρατηρούμενης μεταβλητής σε συγκεκριμένο χρονικό διάστημα. Συνεπώς, η τάση μπορεί να είναι γραμμική (ανοδική ή πτωτική), να μην είναι γραμμική αλλά και σταθερή σε κάθε περίοδο.
- Η εποχικότητα (*seasonality*). Αφορά το φαινόμενο της περιοδικής μεταβολής εντός ενός χρονικού διαστήματος. Με άλλα λόγια, η εποχικότητα μπορεί να ορισθεί σαν μια περιοδική διακύμανση *που έχει σ*ταθερό μήκος.
- Οι ακραίες τιμές (*outliers*), οι οποίες αναφέρονται στις παρατηρήσεις που τείνουν να αποκλίνουν σε μεγάλο βαθμό από το μέσο όρο του συνόλου τιμών.

Αναφέρονται επίσης, ως λιγότερο σημαντικά χαρακτηριστικά που αντλούνται από τη μελέτη των χρονοσειρών οι διαλείψεις και η στασιμότητα. Οι πρώτες έχουν να κάνουν με την έλλειψη τιμών για κάποιο χρονικό διάστημα, ενώ η στασιμότητα αναφέρεται στην σταθερότητα των στατιστικών χαρακτηριστικών της χρονοσειράς, Επίσης, στη περίπτωση της στασιμότητας τόσο η μέση τιμή όσο και η διακύμανση των τιμών παραμένουν σταθερές.

2.1.1 ΑΝΑΛΥΣΗ ΚΑΙ ΠΡΟΒΛΕΨΗ ΧΡΟΝΟΣΕΙΡΩΝ

Η ανάλυση των χρονοσειρών αναφέρεται στην μελέτη της διαχρονικής εξέλιξης των τιμών μιας μεταβλητής οι οποίες περιγράφονται από τη χρονοσειρά και αποσκοπεί στα ακόλουθα (Λισγάρα, 2011):

I. Στον χαρακτηρισμό του φαινομένου που αναπαριστά η χρονοσειρά.
II. Στην εύρεση ενός μοντέλου που περιγράφει επακριβώς τη πορεία του φαινομένου.
III. Στην εξαγωγή συμπερασμάτων σχετιζόμενων με μελλοντική συμπεριφορά αυτού (πρόβλεψη μελλοντικής τιμής του φαινομένου).

Η ανάγκη λήψης σωστής απόφασης προκειμένου για όσο το δυνατόν ασφαλέστερες επενδυτικές αποφάσεις εκ μέρους των επενδυτών οδήγησε στη συστηματικότερη πρόβλεψη των χρονοσειρών. Η πρόβλεψη έγκειται στη δυνατότητα εκτίμησης της μελλοντικής συμπεριφοράς ενός φαινομένου. Θα μπορούσαμε να πούμε, λοιπόν, ότι με τον όρο πρόβλεψη

εννοούμε την εκτίμηση της τιμής ή κατάστασης, αναφερόμενοι σε μελλοντική χρονική περίοδο (Armstrong). Η πρόβλεψη των δεικτών που μας ενδιαφέρουν στη παρούσα εργασία θα διενεργηθεί χρησιμοποιώντας κατάλληλες τεχνικές που θα αναλυθούν εκτενώς στις επόμενες ενότητες.

2.1.2 ΒΑΣΙΚΑ ΣΤΟΙΧΕΙΑ ΠΡΟΒΛΕΨΗΣ

Αρχικά, αξίζει να σημειωθεί ότι καμία πρόβλεψη δεν είναι 100% ακριβής (Βαϊδάνης, 2005). Με άλλα λόγια υπάρχει αβεβαιότητα, δηλαδή σφάλμα το οποίο ισούται με τη διαφορά της πρόβλεψης από τη πραγματικότητα. Συνεπώς, στόχος μας είναι η ελαχιστοποίηση του σφάλματος, για την όσο το δυνατόν ακριβέστερη προσέγγιση της πραγματικότητας.

Επιπλέον, η πρόβλεψη παρουσιάζει μεγαλύτερη ακρίβεια σε περιπτώσεις ομάδων παρά σε μεμονωμένα στοιχεία. Αυτό συμβαίνει καθώς οι μέγιστες και οι ελάχιστες τιμές των στοιχείων αλληλοεξουδετερώνονται με αποτέλεσμα οι ομάδες στοιχείων να έχουν πιο σταθερή συμπεριφορά από τα μεμονωμένα.

Τέλος, η πρόβλεψη είναι ακριβέστερη όταν γίνεται βραχυπρόθεσμα και όχι μακροπρόθεσμα. Πράγματι, όσο μικρότερο είναι το χρονικό διάστημα πρόβλεψης, τόσο μικρότερος είναι ο βαθμός αβεβαιότητας και κατ' επέκταση τόσο μικρότερο το σφάλμα που θα περιέχει.

Όσον αφορά τα πλεονεκτήματα του μοντέλου χρονοσειρών προκειμένου να διενεργηθεί πρόβλεψη, διαπιστώνουμε ότι:

- Δεν υφίσταται (πάντα) η δυνατότητα συσχέτισης ή του προσδιορισμού της αλληλεπίδρασης μεταξύ ενός μεταβαλλόμενου μεγέθους με κάποιους παράγοντες.
- Μας ενδιαφέρει, κυρίως, να προσδιορίσουμε το τι θα συμβεί μελλοντικά και όχι τον λόγο που θα συμβεί.
- Η χρήση του μοντέλου χρονοσειρών παρουσιάζει μικρότερο κόστος.

2.2 ΟΡΙΣΜΟΣ TEXT MINING

Είναι γεγονός ότι η ραγδαία ανάπτυξη του διαδικτύου και του Παγκόσμιου Ιστού (Παχίδη, 2008), καθώς και η εισαγωγή των πληροφοριακών συστημάτων σε υπηρεσίες και οργανισμούς τόσο για την εσωτερική τους λειτουργία όσο και για την εξυπηρέτηση του κοινού, έχουν ως αποτέλεσμα τη συνεχή παραγωγή, διακίνηση και αποθήκευση τεράστιου όγκου πληροφορίας σε τρομερές ταχύτητες καθημερινά, μέσω δεδομένων διαφορετικού περιεχομένου, ακόμη και διαφορετικού τύπου (κείμενα, εικόνες, audio, video). Συνεπώς, παρατηρούμε να αυξάνεται με αλματώδη τρόπο το σύνολο των κειμένων τα οποία τις περισσότερες φορές δεν είναι δομημένα ενώ παράλληλα δύναται να είναι γραμμένα σε διάφορους τύπους κειμένων (άρθρα, e-mail, δημοσιεύσεις) αλλά και σε διαφορετικές γλώσσες.

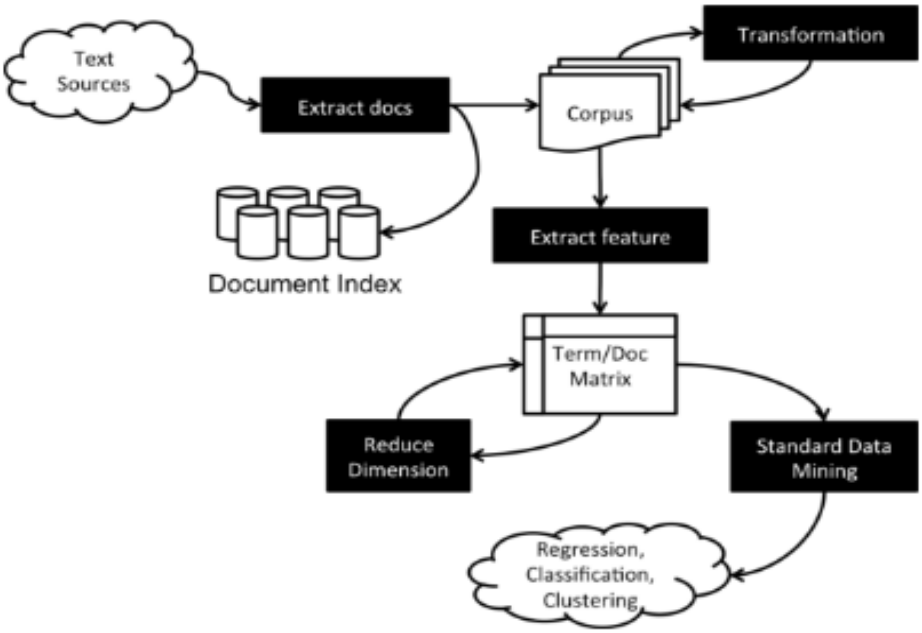
Η υπερφόρτωση της πληροφορίας, όμως, κάνει όλο και πιο εμφανές το πρόβλημα της κατανόησης και του χειρισμού της από τους χρήστες ενώ ταυτόχρονα τα περισσότερα κείμενα δεν μπορούν να υποβληθούν σε αυτόματη επεξεργασία με κάποιο τυποποιημένο τρόπο εξαιτίας της μη σύνδεσής τους με μεταδεδομένα (*metadata*: δεδομένα τα οποία χρησιμοποιούνται για την περιγραφή και αναφορά σε άλλα δεδομένα). Συνεπώς, γίνεται εύκολα αντιληπτή η ανάγκη για ανάπτυξη αυτοματοποιημένων τεχνικών για την ανακάλυψη, ανάλυση και επεξεργασία πληροφοριών από κειμενικά δεδομένα.

Η ανακάλυψη γνώσης σε κείμενο (*Knowledge Discovery in Text - KDT*) καθώς και η εξόρυξη κειμένου (*Text Mining*) περιλαμβάνουν αυτοματοποιημένες τεχνικές για την ανάλυση πολύ μεγάλων συλλογών από δεδομένα αλλά και την εξαγωγή χρήσιμων

πληροφοριών από αυτά, οι οποίες βρίσκονται σήμερα στο επίκεντρο του ενδιαφέροντος τόσο από εμπορική όσο και από επιστημονική πλευρά. Χρησιμοποιώντας τεχνικές από την εξόρυξη δεδομένων, την μηχανική μάθηση, τη στατιστική, την επεξεργασία φυσικής γλώσσας, την ανάκτηση πληροφορίας, την εξαγωγή πληροφορίας και τη διαχείριση γνώσης, οι τεχνικές αυτές προσπαθούν να επιλύσουν το πρόβλημα της μετατροπής των τεραστίων ποσοτήτων από δεδομένα, σε χρήσιμη γνώση.

Μπορούμε να πούμε ότι η εξόρυξη κειμένου αποτελεί ένα στάδιο της ανακάλυψης γνώσης σε κείμενο, η οποία είναι μια διαδικασία που περιλαμβάνει πολλά βήματα για την ανεύρεση χρήσιμης πληροφορίας. Η διαδικασία της εξόρυξης γνώσης από κείμενο χρησιμοποιεί πολύ μεγάλα σύνολα κειμένων (γνωστά και ως *corpora*) που είναι αποθηκευμένα είτε στο διαδίκτυο είτε συμβατικά, και περιλαμβάνει την ανακάλυψη προτύπων (*patterns*) ανάμεσα στα σύνολα δεδομένων (που εμπεριέχονται στα κείμενα), τα οποία προηγουμένως δεν ήταν γνωστά. Η ανάλυση αυτών των προτύπων οδηγεί στην εύρεση μη αναμενόμενων συσχετίσεων ανάμεσα στα δεδομένα και να τα σύνοψή τους με νέους τρόπους που είναι κατανοητοί και χρήσιμοι στους χρήστες.

Η επεξήγηση του όρου ‘πρότυπο’ αποδίδεται θεωρώντας τα δεδομένα μας ως ένα σύνολο γεγονότων *F* (έστω περιπτώσεις σε μια βάση δεδομένων). Το πρότυπο είναι ένας κανόνας *E* ο οποίος περιγράφει γεγονότα σε ένα υποσύνολο *FE* του *F*. Μπορεί να έχουμε είτε πρότυπα πρόβλεψης (*predictive pattern*), με σκοπό την πρόβλεψη ενός ή περισσότερων γνωρισμάτων (*attributes*) από αυτά που υπάρχουν στη βάση, είτε πρότυπα ενημέρωσης (*informative pattern*) τα οποία δεν επιλύουν κάποιο συγκεκριμένο πρόβλημα αλλά παρουσιάζουν στο χρήστη ενδιαφέροντα πρότυπα που θα έπρεπε να γνωρίζει. Έτσι, απλούστερα λέμε ότι το Text Mining εξετάζει μεγάλες συλλογές από έγγραφα μη δομημένων κειμένων προκειμένου να ανακαλύψει τη δομή καθώς και αυτονόητα νοήματα που κρύβονται μέσα στο κείμενο. Συμπερασματικά, όπως η εξόρυξη δεδομένων εντοπίζει συνδέσεις και συσχετίσεις που δεν ήταν προηγουμένως γνωστές ανάμεσα σε δομημένα δεδομένα, έτσι και η εξόρυξη κειμένου βρίσκει συνδέσεις ανάμεσα σε κείμενα, τα οποία όμως αποτελούν μη δομημένα δεδομένα. Στο παρακάτω γράφημα, φαίνεται η διαδικασία Text Mining:



ΕΙΚΟΝΑ 2-1: ΔΙΑΔΙΚΑΣΙΑ TEXT MINING (ΠΗΓΗ:DZONE.COM)

2.1.1 ΕΝΝΟΙΟΛΟΓΙΚΑ ΘΕΜΕΛΙΑ TEXT MINING

Πριν εντρυφήσουμε στις τεχνικές του Text Mining, είναι απαραίτητο να επεξηγηθούν ορισμένες έννοιες της μεθόδου (Μακρής, 2015).

- Αδόμητα δεδομένα:

Με την λέξη αδόμητα εννοούμε ότι το κείμενο δεν έχει κάποια συγκεκριμένη μορφή και αποτελείται από σκόρπιες πληροφορίες και λέξεις. Πολλές φορές υπάρχει περίπτωση να είναι και σε ημιδομημένη μορφή δηλαδή κάποιες πληροφορίες να είναι σε ένα υπολογιστικό φύλλο ή σε μία βάση δεδομένων ανάλογα με τις πληροφορίες που παρέχει και συγχρόνως να έχει και κάποιες σκόρπιες πληροφορίες. Και στις δύο περιπτώσεις, η μέθοδος που ακολουθούμε είναι ίδια (η περεταίρω αξιοποίηση των διαθέσιμων εγγράφων απαιτεί τη χρήση αυτοματοποιημένων μέσων).

- Σύνταξη:

Ένας υπολογιστής μπορεί να εξετάσει τους χαρακτήρες των λέξεων και την θέση τοποθετούνται οι λέξεις σε μία πρόταση αλλά δεν μπορεί να γνωρίζει τις πληροφορίες που ένα κείμενο μεταδίδει ενώ μπορεί να κατανοήσει τη δομή και τη σύνταξή του. Με τον όρο σύνταξη, λοιπόν, εννοούμε τη δομή της γλώσσας και το πως μεμονωμένες λέξεις συνδυάζονται για να κάνουν προτάσεις και παραγράφους. Ειδικοί κανόνες γραμματικής και γλώσσας διέπουν τον τρόπο με τον οποίο χρησιμοποιείται το λεξιλόγιο. Αυτή η δομή αποτελεί μια οργανωμένη διαδικασία και είναι σχετικά εύκολη για έναν υπολογιστή να την επεξεργαστεί. Για την πλήρη κατανόηση της σημασίας ενός κειμένου, παρόλα αυτά, απαιτείται και η σημασιολογία των μεμονωμένων λέξεων.

- Σημασιολογία:

Χρησιμοποιώντας κοινά ιδιώματα τονίζεται διαφορά μεταξύ της σύνταξης και της σημασιολογίας. Ιδιώματα είναι λέξεις ή φράσεις με μεταφορική σημασία (δηλαδή φράσεις με διαφορετική από την κυριολεκτική έννοια των λέξεων). Χρησιμοποιώντας το ακόλουθο παράδειγμα διαπιστώνουμε τη δυσκολία της πλήρους κατανόησης:

«Η Μαρία έχει πεταλούδες στο στομάχι πριν από κάθε παράσταση» είναι συντακτικά σωστή και έχει δύο σημασιολογικές ερμηνείες: Τη γραμματική, που ερμηνεύει ότι η Μαρία πριν από την παράσταση τρώει πεταλούδες και την ιδιωματική ερμηνεία, ότι δηλαδή η Μαρία είναι νευρική και αγχωμένη πριν από κάθε παράσταση. Σαφώς η πλήρης σημασιολογική έννοια του κειμένου είναι δύσκολο να προσδιοριστεί αυτόματα χωρίς την εκτεταμένη κατανόηση της γλώσσας που χρησιμοποιείται.

Ευτυχώς στις περισσότερες περιπτώσεις μόνο η σύνταξη μπορεί να χρησιμοποιηθεί για να εξάγουμε την πρακτική αξία από το κείμενο, χωρίς να είναι απαραίτητη η σημασιολογική κατανόηση. Κατά την ταξινόμηση εγγράφων και την ανάκτηση πληροφοριών συμβαίνει η ενασχόληση των εκάστοτε αλγορίθμων με την κατάταξη ή την εύρεση συγκεκριμένων τύπων εγγράφων σε μία μεγάλη βάση δεδομένων. Η βασική ιδέα πίσω από αυτούς τους αλγορίθμους είναι ότι η συντακτική ομοιότητα (παρόμοιες λέξεις) συνεπάγεται σημασιολογική ομοιότητα (παρόμοια έννοια). Αν και βασίζεται σε συντακτικές πληροφορίες οι προσεγγίσεις αυτές λειτουργούν επειδή είναι έγγραφα που μοιράζονται πολλές λέξεις κλειδιά συχνά για το ίδιο θέμα. Σε άλλες περιπτώσεις ο στόχος είναι η σημασιολογική έννοια.

- Σειρά των λέξεων:

Στους περισσότερους αλγορίθμους το κείμενο αντιπροσωπεύεται από στοιχεία που δείχνουν την εμφάνιση των λέξεων μέσα στο κείμενο, οδηγώντας σε μεγάλο αριθμό διαστάσεων. Υπάρχει, λοιπόν, μία σιωπηρή παραδοχή βάσει της οποίας η σειρά στο έγγραφο δεν έχει καμία σημασία αλλά το ενδιαφέρον μας βρίσκεται στο πόσο συχνά βρίσκονται κοντά δύο ή και παραπάνω λέξεις μέσα στο κείμενο. Δεδομένου όμως ότι η κατανόηση ενός κειμένου προϋποθέτει ότι θα διαβαστούν οι λέξεις που το συντελούν με συγκεκριμένη σειρά, η προηγούμενη πρόταση φαντάζει 'παράξενη'. Για τις περισσότερες εφαρμογές όμως αυτό δεν είναι πρόβλημα, αφού η συλλογή των λέξεων που αναγράφεται στο έγγραφο με οποιαδήποτε σειρά αποτελεί συνήθως αρκετή πληροφορία για να γίνει η σημασιολογική διαφοροποίηση. Η κύρια δύναμη των αλγορίθμων είναι η ικανότητα τους να βρίσκουν όλες τις λέξεις κλειδιά ενός κειμένου. Ισχύει, τέλος, ότι οι προαναφερθέντες λέξεις όντας μόνες δεν διαφοροποιούν το έγγραφο ενώ όταν συνδυάζονται με άλλες δευτερεύουσες έχουν την ικανότητα να το διαφοροποιούν.

2.1.2 ΤΕΧΝΙΚΕΣ ΕΞΟΡΥΞΗΣ ΚΕΙΜΕΝΟΥ

Οι κυριότερες κατηγορίες των μεθόδων που χειρίζεται η εξόρυξη κειμένου είναι (Παχίδη, 2008):

- Εξαγωγή χαρακτηριστικών γνωρισμάτων (*Feature Extraction*)

Έχει ως στόχο τον προσδιορισμό γεγονότων και σχέσεων στο κείμενο, διακρίνοντας (συχνά) εάν κάποια ονομαστική φράση είναι πρόσωπο, θέση, οργανισμός ή άλλο διακριτό αντικείμενο. Οι αλγόριθμοι εξαγωγής χαρακτηριστικών περιλαμβάνουν την εξαγωγή ονόματος (εντοπίζονται εμφανίσεις ονομάτων στο κείμενο και καθορίζεται σε ποιο τύπο οντότητας αναφέρεται το όνομα), την εξαγωγή όρου μιας περιοχής (προσδιορισμός τεχνικών όρων σε ένα κείμενο) αναγνώριση συντμήσεων (προσδιορίζονται συντμήσεις και αρκτικόλεξα και αντιστοιχούνται στην πλήρη μορφή τους). Αυτό περιλαμβάνει την επιλογή σημαντικών όρων και την απόρριψη των μη σημαντικών, καθώς και τον υπολογισμό της συχνότητας εμφάνισης των όρων. Επίσης, οι όροι πρέπει να βρίσκονται σε κανονική ή καθιερωμένη μορφή. Η διαδικασία αυτή μπορεί να χρησιμοποιεί λεξικά για τον προσδιορισμό μερικών όρων καθώς και γλωσσικά υποδείγματα για την ανίχνευση άλλων.

- Πλοήγηση με βάση το κείμενο (*Text Based Navigation*)

Περιλαμβάνει την αναζήτηση σε εσωτερικές συλλογές εγγράφων ή σε συλλογές που βρίσκονται στον Παγκόσμιο Ιστό. Κύριο χαρακτηριστικό αποτελεί η δυνατότητα, αφού αρχικά συνταχθεί ένα ευρετήριο, να προσφέρεται ένα αρκετά ευρύ φάσμα επιλογών αναζήτησης κειμένου, στις οποίες συμπεριλαμβάνονται οι βασικές επιλογές αναζήτησης (η Boolean, η index-based) αλλά και πιο σύνθετες επιλογές αναζήτησης (όπως *relevancy*, έρευνα φυσικής γλώσσας).

- Κατηγοριοποίηση, κατάταξη με επίβλεψη (*Categorization, Supervised Classification*)

Με τον όρο κατηγοριοποίηση εννοούμε τη διαδικασία της κατάταξης εγγράφων σε προκαθορισμένες κατηγορίες. Η χρησιμότητά της έγκειται στον προσδιορισμό των κύριων θεμάτων μιας συλλογής εγγράφων. Οι κατηγορίες είτε έχουν διαμορφωθεί εξαρχής από τον προγραμματιστή είτε μπορούν να προσδιοριστούν από το χρήστη. Υπάρχουν δύο τρόποι για την κατηγοριοποίηση:

1. Ο πρώτος τρόπος περιλαμβάνει τη δημιουργία ενός θησαυρού (*thesaurus*), δηλαδή ενός συνόλου που περιλαμβάνει όρους σχετικούς με το θέμα κάθε κατηγορίας καθώς και συσχετίσεις μεταξύ αυτών των όρων (διευρυμένους όρους, κοντινότερους όρους, συνώνυμα, σχετικούς όρους) και τελικά τον ορισμό του θέματος του κειμένου με βάση τη συχνότητα των όρων σχετικών με το θέμα που υπάρχουν στο έγγραφο.

2. Ο δεύτερος τρόπος περιλαμβάνει την εκπαίδευση (*training*) του εργαλείου κατηγοριοποίησης με κάποια δείγματα από τα έγγραφα, τη στατιστική ανάλυση λεκτικών προτύπων (*linguistic patterns*) όπως είναι οι λεξικολογικές συγγένειες, οι συχνότητες λέξεων των εγγράφων προς εκπαίδευση, το χωρισμό αυτών των προτύπων σε κατηγορίες (με στατιστικό τρόπο), και τέλος την ταξινόμηση των υπόλοιπων εγγράφων. Η δεύτερη προσέγγιση είναι προτιμότερη όταν έχουμε να κάνουμε με μεγάλους τομείς, καθώς τότε είναι αρκετά δύσκολο να δημιουργηθεί κάποιος θησαυρός εννοιών.

Γενικότερα, ο στόχος της διαδικασίας αυτής είναι η ανάπτυξη ενός μοντέλου, το οποίο αργότερα θα μπορεί να χρησιμοποιηθεί για την κατηγοριοποίηση μελλοντικών δεδομένων.

Οι πιο γνωστές μέθοδοι κατηγοριοποίησης είναι τα δέντρα απόφασης (*Decision Trees*), η Bayesian κατηγοριοποίηση , η κατηγοριοποίηση *k* πλησιέστερου γείτονα (*K-Nearest-Neighbor*) και κατηγοριοποίηση με νευρωνικά δίκτυα (*Neural Networks*).

❖ Ομαδοποίηση, μη επιβλεπόμενη κατάταξη (*Clustering, Unsupervised Classification*)

Μία ομάδα (*cluster*) είναι μια συλλογή από σχετικά έγγραφα και συνεπώς η ομαδοποίηση (*clustering*) είναι η διαδικασία της δημιουργίας ομάδων εγγράφων βάσει κάποιου κριτηρίου ομοιότητας, αυτόματα χωρίς να έχουμε προσδιορίσει από πριν τις κατηγορίες και αυτό είναι που την ξεχωρίζει από τη κατηγοριοποίηση. Η ομαδοποίηση κειμένων είναι χρήσιμη για τον προσδιορισμό κρυμμένων ομοιοτήτων, για να διευκολύνει τη διαδικασία του να βρούμε παρόμοιες ή σχετικές πληροφορίες, ενώ επιπλέον δίνει τη δυνατότητα γενικής επισκόπησης μιας νέας συλλογής δεδομένων ενώ εξερευνάται. Οι πιο γνωστοί αλγόριθμοι που χρησιμοποιούνται είναι ιεραρχικοί (*hierarchical*), διαχωριστικοί (*partitional*) και ασαφείς (*fuzzy*). Σημειώνουμε ότι ο πιο σημαντικός παράγοντας στη λειτουργία της ομαδοποίησης είναι το μέτρο ομοιότητας που χρησιμοποιεί ο εκάστοτε αλγόριθμος, καθώς υπάρχουν διάφοροι τύποι μέτρων όπως η θεώρηση λέξεων που εμφανίζονται συχνά μαζί ως κοινά χαρακτηριστικά, ενώ ένας άλλος τύπος μπορεί να περιλαμβάνει χαρακτηριστικά γνωρίσματα που έχουν εξαχθεί (το όνομα ενός προσώπου).

Ο στόχος της ομαδοποίησης είναι να επιλεγούν τα καταλληλότερα γνωρίσματα στα οποία πρόκειται να εφαρμοστεί η ομαδοποίηση ώστε να επιτυγχάνεται η βέλτιστη ομοιογένεια σε κάθε συστάδα. Έτσι η προ-επεξεργασία των δεδομένων πριν την εφαρμογή της διαδικασίας ομαδοποίησης κρίνεται απαραίτητη.

❖ Περιληπτική Παρουσίαση της Πληροφορίας (*Summarization*)

Αποτελεί την εξαγωγή της περίληψης ενός κειμένου, δηλαδή τη μείωση του μεγέθους του κειμένου διατηρώντας όμως τα βασικά στοιχεία του περιεχομένου του. Σε αυτή τη λειτουργία ο χρήστης έχει συνήθως τη δυνατότητα να καθορίσει διάφορες παραμέτρους, όπως το πλήθος των λέξεων που θα εξαχθούν ή το ποσοστό επί του συνολικού κειμένου που θα αποτελεί την περίληψη.

- ❖ Γλωσσικός προσδιορισμός (*Language Identification*) και απόδοση κειμένου στο συγγραφέα

Ένα εργαλείο language identification μπορεί να προσδιορίσει σε ποια γλώσσα είναι γραμμένο ένα κείμενο, ή και τι ποσοστό του κειμένου είναι γραμμένο σε κάθε γλώσσα, εάν αυτό είναι γραμμένο σε περισσότερες. Επιπλέον, υπάρχει η δυνατότητα προσδιορισμού του συγγραφέα στον οποίο ανήκει το κείμενο, χρησιμοποιώντας τεχνικές data mining.

- ❖ Συσχετίσεις (*Associations*)

Στην ανάλυση συσχετίσεων αναγνωρίζονται σχέσεις μεταξύ χαρακτηριστικών γνωρισμάτων που έχουν εξαχθεί από τη συλλογή εγγράφων και ορίζεται ένα πρότυπο με τη χρήση μιας αντικειμενικής συσχέτισης. Για να οριστεί η έννοια του προτύπου θεωρούμε την ύπαρξη δύο υπολέξεων όπου η μια ακολουθεί την άλλη σε συγκεκριμένη απόσταση. Συνεπώς, το πρότυπο εκφράζει έναν κανόνα, σύμφωνα με τον οποίο η αντικειμενική συνθήκη που υφίσταται μεταξύ των δύο υπολέξεων θα διατηρηθεί με μεγάλη συχνότητα. Οι κανόνες αυτοί είναι πολύ ευέλικτοι για την περιγραφή των τοπικών ομοιοτήτων που περιέχονται στα δεδομένα του κειμένου.

- ❖ Απεικόνιση – Οπτικοποίηση (*Visualization*)

Η απεικόνιση χρησιμοποιεί την εξαγωγή χαρακτηριστικών γνωρισμάτων και το ευρετήριο βασικών όρων για να κατασκευάσει μια γραφική αναπαράσταση μιας συλλογής εγγράφων. Η προσέγγιση αυτή δίνει τη δυνατότητα στο χρήστη να αναγνωρίζει πολύ γρήγορα τα κύρια θέματα και τις βασικές έννοιες των κειμένων, με βάση τη σπουδαιότητα τους κατά την αναπαράσταση.

2.1.3 ΑΝΑΠΑΡΑΣΤΑΣΗ ΚΕΙΜΕΝΟΥ

Λόγω της συχνής έλλειψης κάποιας δομής στα αρχεία κειμένων, είναι προφανής η ανάγκη εύρεσης μια αναπαράστασης για την αντιπροσώπευση των στοιχείων-όρων των κειμένων, έτσι ώστε να είναι δυνατή η μετέπειτα επεξεργασία τους (Παχίδη, 2008).

Όταν έχουμε μια συλλογή από αρχεία κειμένου, μπορούμε να θεωρήσουμε καθένα από αυτά ως ένα *bag-of-words*, μια 'σακούλα' η οποία περιλαμβάνει όλες τις λέξεις που βρίσκονται στο κείμενο. Ο συχνότερος τρόπος αναπαράστασης ενός κειμένου είναι η αναπαράσταση διανύσματος (*vector representation*), η οποία προέρχεται από τα συστήματα ανάκτησης πληροφορίας (*information retrieval*). Έτσι, κάθε υποσύνολο δεδομένων από το σύνολο κειμένων που έχουμε είναι και ένα διάνυσμα όρων (*term vector*) στο οποίο κάθε όρος αποτελεί ένα μοναδικό ανεξάρτητο χαρακτηριστικό. Κάθε στοιχείο σε αυτό το διάνυσμα έχει και μια τιμή η οποία αντιστοιχεί στην εμφάνιση του όρου μέσα στο κείμενο. Με βάση αυτό μπορούμε να διακρίνουμε διάφορα μοντέλα διανυσματικής αναπαράστασης των κειμένων όπως:

I. Το λογικό μοντέλο (*Boolean model*) όπου κάθε έγγραφο αναπαρίσταται από ένα σύνολο λογικών τιμών και κάθε μία από τις οποίες δηλώνει εάν ένας συγκεκριμένος όρος εμφανίζεται στο έγγραφο (συνήθως η τιμή 1 σημαίνει ότι εμφανίζεται και η τιμή 0 σημαίνει απουσία του συγκεκριμένου όρου από το κείμενο). Τα πλεονεκτήματα του λογικού μοντέλου είναι η ευκολία και η ταχύτητα λειτουργιών ερώτησης, αναζήτησης κλπ και η δυνατότητα χρησιμοποίησης της Boolean άλγεβρας. Το μειονέκτημα του μοντέλου είναι ότι η απάντηση στο κατά πόσον είναι σχετικό ένα κείμενο με ένα

συγκεκριμένο όρο (και κατ' επέκταση θέμα) είναι μια δυαδική απόφαση, ενώ επιπλέον μία λογική τιμή για κάθε χαρακτηριστικό δεν μπορεί να αποδώσει κατά πόσο σημαντική είναι η παρουσία μίας λέξης σε ένα κείμενο, γεγονός το οποίο συχνά μπορεί να οδηγήσει σε λάθος συμπεράσματα.

II. Το μοντέλο διανυσματικού χώρου (*Vector Space Model*) όπου τα αρχεία αναπαρίστανται ως διανύσματα σε ένα πολυδιάστατο Ευκλείδειο χώρο. Κάθε άξονας στο χώρο αντιστοιχεί σε ένα χαρακτηριστικό (*attribute*), δηλαδή σε έναν όρο-λέξη, με αποτέλεσμα η συντεταγμένη κάθε διανύσματος ως προς έναν άξονα να χαρακτηρίζει την εμφάνιση του όρου (στον οποίο αντιστοιχεί ο άξονας) στο συγκεκριμένο διάνυσμα-αρχείο κειμένου και μάλιστα να αποτελεί ένα 'βάρος' του όρου ως προς το συγκεκριμένο κείμενο (υποδηλώνει τη σημαντικότητα του όρου). Τα βάρη που χρησιμοποιούνται για κάθε χαρακτηριστικό είναι πραγματικές τιμές και συνήθως υποδηλώνουν τη συχνότητα εμφάνισης της λέξης. Τελικά, μια συλλογή εγγράφων αναπαρίσταται από ολόκληρο το διανυσματικό χώρο.

Ας δούμε τώρα εκτενέστερα τα βάρη (*weights*) που χρησιμοποιούνται για τις τιμές των συντεταγμένων (που αντιστοιχούν σε όρους) στο Vector Space Model. Θα θεωρήσουμε ότι έχουμε τη συντεταγμένη του αρχείου *d* που αντιστοιχεί στον άξονα του όρου *t*. Αρχικά, ορίζουμε τις ακόλουθες τιμές για τους όρους και τα αρχεία:

- ***D*** : Είναι ο αριθμός των αρχείων που συγκροτούν τη συλλογή κειμένων που έχουμε (άρα και ο αριθμός των διανυσμάτων).
- **Term Frequency -*TF(d,t)*** : Δηλώνει τη συχνότητα εμφάνισης του όρου t στο στοιχείο *d (n(d,t)).*
- **Document Frequency - *DF(t)*** : Εκφράζει το πλήθος κειμένων από τη συλλογή μας που περιέχουν τον όρο t.
- **Inverse Document Frequency - *IDF(t)*** : Εκφράζει την 'σπανιότητα' (*scarcity*) του όρου στη συλλογή κειμένων. Υπολογίζεται με διάφορους τύπους, οι συνηθέστεροι εκ των οποίων είναι: $IDF(t) = log\left(\frac{D}{DF(t)}\right)$ και $IDF(t) = log\left(\frac{1+D}{DF(t)}\right)$ ή $IDF(t) = log\left(\frac{D-DF(t)}{DF(t)}\right)$.

Βάσει αυτών διακρίνονται διάφοροι τρόποι απόδοσης βάρους w σε κάθε όρο (*term weighting*), και έτσι υπολογίζεται η τιμή της συντεταγμένης της.

Ένας πρώτος τρόπος είναι η θεώρηση *w(d,t) = TF(d,t)*, έτσι ώστε κάθε διάνυσμα να είναι της μορφής d_{tf}= (tf_1, tf_2, tf_3 , ..., tf_n). Η πιο απλή μορφή για την εύρεση των συχνότερων όρων είναι ο αριθμός δηλαδή εμφάνισης μιας λέξης σε κάθε κείμενο *(TF(d,t) = n(d,t))*. Ωστόσο συνήθως υπόκειται σε κάποια κανονικοποίηση (*length normalization*) έτσι ώστε να μειώνεται ο θόρυβος που προκαλείται από το μέγεθος κειμένων (τα οποία εκ των πραγμάτων θα εμφανίζουν περισσότερους όρους με μεγαλύτερη συχνότητα). Έτσι, υπάρχουν ποικίλοι τρόποι υπολογισμού των συχνότερων όρων, δύο από τους ευρέως χρησιμοποιούμενους φαίνεται παρακάτω :

$$TF(d,t) = \frac{n(d,t)}{max_t} \quad \text{ή} \quad TF(d,t) = 1 + log\, n(d,t)$$

Αξίζει να σημειωθεί ότι σε ένα κείμενο δεν είναι όλοι οι όροι σημαντικοί. Για παράδειγμα, λέξεις που εμφανίζονται διαρκώς όπως άρθρα, αντωνυμίες κλπ θα έχουν πολύ μεγάλη

συχνότητα και θα αποτελούν θόρυβο για την εξακρίβωση των σημαντικών όρων που καθορίζουν το περιεχόμενο ενός κειμένου. Για το λόγο αυτό, θεωρούμε ότι η σπανιότητα ενός όρου στη συλλογή κειμένων αποτελεί ένα μέτρο σημαντικότητας του όρου. Θεωρούμε, λοιπόν, ότι η σημαντικότητα είναι αντιστρόφως ανάλογη της εμφάνισης του όρου, και εισάγουμε τον όρο του inverse document frequency στον υπολογισμό του βάρους:

$$w(d,t) = TF(d,t) \times IDF(t)$$

Προκύπτει, λοιπόν, ότι μεγαλύτερη σημασία έχουν οι όροι που εμφανίζονται ούτε υπερβολικά συχνά ούτε σπάνια μέσα στο κείμενο. Ο υπολογισμός του βάρους με την προσέγγιση $TF - IDF$ είναι από τους πιο συνήθεις στον τομέα της εξόρυξης δεδομένων, και υπολογίζεται με ποικίλους τρόπους.

Δεδομένου ότι τα κείμενα μεγάλου μήκους τείνουν να έχουν μεγαλύτερες συχνότητες λέξεων καθώς και περισσότερους όρους, επέρχεται η ανάγκη κανονικοποίησης αυτών. Υπάρχουν διάφοροι τρόποι κανονικοποίησης ως προς το μήκος των αρχείων (*document length normalization*), όπως ο πολλαπλασιασμός του συχνότερου όρου με κάποιο άλλο όρο, ή η κανονικοποίηση της απόστασης μεταξύ των διανυσμάτων.

Συνοπτικά, μπορούμε να πούμε ότι για τον υπολογισμό του μοντέλου VSM, στο οποίο αντιστοιχεί μια συλλογή αρχείων, πρέπει αρχικά να γίνει μια προ-επεξεργασία των κειμένων. Ουσιαστικά, απαιτείται να αναγνωρισθούν οι λέξεις από τις οποίες αποτελείται κάθε κείμενο και για να βελτιστοποιηθεί η διαδικασία εύρεσης των σημαντικών όρων κάθε κειμένου. Οι λέξεις, συνεπώς, υπόκεινται σε διεργασίες αφαίρεσης πολύ κοινών λέξεων οι οποίες δεν έχουν νοηματική αξία (άρθρα αντωνυμίες, κλπ), εύρεσης λέξεων αντιστοιχούν στο ίδιο θέμα αλλά έχουν διαφορετική μορφή (παράγωγα) και εύρεσης των όρων που είναι οι πιο αντιπροσωπευτικοί σε κάθε κείμενο ξεχωριστά (ειδικότερη ανάλυση γίνεται στην επόμενη ενότητα). Μετά από αυτό το βήμα (*document indexing*) προχωράμε στο βήμα της ανάθεσης βάρους σε κάθε όρο για κάθε κείμενο (*term weighting*) σε όλη τη συλλογή που έχουμε, ώστε κάθε βάρος να υποδηλώνει πόσο σημαντικός θεωρείται ο εκάστοτε όρος για το αντίστοιχο κείμενο.

Τα μειονεκτήματα της μεθόδου του Vector Space Model θεωρούνται τα ακόλουθα:

1. Είναι αρκετά αργή ως προς το χρόνο επεξεργασίας λόγω της πληθώρας υπολογισμών που απαιτούνται
2. Δεν εξυπηρετεί ιδιαίτερα την ενημέρωση αλλαγών στα κείμενα εφόσον για κάθε όρο προστίθεται ένας επιπλέον άξονας και πρέπει να γίνουν υπολογισμοί τη συντεταγμένης για όλα τα διανύσματα στο χώρο
3. Η πολυδιάστατη μορφή της απαιτεί κόστος μνήμης και χαμηλή ταχύτητα σε υπολογισμούς.

2.1.4 ΠΡΟ-ΕΠΕΞΕΡΓΑΣΙΑ ΚΕΙΜΕΝΟΥ

Η προ-επεξεργασία όλων των δεδομένων που θέλουμε να αναλύσουμε, δηλαδή η μετατροπή των αδόμητων (ή και ημίδομημένων) κειμένων σε δομημένα έγγραφα ικανά για ανάλυση, αποτελεί αναγκαίο και αναπόσπαστο κομμάτι της μεθόδου Text Mining και προηγείται πριν οποιαδήποτε περεταίρω επεξεργασία. Τα βασικά βήματα της διαδικασίας καταγράφονται ως εξής (Μακρής, 2015):

1. Επιλογή των εγγράφων

Υπάρχουν περιπτώσεις στις οποίες είναι η συλλογή και η διαλογή των εγγράφων που θα υποστούν ανάλυση αποτελεί εύκολη διαδικασία. Αυτό, βέβαια, εξαρτάται κυρίως από τον λόγο για τον οποίο επιτελείται εξόρυξη κειμένου. Για του λόγου το αληθές, αν ο σκοπός μας είναι η ομαδοποίηση ή η ταξινόμηση των εγγράφων συνήθως χρησιμοποιούμε όλα τα δεδομένα που έχουμε στην διάθεση μας. Αντίθετα, αν ο λόγος του Text Mining είναι η πρόβλεψη, όπως για παράδειγμα στη περίπτωση πρόβλεψης της πιθανότητας κέρδους από τη σωστή διαχείριση χρηματιστηριακών δεικτών, τότε είναι αναγκαίο τα δεδομένα να χωριστούν και να επιλεχθούν τα πιο χρήσιμα για την εν λόγω διαδικασία.

2. Σημεία στίξης και κεφαλαία μικρά

Το πρώτο βήμα μετά την επιλογή των εγγράφων είναι η διαγραφή όλων των σημείων στίξης τα οποία δεν προσφέρουν καμία πληροφορία στο εννοιολογικό κομμάτι κάποιου κειμένου κατά συνέπεια ούτε και στην οποιαδήποτε μορφής ανάλυση του. Πέρα από αυτό, η παραμονή τους στο κείμενο μπορεί να προκαλέσει πολλά προβλήματα στην επεξεργασία καθώς μία λέξη που έχει κόμμα μπορεί να θεωρηθεί διαφορετική από την ίδια λέξη χωρίς κόμμα. Με την ίδια λογική είναι απαραίτητο να μετατρέψουμε όλα τα κεφαλαία σε μικρά ή το αντίθετο.

3. Tokenize

Το επόμενο βήμα είναι ο διαχωρισμός του κειμένου σε μεμονωμένες λέξεις. Σε ένα αδόμητο κείμενο πολλές φορές μπορεί δύο, τρεις ή και παραπάνω λέξεις να είναι ενωμένες μεταξύ τους δηλαδή να μην έχουν κενό μεταξύ τους. Ο ρόλος του Tokenize, λοιπόν, είναι να αναγνωρίζει τις λέξεις, σε συνδυασμό με κάποιο λεξικό, και να τις διαχωρίζει με κενά προκειμένου να αναλυθούν ξεχωριστά.

4. Διαγραφή των stopwords

Κατά τη προ-επεξεργασία κειμένου κρίνεται επίσης απαραίτητο να αφαιρεθούν κάποιες λέξεις οι οποίες παρουσιάζονται πολύ συχνά προκειμένου για εξοικονόμηση χώρου αποθήκευσης αλλά και για επιτάχυνση της διαδικασίας επεξεργασίας του κειμένου. Οι λέξεις αυτές ονομάζονται ‘stopword’ και η διαγραφή των stopwords γίνεται χωρίς την απώλεια σημαντικών πληροφοριών γιατί στα περισσότερα κείμενα αυτές οι λέξεις δεν έχουν καμία επίδραση στα τελικά αποτελέσματα. Αντίστοιχα, η διαδικασία αφαίρεσης ονομάζεται ‘stopping’ και κάθε αλγόριθμος Text Mining περιλαμβάνει αυτή την διαδικασία.

5. Stemming

Επόμενο βήμα στην προ-επεξεργασία κειμένου ονομάζεται stemming και σκοπό έχει λέξεις οι οποίες έχουν την ίδια βάση (δηλαδή προέρχονται από την ίδια λέξη) να αποτελέσουν μία οντότητα. Αυτή η διαδικασία περιλαμβάνει τον εντοπισμό και την αφαίρεση των προθεμάτων, των επιθεμάτων και του πληθυντικού πάλι σε συνδυασμό με λεξικό και οδηγεί στη συρρίκνωση του κειμένου, στην βελτίωση του αλγορίθμου αλλά και της ακρίβειας των αποτελεσμάτων μας.

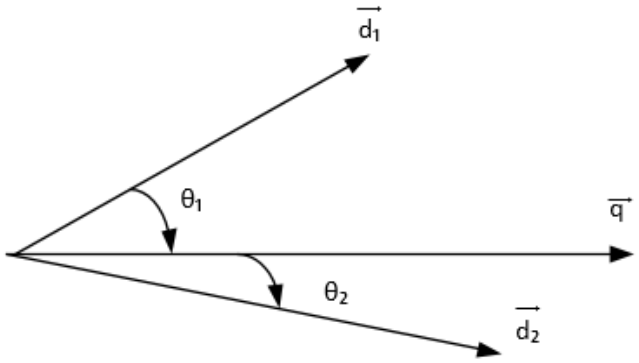
6. Διόρθωση της ορθογραφίας

Τελευταίο βήμα της διεργασίας αυτής είναι η διόρθωση της ορθογραφίας. Πράγματι, οι ανορθόγραφες λέξεις μπορούν να οδηγήσουν σε περιττή αύξηση του μεγέθους του χώρου που απαιτείται για την αποθήκευση του κειμένου κατά την επεξεργασία του. Είναι απαραίτητη, λοιπόν, η διόρθωση τέτοιων λαθών που συνήθως γίνεται με την βοήθεια λεξικών τα οποία συνδυάζονται με τον αλγόριθμο. Με άλλα λόγια, εντοπίζονται λέξεις που δεν υπάρχουν στο

λεξικό και γίνεται προσπάθεια να συνδυασθούν με άλλες που υπάρχουν με βάση τα μεμονωμένα γράμματα από τα οποία αποτελούνται.

2.1.5 ΥΠΟΛΟΓΙΣΜΟΣ ΟΜΟΙΟΤΗΤΑΣ / ΑΠΟΣΤΑΣΗΣ ΑΡΧΕΙΩΝ

Στην εξόρυξη κειμένου είναι αρκετά συνήθης η προσπάθεια εύρεσης νοηματικής ομοιότητας των αρχείων κειμένου είτε με κάποια θεματική περιοχή (κατηγοριοποίηση) είτε και μεταξύ τους (ομαδοποίηση). Δεδομένου ότι έχουμε αναπαράσταση των αρχείων κειμένου με διανύσματα, η σύγκριση της ομοιότητας μεταξύ τους ανάγεται στην σύγκριση μεταξύ των διανυσμάτων στα οποία αντιστοιχούν στο μοντέλο διανυσματικού χώρου (Παχίδη, 2008). Στο ακόλουθο σχήμα, θεωρούμε δύο διανύσματα d_1 και d_2 στο μοντέλο του διανυσματικού χώρου τα οποία αντιστοιχούν σε αρχεία κειμένου, καθώς και το διάνυσμα q το οποίο αντιστοιχεί σε ένα σύνολο όρων.



ΕΙΚΟΝΑ 2-2: ΜΟΝΤΕΛΟ ΔΙΑΝΥΣΜΑΤΙΚΟΥ ΧΩΡΟΥ

Μπορούμε εύκολα να παρατηρήσουμε λοιπόν ότι θεωρώντας ως μέτρο σύγκρισης την ευκλείδεια απόσταση, το έγγραφο d_2 είναι πιο 'κοντά' στην ερώτηση (*query*), ενώ αν θεωρήσουμε ως μέτρο το συνημίτονο της γωνίας δύο διανυσμάτων, το έγγραφο d_1 είναι πιο κοντά στο q. Αν θεωρήσουμε ως μέτρο σύγκρισης της ομοιότητας μεταξύ των αρχείων την απόσταση μεταξύ των διανυσμάτων τους στο χώρο, μπορούμε εύκολα να συμπεράνουμε ότι όσο μεγαλύτερη είναι η απόσταση μεταξύ των διανυσμάτων τόσο πιο ανόμοια είναι τα έγγραφα μεταξύ τους. Από την άλλη, αν θεωρήσουμε ως μέτρο ομοιότητας το πόσο σημαντικά είναι τα έγγραφα μεταξύ τους (δίνοντας τιμές από 0 έως 1), τότε συμπεραίνουμε ότι όσο μεγαλύτερη είναι η τιμή του μέτρου τόσο μεγαλύτερη θα είναι η ομοιότητα των εγγράφων.

Παρατηρούμε συνεπώς ότι υπάρχουν ποικίλα μέτρα για τη σύγκριση της ομοιότητας μεταξύ των διανυσμάτων και ανάλογα με τη φύση του συνόλου δεδομένων του μοντέλου που έχουμε κατασκευάσει θα πρέπει να επιλέξουμε το πιο ιδανικό μέτρο ομοιότητας (αν για παράδειγμα έχουμε binary δεδομένα δεν είναι ιδιαίτερα αποδοτική η χρήση της Ευκλείδειας απόστασης).

Επισημαίνουμε και πάλι ότι συχνά η ομοιότητα των εγγράφων ή όρων είναι αντίστοιχη της απόστασης, αυτό όμως δε συμβαίνει πάντα γι' αυτό και πρέπει να έχουμε υπόψη μας τι υπολογίζει κάθε φορά το μέτρο που χρησιμοποιούμε. Συγκεκριμένα, για να μπορεί να χρησιμοποιηθεί η απόσταση θα πρέπει η μετρική που χρησιμοποιούμε να έχει τις ιδιότητες που έχουν και τα διανύσματα στον Ευκλείδειο χώρο.

Δεδομένης μιας συλλογής από αρχεία S, εάν το $d: S \times S \rightarrow R$ είναι ένα μέτρο απόστασης, πρέπει να ικανοποιεί τις ακόλουθες προδιαγραφές:

I. $d(x,x) = 0$
II. $d(x,y) \geq 0$ όταν $x \neq y$
III. $d(x,y) = d(y,x)$ (συμμετρία)
IV. $d(x,z) \leq d(x,y) + d(y,z)$ (τριγωνική ανισότητα)

Έστω ότι έχουμε n αντικείμενα, τότε τοποθετούμε τις αποστάσεις τους $d_{jk} = d(x_j, x_k)$ σε έναν πίνακα $D = [d_{jk}]$ ο οποίος θα έχει n γραμμές και n στήλες. Ο πίνακας αυτός ονομάζεται πίνακας αποστάσεων ή πίνακας εγγύτητας των n σημείων ενώ τα διαγώνια στοιχεία του πίνακα ισούται με μηδέν (δηλαδή d_{jk}=d_{kj}). Τα πιο γνωστά μέτρα απόστασης και ομοιότητας που χρησιμοποιούνται όταν έχουμε να κάνουμε με ποσοτικά δεδομένα (συχνότητα εμφάνισης λέξεων κλπ) απεικονίζονται παρακάτω (Νασίκας, 2006; Σταυλιώτης, 2008):

1. **Μετρική Manhattan ή City Block distance (L1 Norm):**

$$d(x_j, x_k) = \sum_{i=1}^{n} |x_{ji} - x_{ki}|$$

Είναι το άθροισμα, δηλαδή, των απόλυτων τιμών των διαφορών ανάμεσα στα δύο διανύσματα των εγγράφων.

2. **Euclidean distance (L2 norm) ή απόσταση του Pearson:**

$$d(x_j, x_k) = \sqrt{\sum_{i=1}^{n} (x_{ji} - x_{ki})^2}$$

Είναι η ευκλείδεια απόσταση και ορίζεται ως η ρίζα του αθροίσματος των τετραγώνων. Όσο μικρότερη είναι η απόσταση μεταξύ των διανυσμάτων (όσο πιο κοντά βρίσκονται δηλαδή μέσα στο διανυσματικό χώρο) τόσο πιο όμοια θεωρούνται τα αντίστοιχα έγγραφα μεταξύ τους. Αποτελεί τη πιο συχνή μορφή απόστασης ονομάζεται και είναι γνωστή περισσότερο για την επίλυση γεωμετρικών προβλημάτων, εξαρτώνται από τη κλίμακα μέτρησης ενώ επηρεάζονται αρκετά από τις έκτροπες παρατηρήσεις (*outliers*). Ικανοποιεί τους τέσσερις βασικούς κανόνες που προαναφέραμε και χρησιμοποιείται ευρέως στην ομαδοποίηση του Text Mining. Χρησιμοποιείται ιδιαίτερα και στην K-means.

3. **Minkowski distance:**

Μία γενίκευση της Ευκλείδειας απόστασης και της απόστασης Manhattan ή City Block είναι η λεγόμενη Απόσταση Minkowski και ο τύπος υπολογισμού της να είναι ο εξής:

$$d(x_j, x_k) = \left[\sum_{i=1}^{n} (x_{ji} - x_{ki})^{\lambda}\right]^{\frac{1}{\lambda}}, \text{ όπου } \lambda \geq 1.$$

Για $\lambda = 1$ η απόσταση Minkowski μας οδηγεί στην απόσταση Manhattan ενώ για $\lambda = 2$ μας οδηγεί στην Ευκλείδεια απόσταση.

4. **Cosine distance:**

$$d(x_j, x_k) = \frac{\sum_{i=1}^{n}(x_{ji} \times x_{ki})}{\sqrt{\sum_{i=1}^{n}(x_{ji})^2} \times \sqrt{\sum_{i=1}^{n}(x_{ki})^2}}$$

Είναι το συνημίτονο της γωνίας μεταξύ των δύο διανυσμάτων και για το λόγο αυτό αποτελεί ένα από τα πιο δημοφιλή κριτήρια ομαδοποίησης. Ισχύει ότι όσο πιο κοντά είναι η τιμή ενός συνημίτονου στην τιμή 1, τόσο μεγαλύτερη είναι η γωνία. Συνεπώς, θεωρούμε ότι όσο πιο κοντά στην τιμή 1 είναι η απόσταση των συνημίτονων τόσο πιο κοντά είναι μεταξύ τους τα διανύσματα και οπότε τόσο ομοιότερα μεταξύ τους τα αντίστοιχα έγγραφα. Τέλος, είναι ανεξάρτητο του μεγέθους ενός κειμένου υπό την έννοια ότι αν ένα κείμενο έχει το ίδιο λεξιλόγιο με ένα άλλο τότε με βάση αυτό το κριτήριο τα δύο κείμενα δείχνουν πανομοιότυπα ανεξάρτητα από το μέγεθος του ενός και του άλλου κειμένου.

5. **Cluster**:

$$d(x_j, x_k) = \frac{\sum_{i=1}^{n}(x_{ji} \times x_{ki})}{\sqrt{\sum_{i=1}^{n} x_{ji}}}$$

Και

$$d(x_k, x_j) = \frac{\sum_{i=1}^{n}(x_{ji} \times x_{ki})}{\sqrt{\sum_{i=1}^{n} x_{ki}}}$$

Αποτελεί ένα ενδιαφέρον μέτρο καθώς δεν ισχύει η συμμετρική ιδιότητα. Δηλαδή, αν επιλέξουμε τυχαία δύο λέξεις τότε η ομοιότητά τους θα εξαρτάται από τη σειρά με την οποία συλλέχθηκαν. Για παράδειγμα, αν χρησιμοποιήσουμε πρώτα τη λέξη μπάσκετ, αυτή φαίνεται να έχει μεγαλύτερη ομοιότητα με την λέξη μπάλα, ενώ αν χρησιμοποιηθεί πρώτα η λέξη μπάλα δεν καταλήγουμε στο ίδιο αποτέλεσμα.

6. **Camberra distance:**

$$d(x_j, x_k) = \sum_{i=1}^{n} \frac{|x_{ji} - x_{ki}|}{|x_{ji}| + |x_{ki}|}$$

Η απόσταση Camberra ορίζεται ως το άθροισμα του πηλίκου των διαφορών ως προς τα αθροίσματα των συντεταγμένων.

7. **Απόσταση max ή Chebyshev distance:**

$$d(x_j, x_k) = \max_{i=1} |x_{ji} - x_{ki}|$$

Η απόσταση Chebyshev, σε αντίθεση με την L1 norm, ορίζεται ως τη μέγιστη απόσταση ανάμεσα σε συντεταγμένες και όχι σαν το άθροισμα αυτών. Συγκεκριμένα, θεωρεί δύο παρατηρήσεις διαφορετικές εάν διαφέρουν τουλάχιστον σε μια μεταβλητή.

Αξίζει να σημειωθεί ότι η επιλογή του μέτρου ομοιότητας εξαρτάται σημαντικά από τη φύση των δεδομένων. Εάν για παράδειγμα μετράμε τα βάρη των όρων με βάση το *TF-IDF* χωρίς να λαμβάνουμε υπόψη το μέγεθος του αρχείου, και χρησιμοποιήσουμε την Ευκλείδεια απόσταση ως μέτρο ομοιότητας, τότε το ότι αρχεία μεγαλύτερου μήκους τείνουν να

εμφανίζουν περισσότερους όρους και μεγαλύτερες συχνότητες εμφάνισης των όρων, θα αποτελεί ένα είδος θορύβου για την εύρεση της ομοιότητας μεταξύ των αρχείων. Σε αυτή την περίπτωση ένα μέτρο που προτείνεται είναι η ομοιότητα βάσει συνημίτονου (*cosine distance*) καθώς με τον υπολογισμό του συνημίτονου επιτελείται κανονικοποίηση ως προς το μήκος των κειμένων (*document length normalization*). Συμπληρωματικά, ο υπολογισμός απόστασης μεταξύ διατάξιμων μεταβλητών εναπόκειται στους παραπάνω τύπους, καθώς αντιμετωπίζονται ως συνεχείς, όμως προκειμένου να έχουμε σωστό αποτέλεσμα απαιτείται όλες οι μεταβλητές να υπολογίζονται στην ίδια κλίμακα.

Έστω ότι x_{ji} είναι η τιμή της *i*-στης δίτιμης μεταβλητής (με *k=1, ..., p*) και ισούται με τη μονάδα ή με το μηδέν ανάλογα από το αν το *j*-στο υποκείμενο (με *j=1, ..., n*) εμφανίζει ή όχι το υπό μελέτη χαρακτηριστικό. Συγκρίνοντας τα ζεύγη παρατηρήσεων (εγγράφων ή όρων) έχουμε (Νασίκας, 2006; Σταυλιώτης, 2008):

$$\left(x_{ji} - x_{ki}\right)^2 = \begin{cases} 0, & \alpha\nu\ x_{ji} = x_{ki} = 0\ \acute{\eta}\ x_{ji} - x_{ki} = 1 \\ 1, & \alpha\nu\ x_{ji} \neq x_{ki} \end{cases}$$

Στη περίπτωση αυτή, η συμφωνία αντιμετωπίζεται με τον ίδιο τρόπο, είτε πρόκειται για συμφωνία στο 'όχι' είτε στο 'ναι. Θεωρώντας ότι έχουμε ένα ζεύγος παρατηρήσεων $x_j = \left(x_{j1}, x_{j2}, \dots, x_{jp}\right)$ και $x_\kappa = \left(x_{k1}, x_{k2,} \dots, x_{kp}\right)$. Τότε το μέτρο ομοιότητας, έστω $s_{jk} = s(x_j, x_k)$, είναι ένας πραγματικός αριθμός έτσι ώστε να ισχύουν οι παρακάτω ιδιότητες (Κούτρας, 2007):

1. $s_{jk} \geq 0\ \forall\ k, j$ και $k = j \leftrightarrow s_{kj} = 1$
2. $s_{jk} \leq 1$
3. $s_{jk} = s_{kj}$ (συμμετρική ιδιότητα)

Βάσει των παραπάνω, είμαστε πλέον σε θέση να καθορίσουμε μερικούς συντελεστές ομοιότητας (*similarity coefficients*) που επιτρέπουν το διαφορετικό χειρισμό της συμφωνίας στο 0 και στο 1. Πριν όμως, είναι απαραίτητη η παράθεση ενός πίνακα πλήθους ομοιοτήτων - ανομοιοτήτων για δύο υποκείμενα *j* και *k*, με *p* συνολικό πλήθος μεταβλητών.

	Υποκείμενο k		
Υποκείμενο j	**1**	**0**	***Sum***
1	*a*	*b*	*a+b*
0	*c*	*d*	*c+d*
Sum	*a+c*	*b+d*	*a+b+c+d=p*

ΠΙΝΑΚΑΣ 2-1: ΠΙΝΑΚΑΣ ΠΛΗΘΟΥΣ ΟΜΟΙΟΤΗΤΩΝ-ΑΝΟΜΟΙΟΤΗΤΩΝ ΓΙΑ ΤΑ ΥΠΟΚΕΙΜΕΝΑ ***j*** ΚΑΙ ***k***

Τα βασικότερα μέτρα ομοιότητας παρατίθενται ακολούθως:

1. **Simple Matching**:

$$s\left(x_j, x_k\right) = \frac{a + d}{p}$$

Ορίζεται ως ο αριθμητικός μέσος των όμοιων συντεταγμένων των δύο διανυσμάτων, ή αλλιώς εκφράζει το ποσοστό των όμοιων συντεταγμένων των διανυσμάτων στο σύνολο p. Εκφράζει ίσα βάρη για συμφωνίες '1-1' και '0-0'.

Παρομοίως, ορίζονται και τα υπόλοιπα μέτρα:

2. **Rogers and Tanimoto**:

$$s(x_j, x_k) = \frac{a+d}{a+d+2(b+c)}$$

Σε αυτόν τον τύπο δίνεται διπλάσιο βάρος για τις συμφωνίες '1-1' και '0-0'.

3. **Sokal and Sneath**:

$$s(x_j, x_k) = \frac{2(a+d)}{2(a+d)+(b+c)}$$

Ο τύπος δίνει διπλάσια βαρύτητα στις συμφωνίες '1-1' και '0-0'.

4. **Jaccard Coefficient:**

$$s(x_j, x_k) = \frac{\sum_{i=1}^{n}(x_{ji} \times x_{ki})}{\sum_{i=1}^{n}(x_{ji})^2 + \sum_{i=1}^{n}(x_{ki})^2 - x_{ji} \times x_{ki}} = \frac{a}{a+b+c}$$

Δηλώνει την απουσία συμφωνιών '0-0' από αριθμητή και παρονομαστή.

Αξίζει να σημειωθεί ότι ο τύπος Simple Matching που προαναφέρθηκε γενικεύεται και για ονομαστικές. Δηλαδή, θεωρώντας x και y δύο παρατηρήσεις του δείγματός μας, υπολογίζουμε τη μεταξύ τους απόσταση βάσει του τύπου:

$$s(x, y) = \frac{u}{p}$$

Όπου u το σύνολο των μεταβλητών με την ίδια τιμή για τα x και y και p το σύνολο των μεταβλητών.

Η περίπτωση υπολογισμού της απόστασης μεταξύ κατηγορικών δεδομένων αντιμετωπίζεται διαφορετικά. Καταρχήν, η απόσταση στα δεδομένα αυτά συνεπάγεται τη ποσοτικοποίηση της κλίμακας και την έκφραση της διαφορετικότητας μεταξύ των τιμών μιας κατηγορικής μεταβλητής. Οπότε έχουμε τρείς δυνατές επιλογές (M.Bramer, 2007):

I. Αυθαίρετα θεωρούμε ότι παρατηρήσεις που κατηγοριοποιούνται στην ίδια κλάση απέχουν μεταξύ τους 0, ενώ αν διαφέρουν 1.

II. Η (μερική) διαταξιμότητα μεταξύ των μεταβλητών δίνει τη δυνατότητα ορισμού της απόστασης βάσει της κρίσης μας. Έστω ότι έχουμε μια μεταβλητή με τρεις κατηγορίες 'υψηλό', 'μέτριο' και 'χαμηλό'. Τότε η απόσταση μεταξύ 'υψηλού' – 'μέτριου' ή 'μέτριου'- 'χαμηλού' θα μπορούσε να οριστεί ίση με 0.5 ενώ η απόσταση 'υψηλού'- 'χαμηλού' ίση με 1.

III. Χρήση αντί της απόστασης μεταξύ δύο παρατηρήσεων, το άθροισμα των γνωρισμάτων που δεν είναι κοινά μεταξύ τους. Δηλαδή, η απόσταση δύο όρων θα ισούται με το σύνολο των διαφορετικών γνωρισμάτων τους (Σταυλιώτης, 2008).

ΚΕΦΑΛΑΙΟ 3
ΑΝΑΣΚΟΠΗΣΗ ΣΥΣΤΗΜΑΤΟΣ ΕΞΟΡΥΞΗΣ ΓΝΩΣΗΣ ΚΕΙΜΕΝΙΚΩΝ ΔΕΔΟΜΕΝΩΝ

3.1 ΣΥΣΤΗΜΑ NEWSCATS

Είναι γεγονός ότι η πρόβλεψη των τιμών των μετοχών στο Χρηματιστήριο βασιζόμενη σε συγκροτημένα δεδομένα είναι αρκετά δημοφιλής. Πράγματι, πολυάριθμες δημοσιεύσεις περιγράφουν εφαρμογές εξόρυξης γνώσης χρησιμοποιώντας συγκροτημένα δεδομένα. Ο λόγος που χρησιμοποιείται η δομημένη πληροφορία και όχι η αδόμητη προκύπτει από το ότι οι προσδοκίες των επενδυτών πηγάζουν ως ένα βαθμό από την αδόμητη πληροφορία (Mittermayer, 2004).

Δελτία τύπου που αφορούν στοιχεία απολαβών, εξαγορών, εκποιήσεων επιχειρήσεων κλπ, επηρεάζουν το 99% της αγοράς. Αυτό συμβαίνει γιατί οι επενδυτές μέσω αυτών αποκτούν χρήσιμές πληροφορίες με αποτέλεσμα η επόμενη δράση τους να είναι απόρροια του περιεχομένου των δελτίων. Για του λόγου το αληθές, δελτία αρνητικού περιεχομένου σημαίνει ότι οι επενδυτές θα πουλήσουν τις μετοχές τους και επεκτατικά θα μειωθεί η τιμή των μετοχών. Ομοίως, δελτία θετικού περιεχομένου θα οδηγήσει τους επενδυτές σε αγορά μετοχών και κατ' επέκταση η τιμή των μετοχών θα αυξηθεί.

Το κατάλληλο μοντέλο για την περιγραφή της κίνησης των τιμών της μετοχής είναι ο Τυχαίος Περίπατος. Σύμφωνα με την εν λόγω θεωρία, οι μελλοντικές τιμές της μετοχής δεν εξαρτάται από την τιμή που είχε στο παρελθόν. Δηλαδή, δεν υπάρχει αυτοσυσχέτιση και η πιθανότητα να αυξηθεί η τιμή ή να μειωθεί είναι η ίδια (Λισγάρα, 2011). Ο Marc-André Mittermayer, στη μελέτη του θεωρεί ότι οι πιθανότητες των μονοπατιών στο μοντέλο του Τυχαίου Περίπατου δεν είναι ίδιες αμέσως μετά την δημοσίευση του δελτίου και ότι η συμπεριφορά των τιμών επηρεάζεται και από άλλους παράγοντες (προσδοκίες επενδυτών).

Δεδομένης της θεωρίας αυτής, Marc-André Mittermayer εισήγαγε ένα σύστημα, το NewsCATS, για την πρόβλεψη των τάσεων των τιμών των μετοχών για το χρονικό διάστημα αμέσως μετά τη δημοσίευση ενός δελτίου τύπου (χρηματοοικονομικού περιεχομένου). Το σύστημά του αποτελείται από τρεις συνιστώσες:

- Η πρώτη συνιστώσα αφορά την ανάκτηση σχετικών πληροφοριών από τα δελτία τύπου εφαρμόζοντας τεχνικές προ-επεξεργασίας κειμένου.
- Η δεύτερη, αφορά την ταξινόμηση των πληροφοριών που πάρθηκαν σε προκαθορισμένες κατηγορίες.
- Τέλος, η τρίτη συνιστώσα εμπεριέχει κατάλληλες στρατηγικές που εφαρμόζονται ανάλογα από τη κατηγορία στην οποία έχουν ταξινομηθεί οι πληροφορίες που διατέθηκαν.

Ο Mittermayer υποστηρίζει ότι τα αποτελέσματα του συστήματός του είναι σε θέση να παρέχουν τη πρόσθετη πληροφορία που απαιτείται προκειμένου για πρόβλεψη της τάσης των χρηματιστηριακών δεικτών, χωρίς βέβαια αυτό να σημαίνει ότι δε χρειάζεται περεταίρω μελέτη του προβλήματος.

Σκοπός της ενότητας αυτής είναι η παράθεση της μελέτης του Mittermayer, όσον αφορά τη χρήση εξόρυξης κειμένου προκειμένου για διενέργεια πρόβλεψης των τιμών των μετοχών.

3.2 ΑΝΑΛΥΤΙΚΗ ΠΕΡΙΓΡΑΦΗ ΤΩΝ ΣΥΝΙΣΤΩΣΩΝ ΤΟΥ NEWSCATS

Το σύστημα του Mittermayer έχει τις ακόλουθες ιδιότητες:

1. Δημιουργήθηκε προκειμένου να είναι σε θέση να αναλύει και να κατηγοριοποιεί αυτόματα τα δελτία τύπου, ανάλογα το περιεχόμενό τους, και να εξάγει συστάσεις χρηματιστηριακών συναλλαγών βάσει αυτών.
2. Διαφέρει σημαντικά από αντίστοιχα προηγούμενα που έχουν δημιουργηθεί ως προς τον τρόπο που συλλέγονται τα παραδείγματα μάθησης και ως προς την απόφαση της πιο ευνοϊκής στρατηγικής χρηματιστηριακών συναλλαγών.
3. Δοκιμάζεται σε δελτία τύπου και δεδομένα τιμών από το 2002 και τα αποτελέσματα δείχνουν ότι οδηγεί τους ενδιαφερόμενους σε καλύτερα αποτελέσματα από ότι θα είχαν αν ακολουθούσαν τυχαία μια στρατηγική.

3.2.1 ΠΡΟ-ΕΠΕΞΕΡΓΑΣΙΑ ΚΑΙ ΑΥΤΟΜΑΤΗ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ ΚΕΙΜΕΝΩΝ

Οι περισσότεροι αλγόριθμοι (ATC) που χρησιμοποιούνται σε αυτή τη κατηγορία είναι γνωστοί από τις εφαρμογές Data Mining. Το γεγονός ότι χρησιμοποιούνται αλγόριθμοι σημαίνει ότι τα δεδομένα που χρησιμοποιούνται είναι αριθμητικά. Επομένως, πρωταρχικό βήμα είναι η μετατροπή οποιουδήποτε δεδομένου σε αριθμητικό.

Όπως, προαναφέρθηκε στη προηγούμενη ενότητα η προ-επεξεργασία κειμένου ακολουθεί κάποια βήματα. Στη παρούσα φάση, λοιπόν, θα αναλυθεί ο τρόπος με τον οποίο λειτουργεί το NewsCATS σε κάθε βήμα:

1) **Εξαγωγή χαρακτηριστικών:** Σκοπός του συστήματος στη φάση αυτή είναι η δημιουργία ενός λεξικού το οποίο πέρα από λέξεις περιλαμβάνει και φράσεις (χαρακτηριστικά) που περιγράφουν επαρκώς τη συλλογή των εγγράφων που έχουν συλλεχθεί. Όσον αφορά τα λεξικά ισχύουν τα εξής:
   - Μπορεί να δημιουργηθούν τοπικά λεξικά, δηλαδή λεξικά για κάθε κατηγορία ή ένα ενιαίο που θα ανταποκρίνεται σε όλες τις επιμέρους κατηγορίες.
   - Προκειμένου να δημιουργηθεί το λεξικό, οι υποψήφιες λέξεις και φράσεις συγκρίνονται με τερματικές λέξεις προκειμένου να αποφευχθεί ο θόρυβος (προθέσεις, αριθμοί κλπ).
   - Επιπλέον, σε λέξεις που διαφέρουν στη κατάληξη ή στο πρόθεμα (δηλαδή έχουν κοινό στέλεχος), εφαρμόζονται τεχνικές αποκοπής ώστε να αντιμετωπίζονται σαν ενιαία χαρακτηριστικά.
2) **Επιλογή χαρακτηριστικών:** Στόχο εδώ αποτελεί η εξάλειψη χαρακτηριστικών που παρέχουν λίγες ή λιγότερο σημαντικές πληροφορίες. Η εξάλειψη πραγματοποιείται χρησιμοποιώντας τους δείκτες *TF* (όρος συχνότητας), *IDF* (αντίστροφη συχνότητα) και $TF \times IDF$ (το γινόμενο αυτών).

   Όταν χρησιμοποιείται ο όρος συχνότητας (*TF*) διαπιστώνουμε ότι στο σύνολο εγγράφων είναι συχνότεροι οι σημαντικοί όροι. Όταν, χρησιμοποιείται η αντίστροφη συχνότητα (*IDF*) διαπιστώνουμε ότι περισσότεροι είναι οι μη-σημαντικοί ενώ όταν εφαρμόζεται ο τρίτο δείκτης σημαίνει ότι μια μεταβλητή (το γινόμενο των δύο πρώτων δεικτών) καθορίζει τη σημαντικότητα των εγγράφων. Σε κάθε περίπτωση, στο τέλος της διαδικασίας παραμένουν μόνο εκείνα τα έγγραφα που έχουν το μεγαλύτερο score, δηλαδή περιλαμβάνουν τις πιο σημαντικές λέξεις.

   Έχει αποδειχθεί, ότι η ύπαρξη αυτών των δεικτών, και ειδικότερα ο δείκτης *TF*, είναι αρκετά αποτελεσματικές ως προς την επιλογή των κατάλληλων χαρακτηριστικών.

3) **Αναπαράσταση εγγράφου:** Σκοπός της τρίτης φάσης είναι η αναπαράσταση των εγγράφων ως ένα διάνυσμα *n* χαρακτηριστικών, όπου το *n* εκφράζει το πλήθος των χαρακτηριστικών που παρέμειναν μετά την δεύτερη φάση. Συνεπώς, το σύνολο πληροφοριών που προέκυψε από τη διαδικασία μπορεί να θεωρηθεί ως ένας πίνακας *F*, $m \times n$ χαρακτηριστικών, όπου *m* το πλήθος των εγγράφων. Το στοιχείο f_{ij} αντιπροσωπεύει τη συχνότητα του χαρακτηριστικού *j* στο έγγραφο *i*. Ενώ τα μέτρα συχνότητας είναι ίδια με της προηγούμενης φάσης, η διαφορά είναι ότι κάθε συχνότητα υπολογίζεται ανά έγγραφο.

Συχνά, τα μέτρα συχνότητας ακολουθούν τη δυαδική αναπαράσταση. Συγκεκριμένα, τα μέτρα περιορίζονται στις τιμές {0,1} όπου κάθε τιμή υποδηλώνει εάν το εκάστοτε χαρακτηριστικό ανήκει ή όχι στο έγγραφο. Στο τέλος της διαδικασίας, το διάνυσμα κανονικοποιείται βάσει συνημιτόνου.

Προκειμένου να μειωθεί το μέγεθος του πίνακα *F* έχουν αναπτυχθεί τεχνικές που βασίζονται κυρίως στο ότι πολλά χαρακτηριστικά μπορούν να θεωρηθούν συνώνυμα και επομένως να εξαλειφθούν. Οι κύριες προσεγγίσεις των ταξινομητών ATC αφορούν τη χρήση δέντρων απόφασης, νευρωνικών δικτύων, SVM κ.ά. Το σύστημα NewsCATS βασίζεται στον ταξινομητή SVM (*Support Vector Machine*). Ο ταξινομητής SVM έχει τα εξής χαρακτηριστικά:

- Περιλαμβάνει θετικά αλλά και αρνητικά έγγραφα (δηλαδή έγγραφα θετικού και αρνητικού περιεχομένου), τα οποία διαχωρίζει στον *n*-διάστατο χώρο αναζητώντας την καταλληλότερη απόφαση, δηλαδή χρησιμοποιώντας φορείς υποστήριξης.
- Η απόφαση διαχωρισμού ή με άλλα λόγια η κατηγοριοποίηση των εγγράφων, γίνεται αρκετά γρήγορα καθώς μόνο μια 'προϊόντος' ανά έγγραφο απαιτείται να υπολογιστεί.

Δεδομένου ότι τα μέτρα ομοιότητας αφορούν κάθε κατηγορία ξεχωριστά είναι πιθανό ένα έγγραφο να ανήκει σε πολλές κατηγορίες, γεγονός που δυσκολεύει τη διεργασία. Εντούτοις, ο ταξινομητής SVM είναι ο καταλληλότερος ταξινομητής για το NewsCATS.

3.2.2 ΔΙΕΝΕΡΓΕΙΑ ΠΡΟΒΛΕΨΗΣ ΑΠΟ ΑΔΟΜΗΤΑ ΔΕΔΟΜΕΝΑ

Ο Wüthrich και άλλοι, το 1998, ανέλυσαν πηγές χρηματοοικονομικών ιστοσελίδων προκειμένου, χρησιμοποιώντας τεχνικές Text Mining προκειμένου να διαπιστώσουν αν η τιμή μιας συγκεκριμένης μετοχής του Χρηματιστηρίου του Hong Kong θα ανέβαινε (>0.5%), θα κατέβαινε (<0.5%) ή θα παρέμενε σταθερή. Τα αποτελέσματα έδειξαν ότι επετεύχθη ακρίβεια κατά την εφαρμογή των τεχνικών αυτών της τάξης του 46%, εν συγκρίσει με τα αποτελέσματα επιτυχίας που θα παίρναμε αν χρησιμοποιούνταν ένας τυχαίος δείκτης πρόβλεψης όπου το ποσοστό επιτυχίας θα ήταν 33%.

Οι τεχνικές που ακολουθήθηκαν συνοψίζονται ακολούθως:

i. Η πρώτη τεχνική είχε ως βασικό στοιχείο τη χρήση μιας a priori κυρίαρχης γνώσης. Συγκεκριμένα, χρησιμοποιήθηκε ένα λεξικό που περιλάμβανε 392 λέξεις, όπου κάθε μια θεωρούνταν ως ευρέως χρησιμοποιούμενη που έχει την ικανότητα να επηρεάζει τη τιμή της μετοχής προς οποιαδήποτε κατεύθυνση.
ii. Η δεύτερη τεχνική, εν αντιθέσει με την πρώτη, βασιζόταν κυρίως στην ενδοημερήσια τιμή της μετοχής σε χρονικό διάστημα 10 λεπτών. Ειδικότερα, οι ερευνητές μέτρησαν την απόδοση της τεχνικής προσομοιώνοντας την αγορά σκοπεύοντας σε κέρδος της τάξης του 1% και πάνω αμέσως ή σε διάστημα μιας ώρας, με ενδεχόμενες απώλειες. Η στρατηγική που ακολουθήθηκε αποσκοπούσε πρώτον, σε ένα μέσο κέρδος ανά συναλλαγή της τάξης του 0.23% και δεύτερον στο καθορισμό του μέσου όρου αναμονής μέχρι να επηρεαστεί (θετικά ή αρνητικά) η

τιμή αναμονής. Τα ευρήματα έδειξαν ότι ο μέσος όρος αναμονής ήταν τα 20 λεπτά (χωρίς να επιβεβαιωθεί πλήρως το αποτέλεσμα αυτό).

3.2.3 ΔΟΜΗ ΚΑΙ ΕΚΤΕΛΕΣΗ ΤΟΥ NEWSCATS

Προκειμένου το NewsCATS να εκπληρώσει τους σκοπούς για τους οποίους προοριζόταν χρειάστηκε να περιλαμβάνει τρεις μηχανές:

- Τη Μηχανή Προ-επεξεργασίας Κειμένου

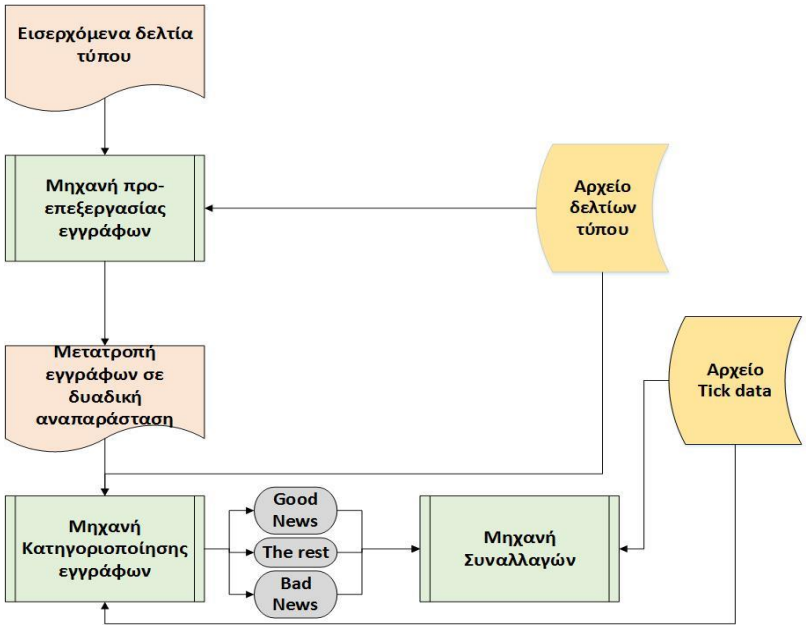
Η Μηχανή αυτή λειτουργεί χρησιμοποιώντας γλώσσα JAVA. Κατά την φάση εξαγωγής χαρακτηριστικών, λοιπόν, η μηχανή καλείται να επιλέξει έναν από τους διάφορους απορρέοντες αλγόριθμους (Porter, Peak & Platen) για την αφαίρεση των λέξεων τερματισμού, ενώ για τη φάση επιλογής χαρακτηριστικών επιλέγεται ως μέτρο συχνότητας μια δυαδική πράξη ή εναλλακτικά οι μηχανές TF, IDF, $TF \times IDF$. Τέλος, έχει τη δυνατότητα δημιουργίας λεξικού.

- Τη Μηχανή Κατηγοριοποίησης

Η Μηχανή Κατηγοριοποίησης χρησιμοποιεί τον ταξινομητή SVM Light ενώ η εφαρμογή υποδοχής των αρχείων χρησιμοποιεί γλώσσα Visual Basic και εμπεριέχει και την Μηχανή Συναλλαγών.

- Τη Μηχανή Συναλλαγών

Ουσιαστικά, λοιπόν, στη πρώτη φάση αρχειοθετούνται τα δελτία ειδήσεων, στη δεύτερη τα αρχεία κατηγοριοποιούνται σε προκαθορισμένο αριθμό κατηγοριών, όπου κάθε κατηγορία έχει διαφορετικό αντίκτυπο στις τιμές των μετοχών. Τέλος, στη τρίτη φάση η Μηχανή συναλλαγών παράγει σήματα συναλλαγών (ξεχωριστά για κάθε κατηγορία που δημιουργήθηκε), τα οποία εκτελούνται μέσω ενός διαδικτυακού μεσίτη (*broker*) ή άλλου είδους μεσάζοντες. Με το παρακάτω γράφημα δίνεται η δυνατότητα οπτικής παρουσίασης του συστήματος NewsCATS:



ΕΙΚΟΝΑ 3-1: ΣΥΣΤΗΜΑ NEWSCATS (ΠΗΓΗ: MITTERMAYER,2004)

Το NewsCATS δεν εφαρμόζεται σε δελτία τύπου με τα εξής χαρακτηριστικά:

1. Δεν έχουν σύμβολο επιλογής ή έχουν δύο και περισσότερα σύμβολα επιλογής. Ειδικότερα, τα δελτία με δύο ή περισσότερα σύμβολα αποκλείονται διότι είναι αρκετά δαπανηρά για το σύστημα.
2. Δε κάνουν αναφορά στα Χρηματιστήρια όπου ανήκουν οι μετοχές εταιριών. Συγκεκριμένα, η απουσία οποιασδήποτε αναφοράς στο Χρηματιστήριο στο οποίο ανήκουν δε δίνει τις κατάλληλες πληροφορίες στο σύστημα ώστε να συγκεντρώσει (ιστορικά) τις τιμές των μετοχών.
3. Κάνουν αναφορά σε Χρηματιστήρια που δεν είναι αντικείμενο μελέτης.
4. Δεν έχουν κωδικό θέματος. Ουσιαστικά, αποκλείονται δελτία τύπου που αφορούν επιχειρήσεις με τζίρο έως και $5,000 την ημέρα, γιατί οι μετοχές τέτοιων εταιρειών θεωρούνται μη διαπραγματεύσιμες. Επίσης, αποκλείονται δελτία μετοχών που εκδίδονται εκτός των ωρών λειτουργίας του Χρηματιστηρίου ενδιαφέροντος.

Εδώ να σημειωθεί ότι οι περιορισμοί του συστήματος προκύπτουν από τις καταγεγραμμένες ώρες των tick by tick δεδομένων καθώς και από την απαίτηση να αξιοποιούνται δεδομένα που εκδόθηκαν tick by tick 60 λεπτά μετά την δημοσίευση του δελτίου τύπου. Συνεπώς, ο αριθμός των δελτίων τύπου που χρησιμοποιεί ο Mittermayer στο παράδειγμά του είναι συγκεκριμένος (6.602). Η απόδοση των τιμών των μετοχών για 60 λεπτά πριν την έκδοση των 6.602 δελτίων είναι -0.01% κατά μέσο όρο, ενώ μετά την έκδοσή τους 1.41%. Αντίστοιχα, για την τυπική απόκλιση ισχύει ότι είναι 1.41% πριν και 2.67% μετά. Αυτή η σημαντική διαφορά υποδηλώνει ότι ο διαχωρισμός αυτός μας αφήνει με δελτία τύπου που έχουν την ικανότητα να επηρεάσουν τις τιμές των μετοχών, ανεξαρτήτως κατεύθυνσης.

3.2.4 ΡΥΘΜΙΣΕΙΣ ΣΥΣΤΗΜΑΤΟΣ

Για να λειτουργήσει το σύστημα, διαχωρίζουμε τα δελτία τύπου σε τρεις κατηγορίες:

1. Η κατηγορία ’Good News’, όπου περιλαμβάνονται τα δελτία που οδηγούν σε αύξηση των τιμών της μετοχής. Συγκεκριμένα, τα δελτία αυτά οδηγούν σε αύξηση το λιγότερο 3% σε κάποιο χρονικό σημείο εντός του χρονικού διαστήματος 60 λεπτών μετά τη δημοσίευσή τους, ενώ παράλληλα η μέση τιμή αύξησης στα 60 λεπτά είναι το λιγότερο 1% παραπάνω από τη τιμή δημοσίευσης.
2. Η κατηγορία ‘Bad News’, όπου εμπεριέχονται τα δελτία που οδηγούν σε μέγιστη πτώση της τιμής 3% και σε μια μέση τιμή μείωσης 1% από την τιμή που δημοσιεύτηκε το αντίστοιχο άρθρο για κάθε μετοχή.
3. Η κατηγορία ‘No Movers’, όπου ανήκουν τα αρχεία που δεν ανήκουν στις άλλες δύο.

Στη προκειμένη, διαχωρίστηκαν 347 δελτία ως ‘Good’, 357 ως ΄Bad’ και 5.898 ως ‘No Movers’. Παρατηρούμε ότι υπάρχει διαφορά στη συχνότητα μεταξύ των κατηγοριών οπότε, επιλέγονται τυχαία 200 αρχεία από κάθε κατηγορία και συνεχίζεται η διαδικασία εκπαίδευσης της μηχανής κατηγοριοποίησης. Όσον αφορά την τρίτη κατηγορία, η επιλογή γίνεται βασιζόμενη στο ότι χρησιμοποιείται ένα υποσύνολο 1.166 αρχείων (που έχει επιλεγεί τυχαία) και από εκεί προηγούνται αυτά που εμφανίζουν τη χαμηλότερη μέγιστη αλλαγή τιμής στη μετοχή και το μεγαλύτερο πλήθος αλλαγών της τιμής των αντίστοιχων μετοχών μέσα στα 60 λεπτά μετά την δημοσίευση του δελτίου. Επομένως, τα αρχεία που επιλέγονται δεν εμφανίζουν υψηλή μεταβλητότητα στη τιμή τους.

Η προ-επεξεργασία των δελτίων τύπου συνεχίζει ως εξής: Κατά την φάση εξαγωγής χαρακτηριστικών δημιουργούνται τρία τοπικά λεξικά που περιέχουν μόνο λέξεις. Οι εγγραφές στα λεξικά προέρχονται από τον αλγόριθμο Porter, από όπου αφαιρούνται ανούσιες

λέξεις τερματισμού (ετικέτες xml για παράδειγμα) και αριθμοί. Συνεπώς, για τη περιγραφή των όρων $TF \times IDF$ χρησιμοποιούνται οι λέξεις με το περισσότερο νόημα. Στη φάση αναπαράστασης εγγράφου, τέλος, χρησιμοποιείται ένα δυαδικό μέτρο συχνότητας και τα διανύσματα των χαρακτηριστικών που προκύπτουν κανονικοποιούνται βάσει συνημιτόνου για περαιτέρω επεξεργασία.

3.2.5 ΑΠΟΤΕΛΕΣΜΑΤΑ ΕΦΑΡΜΟΓΗΣ ΣΕ ΠΡΑΓΜΑΤΙΚΑ ΔΕΔΟΜΕΝΑ

Τα αποτελέσματα του συστήματος του Mittermayer, είναι αναγκαία η παράθεση ενός πίνακα που αναφέρεται στο παράδειγμα που εκείνος χρησιμοποίησε στη μελέτη του. Ο πίνακας που χρησιμοποιήθηκε φαίνεται παρακάτω:

	Good News (N=147)		***No Movers (N=5,698)***		***Bad News (N=157)***		***Overall (Weighted Recall)***
	Prec.	*Rec.*	*Prec.*	*Rec.*	*Prec.*	*Rec.*	
Avg.	6%	43%	98%	59%	5%	47%	58%
Min.	5%	37%	98%	54%	4%	38%	54%
Max.	7%	50%	98%	61%	5%	54%	60%
StDev.	0%	4%	0%	2%	1%	5%	2%

ΠΙΝΑΚΑΣ 3-1: ΠΕΡΙΓΡΑΦΙΚΑ ΣΤΑΤΙΣΤΙΚΑ ΓΙΑ ΤΑ ΔΕΔΟΜΕΝΑ ΕΚΠΑΙΔΕΥΣΗΣ ΤΟΥ MITTERMAYER (ΠΗΓΗ: MITTERMAYER,2004)

Προκειμένου να κατανοηθεί ο παραπάνω πίνακας είναι απαραίτητη η παράθεση των ορισμών ακρίβεια και ανάκληση:

- Η ακρίβεια είναι ο λόγος του αριθμού των εγγράφων που κατηγοριοποιήθηκαν υπό εξονυχιστικό έλεγχο προς το συνολικό αριθμό των εγγράφων που κατηγοριοποιήθηκαν σε συσχετιζόμενα και μη έγγραφα.
- Η ανάκλαση είναι ο λόγος του αριθμού των σχετικών εγγράφων που έχουν κατηγοριοποιηθεί προς το συνολικό αριθμό σχετικών εγγράφων που έπρεπε να είχαν κατηγοριοποιηθεί.

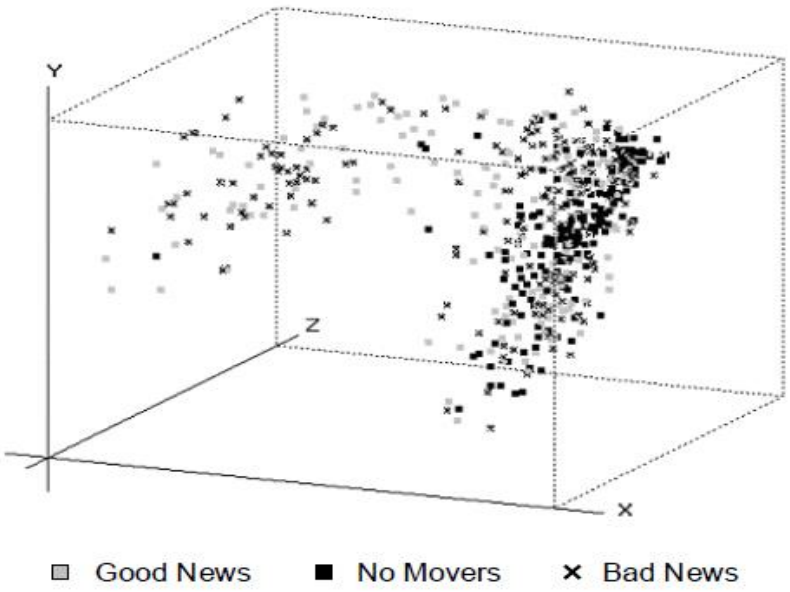
Η ακρίβεια υπολογίζεται από την κατηγοριοποίηση των εναπομείναντα 6002 (147+157+5698) δεδομένων εκπαίδευσης και η διαδικασία κατηγοριοποίησης αυτών συμβαίνει 50 φορές.

Προσαρμόζοντας τις παραπάνω έννοιες στο σύστημα, ο Mittermayer δηλώνει ότι όταν ο αλγόριθμος αδυνατεί να εντοπίσει πρότυπα στα δεδομένα εκπαίδευσης, ο μέσος ανάκλασης για κάθε κατηγορία είναι 33%. Στο παράδειγμα του, όλες οι κατηγορίες έχουν μέσο όρο μικρότερο αυτής της τιμής, ενώ η συνολική ακρίβεια (υπολογίζεται ως σταθμισμένη ανάκλαση) σχεδόν ισούται με την ανάκλαση της κατηγορίας 'No Movers', γιατί η πλειοψηφία των δελτίων ανήκουν στη κατηγορία αυτή.

Ο μέσος όρος ακρίβειας για τις κατηγορίες 'Good News' και 'Bad News' είναι χαμηλά, 6% και 5% αντίστοιχα. Αυτό συμβαίνει γιατί η μετρική ακρίβειας δεν λαμβάνει υπόψη τα λάθη κατά την κατηγοριοποίηση των δελτίων τύπου. Αυτό συνεπάγεται και τη χαμηλή επιλεκτικότητα των κατηγοριών αυτών, της οποίας πιθανή εξήγηση μπορεί να θεωρηθεί το γεγονός ότι τα λεξιλόγια που χρησιμοποιούνται προκειμένου να κατηγοριοποιηθούν τα δελτία τύπου, διαφέρουν αρκετά από το λεξιλόγιο της κατηγορίας 'No Movers' και ελάχιστα μεταξύ

τους. Μια ενδεχόμενη λύση για την αποφυγή αυτού του προβλήματος είναι η χρήση φράσεων και όχι λέξεων, πρόταση που τίθεται υπό έρευνα.

Αντίθετα, τα χαρακτηριστικά ακρίβειας και ανάκλησης της τρίτης κατηγορίας υποδηλώνουν την επιλεκτικότητα αυτής. Παρατηρώντας την παρακάτω εικόνα παρατηρούμε ότι σε ένα χώρο τριών διαστάσεων όπου εμπεριέχονται και οι τρεις κατηγορίες τα έγγραφα της κατηγορίας 'No Movers' συγκεντρώνονται σε μια γωνία εν αντιθέσει με τα υπόλοιπα.



ΕΙΚΟΝΑ 3-2: 3D ΑΠΕΙΚΟΝΙΣΗ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗΣ ΕΓΓΡΑΦΩΝ (ΠΗΓΗ: MITTERMAYER,2004)

Στη συνέχεια, τα αποτελέσματα του πειράματος του Mittermayer διαβιβάζονται στη Μηχανή Συναλλαγών. Αυτή η μηχανή μεταφράζει το αποτέλεσμα κατηγοριοποίησης σε σήματα συναλλαγών τύπου 'Buy Stock', 'Short Stock' και 'Do Nothing'. Τα αποτελέσματα αφορούν τη καλύτερη αντοχή κάθε δελτίου ανά κατηγορία στην ελεγχόμενη περίοδο (*holding*) και υποδηλώνουν ότι είναι καλύτερο να επιλέγεται μεγάλη (μικρή) περίοδος ελέγχου με μεγάλο (μικρό) καθημερινό τζίρο.

Επιστρέφοντας στο παράδειγμα εφαρμογής του NewsCATS, παρακάτω φαίνεται ο πίνακας συναλλαγών όπου συνοψίζονται οι συστάσεις συναλλαγών που δημιουργήθηκαν.

	Buy Recommendations	***Short Recommendations***
Avg.	1,330 (22%)	1,272 (21%)
Min.	1,158 (19%)	997 (17%)
Max.	1,581 (26%)	1,409 (23%)
StDev.	110 (2%)	101 (2%)

ΠΙΝΑΚΑΣ 3-2: ΠΙΝΑΚΑΣ ΣΥΝΑΛΛΑΓΩΝ (ΠΗΓΗ: MITTERMAYER,2004)

Συγκεκριμένα, παρόλο που επισημαίνονται μόνο 147 δελτία τύπου από το υπολειπόμενο σύνολο ως 'Good News' και 157 ως 'Bad News', η Μηχανή Συναλλαγών κατά μέσο όρο προτείνει την αγορά των επικείμενων μετοχών 1330 φορές (ή αλλιώς το 22% των 6002 εναπομείναντα) και την πώληση των αντίστοιχων 1272 φορές (ή αλλιώς το 21% των 6002). Συμπεραίνουμε, λοιπόν, ότι ενώ εκ πρώτης όψεως φαίνεται δελτία της κατηγορίας 'No

Movers' να έχουν κατηγοριοποιηθεί λάθος, η επανεξέταση του παραδείγματος χρησιμοποιώντας δελτίο τύπου με αντίκτυπο μικρότερο από 3% στην επικείμενη τιμή της μετοχής δείχνει ότι αυτό το λάθος είναι τελικά ωφέλιμο. Αυτό συμβαίνει καθώς η μετοχή έχει ήδη αποκτηθεί. Ο συνολικός αριθμός συναλλαγών που προτείνει ο Mittermayer για το σύστημα είναι κατά μέσο 2,602 (δηλαδή το 43%). Ένα σύστημα το οποίο αδυνατεί να εντοπίσει πρότυπα στα δεδομένα εκπαίδευσης θα έδινε περίπου 4,000 (2/3) συστάσεις αγοράς και πώλησης, δεδομένου ότι εκπαίδευση διενεργήθηκε χρησιμοποιώντας ομοιόμορφη κατανομή.

3.2.6 ΠΡΟΣΟΜΟΙΩΣΗ ΧΡΗΜΑΤΙΣΤΗΡΙΟΥ

Ο Mittermayer προκειμένου να αξιολογήσει την απόδοση του συστήματος το οποίο δημιούργησε, εκτέλεσε συστάσεις αγοράς και πώλησης μετοχών πρακτικά, χρησιμοποιώντας δεδομένα από το 2002. Αρχικά, υπέθεσε ότι οι μετοχές μπορούν να αγοραστούν ή να πουληθούν το γρηγορότερο μετά το πέρας 2 λεπτών από τη δημοσίευση του ανάλογου δελτίου τύπου. Η αναμονή 2 λεπτών είναι λογική, καθώς συνυπολογίζεται η διαθεσιμότητα του δελτίου στο ανάλογο site καθώς και η κατηγοριοποίηση αυτού.

Η περίοδος αναμονής μέχρι τη θέση long ή short είναι τα επόμενα 58 λεπτά. Ο ακόλουθος πίνακας απεικονίζει τα περιγραφικά στατιστικά για τις εκτελεσμένες συναλλαγές:

	NewsCATS		***Random Trader***	
	Trades Executed	*Avg. Profit per Trade*	*Trades Executed*	*Avg. Profit per Trade*
Avg.	2,602	0.11%	2,599	0.00%
Min.	2,477	0.03%	2,475	-0.05%
Max.	2,864	0.18%	2,860	0.06%
StDev.	96	0.06%	96	0.03%

ΠΙΝΑΚΑΣ 3-3: ΑΝΑΠΑΡΑΣΤΑΣΗ ΠΕΡΙΓΡΑΦΙΚΩΝ ΣΤΑΤΙΣΤΙΚΩΝ ΤΩΝ ΣΥΝΑΛΛΑΓΩΝ (ΠΗΓΗ: MITTERMAYER,2004)

Ειδικότερα, οι στήλες στη δεξιά πλευρά του πίνακα αντιπροσωπεύουν το αποτέλεσμα του καλύτερου σεναρίου τυχαίας συναλλαγής. Το αποτέλεσμα αυτό, βέβαια, δεν είναι σημαντικά μεγαλύτερο από τη μέση απόδοση της τιμής της μετοχής στα 60 λεπτά μετά τη δημοσίευση του δελτίου, αλλά το κέρδος που επιτυγχάνεται χρησιμοποιώντας το NewsCATS είναι σημαντικά μεγαλύτερο (1%). Η διαπίστωση αυτή υποδεικνύει ότι η πιθανότητα επιτυχίας ενός μοντέλου που ακολουθεί το τυχαίο περίπατο δεν είναι ίδια αμέσως μετά την έκδοση του δελτίου και επίσης, η ασυμμετρία που ενδέχεται να προκληθεί, οφείλεται στο περιεχόμενο του δελτίου.

Η Μηχανή Συναλλαγών προκειμένου να αντιμετωπίσει το πρόβλημα, καθορίζει φράγματα τιμών τα οποία αν ξεπεραστούν δημιουργούν θέσεις αγοράς και πώλησης για τις επικείμενες μετοχές. Συγκεκριμένα, μόλις αποκτηθεί κέρδος (ή απώλεια) της τάξης του *d%* σε χρονικό διάστημα [2,60] λεπτά μετά τη δημοσίευση του δελτίου, καταχωρείται. Εναλλακτικά, αναμονή μέχρι το τέλος της ώρας και καταχώρηση του κέρδους ή της απώλειας αν αυτό είναι απαραίτητο. Άλλοι κανόνες απόφασης για το επίπεδο αυτό είναι ακόμη σε εξέλιξη. Παρακάτω φαίνεται το αποτέλεσμα 50 τρεξιμάτων εισάγοντας διάφορα φράγματα:

Upper	Lower	Num	Rate	Upper	Lower	Num	Rate	Upper	Lower	Num	Rate
Infinite	***Infinite***	**0.11%**	**0.0%**	***Infinite***	**-3.0%**	**0.05%**	**-0.03%**	**3.0%**	***Infinite***	**0.15%**	**0.03%**
3.0%	- 3.0%	0.09%	0.0%	*Infinite*	- 2.0%	- 0.01%	- 0.06%	2.0%	*Infinite*	0.17%	0.06%
2.0%	- 2.0%	0.07%	0.0%	*Infinite*	- 1.0%	- 0.05%	- 0.08%	1.0%	*Infinite*	0.21%	0.07%
1.5%	- 1.5%	0.06%	-0.01%	*Infinite*	- 0.5%	- 0.05%	- 0.07%	1.0%	*Infinite*	0.19%	0.06%
1.0%	- 1.0%	0.05%	-0.01%	3.0%	- 1.0%	- 0.01%	- 0.05%	1.0%	- 3.0%	0.15%	0.04%
0.5%	- 0.5%	0.05%	-0.01%	3.0%	- 0.5%	- 0.02%	- 0.05%	0.5%	- 3.0%	0.15%	0.04%
0.2%	- 0.2%	0.04%	-0.01%	1.0%	- 0.5%	- 0.04%	- 0.01%	0.5%	- 1.0%	0.16%	0.01%

ΠΙΝΑΚΑΣ 3-4: ΚΑΤΑΓΡΑΦΗ 50 ΑΠΟΤΕΛΕΣΜΑΤΩΝ (ΠΗΓΗ: MITTERMAYER,2004)

Τα αποτελέσματα δείχνουν ότι το σύστημα συμπεριφέρεται καλύτερα από τη τυχαία συναλλαγή. Για παράδειγμα, χρησιμοποιώντας συμμετρικά φράγματα, το κέρδος ανά συναλλαγή είναι μεγαλύτερο από 0.05% έως 0.11% (Η υπόθεση άπειρο/άπειρο αφορά τη περίπτωση μηδενικού φράγματος που φαίνεται στον προηγούμενο πίνακα).

Επίσης, το γεγονός ότι το σύστημα οδηγεί σε κέρδος πολύ πιο σύντομα από απώλεια, δείχνει ότι το NewsCATS υπερέχει του τυχαίου. Βασιζόμενοι στα φράγματα, ο κέρδος φτάνει το 0.21% ανά συναλλαγή (το ποσοστό είναι ο μέσος όρος που προέκυψε από τις 50 επαναλήψεις) και έτσι καθίσταται κατά 0.14% υψηλότερο από οποιαδήποτε τυχαία συναλλαγή. Στη περίπτωση του καλύτερου σεναρίου, το κέρδος κυμαίνεται από 0.13% έως 0.28%.

Κατά το μοντέλο που παρουσιάστηκε από τον Lavrenko et al, εφαρμόστηκε παρόμοια προσομοίωση, χρησιμοποιώντας τα ίδια φράγματα, αλλά δε δόθηκε εξήγηση για το λόγο χρήσης των συγκεκριμένων διαστημάτων.

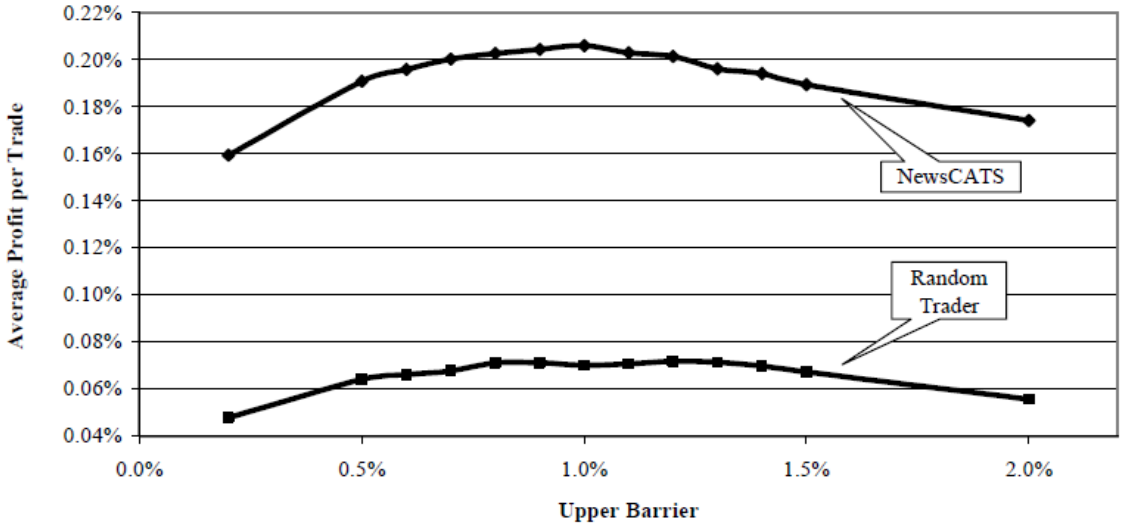
Χρησιμοποιώντας ασύμμετρα φράγματα, ο Mittermayer διαπίστωσε όχι μόνο ότι η Μηχανή Συναλλαγών οδηγεί σε απώλεια νωρίτερα από το κέρδος, αλλά και ότι τόσο το σύστημά του όσο και η τυχαία συναλλαγή εμφανίζουν τα χειρότερα στατιστικά τους.

Οι αξιοσημείωτες διαφορές που εντοπίζονται ανά συναλλαγή λόγω χρήσης διαφορετικού φράγματος μπορούν να επεξηγηθούν από το γεγονός ότι οι ενδοημερήσιες αυξομοιώσεις στη τιμή των μετοχών δεν ακολουθούν απλά το μοντέλο τυχαίου περίπατου, αλλά είναι το αποτέλεσμα της αλληλεπίδρασης ενός τυχαίου περιπάτου και ενός προσωρινού κατωφλιού που παράγεται από περιορισμένες εντολές. Επιπρόσθετα, η χρήση ακατάλληλων φραγμάτων έχει ως αποτέλεσμα τη δημιουργία μικρών κερδών ακόμα και αν οι μετοχές αγοράστηκαν και πουλήθηκαν εντελώς τυχαία.

Η γκρίζα περιοχή του παραπάνω πίνακα, όπου απεικονίζονται οι περιπτώσεις καλύτερου σεναρίου χωρίς κάτω γράφημα, δίνει το έναυσμα για περαιτέρω διερεύνηση του μέσου κέρδους που επιτυγχάνεται χρησιμοποιώντας ποικίλα άνω φράγματα. Έτσι, η προσομοίωση συνεχίζεται χρησιμοποιώντας τιμές για άνω φράγμα που κυμαίνονται από +0.02% έως +2.0%. Η Μηχανή Συναλλαγών δύναται επίσης να ενσωματώσει τα ευρήματα αυτά σε συστάσεις του τύπου:

'Αγόρασε τη μετοχή X και κράτησέ τη μέχρι η τιμή της μετοχής να χτυπήσει το +d% άνω φράγμα'.

Το παρακάτω γράφημα απεικονίζει το μέσο κέρδος ανά συναλλαγή για ποικίλα άνω φράγματα και κανένα κάτω φράγμα. Προφανώς, η διαφορά ανάμεσα στο μέσο κέρδος ανά συναλλαγή του συστήματος και του μέσου κέρδους που επιτυγχάνεται από τυχαία συναλλαγή είναι διαρκής και στατιστικά σημαντική. Το μεγαλύτερο κέρδος, δε, ανά συναλλαγή, χρησιμοποιώντας άνω φράγμα έχει οριστεί στο +0.9%, αλλά δε μπορεί να επιβεβαιωθεί ότι φράγματα κοντά στο +1% οδηγούν σε χαμηλότερα κέρδη.



ΕΙΚΟΝΑ 3-3: ΑΠΕΙΚΟΝΙΣΗ ΓΡΑΦΗΜΑΤΟΣ ΣΥΓΚΡΙΣΗΣ NEWSCATS ΜΕ RANDOM TRADER (ΠΗΓΗ: MITTERMAYER,2004)

Επιπρόσθετα, το σύστημα είναι σε θέση να αποφέρει περισσότερα κέρδη λαμβάνοντας υπόψη το κόστος συναλλαγών. Προς κατανόηση της παραπάνω πρότασης ο Mittermayer, υποθέτει ότι το κόστος είναι US $10 για αγορά και US $10 για πώληση μετοχών. Το σύστημα παραμένει στάσιμο (δεν έχει κέρδος ή απώλεια) αν κάθε συνιστώμενη συναλλαγή εκτελείται με ένα ποσό της τάξης

(US $10+ US $10)/0.21%= US $9,524

Και παραμένει τόσο μέχρι να αποκτηθεί +0.9% ή και παραπάνω. Το ποσό αυτό δε θεωρείται εμπόδιο για τη σωστή λειτουργία του NewsCATS δεδομένου ότι το πείραμα επικεντρώνεται στη προσομοίωση του Χρηματιστηρίου όπου ο ημερήσιος κύκλος συναλλαγών είναι τουλάχιστον US 5,000,000, αγοράς ή πώλησης μετοχών.

3.2.7 ΠΡΟΟΠΤΙΚΕΣ ΕΞΕΛΙΞΗΣ

Το σύστημα NewsCATS δημιουργήθηκε προκειμένου αυτόματα να αναλύει και να κατηγοριοποιεί τα δελτία τύπου και βάσει αυτής της ανάλυσης να προτείνει χρηματιστηριακές συναλλαγές. Αυτό που το διαφοροποιεί από τα προηγούμενα συστήματα είναι ο τρόπος με τον οποίο επιλέγονται τα δεδομένα εκπαίδευσης καθώς και ο τρόπος με τον οποίο καταρτίζονται οι χρηματιστηριακές συστάσεις.

Όπως ήδη αναφέρθηκε, ξεπερνά σημαντικά σε επιτυχία αποφάσεις που επιλέχθηκαν τυχαία μετά τη δημοσίευση δελτίων, εντούτοις απαιτεί βελτίωση. Συγκεκριμένα, είναι αναγκαία η ενίσχυση στη Μηχανή Κατηγοριοποίησης που παρουσιάζει αρκετά μειωμένη επιλεκτικότητα στις κατηγορίες ‘Good News’ και ‘Bad News’ και επιτυγχάνεται

διαχωρίζοντας τα αρχεία σε ‘Movers’ και ‘No Movers’ και στη συνέχεια τη κατηγορία ‘Movers’ σε ‘Good News’ και ‘Bad News’.

Επιπλέον, η έκβαση της διαδικασίας κατηγοριοποίησης εξαρτάται σε μεγάλο βαθμό από τον πίνακα χαρακτηριστικών που παράγεται από τη Μηχανή Προ-επεξεργασίας Εγγράφου. Συνεπώς, μια δεύτερη σκέψη βελτίωσης μπορεί να θεωρηθεί η εφαρμογή μιας προηγούμενης (*a priori*) γνώσης. Ουσιαστικά, η βελτίωση αφορά τις φάσεις εξαγωγής χαρακτηριστικών και επιλογής χαρακτηριστικών που θα περιλαμβάνουν λέξεις ή φράσεις που θεωρούνται ευρέως χρησιμοποιούμενοι όροι, ικανοί να επηρεάσουν τις τιμές των μετοχών και θα ορίζονται από ειδικούς. Ο καθορισμός ενός τόσο περιορισμένου λεξικού ίσως μειώσει την ευελιξία συστημάτων όπως το NewsCATS, το οποίο επεξεργάζεται τη βασική γνώση και επομένως είναι ικανό να λάβει υπόψη του αλλαγές στο λεξιλόγιο.

ΚΕΦΑΛΑΙΟ 4
ΥΒΡΙΔΙΚΗ ΠΡΟΒΛΕΨΗ ΧΡΗΜΑΤΙΣΤΗΡΙΑΚΩΝ ΔΕΔΟΜΕΝΩΝ

Βασιζόμενοι στο σύστημα που προτάθηκε από τον Mittermayer, σκοπός της παρούσας εργασίας είναι να παρουσιαστεί το ανάλογο σύστημα, το οποίο θα αναλύει και μετέπειτα θα

κατηγοριοποιεί άρθρα/δελτία τύπου ούτως ώστε να προτείνονται οι κατάλληλες χρηματιστηριακές συστάσεις.

Θα χρησιμοποιηθούν δεδομένα εκπαίδευσης που αφορούν το σύνολο τραπεζικών μετοχών που συμπεριλαμβάνονται στο Χρηματιστήριο Αθηνών και ορίζονται σε ενεργή κατάσταση. Τα δεδομένα εκπαίδευσης που χρησιμοποιούνται πάρθηκαν διαδικτυακά (ΝΑΥΤΕΜΠΟΡΙΚΗ).

4.1. ΔΙΑΔΙΚΑΣΙΑ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗΣ ΔΕΔΟΜΕΝΩΝ ΕΚΠΑΙΔΕΥΣΗΣ

Η κατηγοριοποίηση των δεδομένων εκπαίδευσης που θα εισάγουμε στο σύστημά μας, αφορά το διαχωρισμό τους σε τρεις κατηγορίες δεδομένων, τα αριθμητικά, τα κειμενικά και τα υβριδικά. Η προσαρμογή των δεδομένων σε κάθε μια από τις προαναφερθέντες απαιτεί την εισαγωγή κανόνων, οι οποίοι παρουσιάζονται στις υποενότητες που ακολουθούν.

Αξίζει να σημειωθεί ότι δεδομένου ότι χρειαζόμαστε δεδομένα ενός έτους, χρησιμοποιήθηκαν δεδομένα για το έτος 2014 ενώ το αποτέλεσμα κάθε απόφασης προέρχεται ως συνισταμένη του περιεχομένου τους, του γραφήματος ημερήσιας κίνησης για κάθε τραπεζική μετοχή και του ημερήσιου όγκου συναλλαγών.

Από τα γραφήματα, παρατηρώντας τα *candlesticks*, δηλαδή γραφικές απεικονίσεις κίνησης των τιμών μπορούμε να εξάγουμε σημαντικές πληροφορίες για την υποκείμενη μετοχή, δηλαδή τιμή ανοίγματος, κλεισίματος, τη χαμηλότερη τιμή κατά τη διάρκεια της ημέρας καθώς και την υψηλότερη. Εκτός από τις πληροφορίες αυτές, τα *candlesticks* χρησιμοποιούνται για να συμπεράνουμε αν οι αγοραστές ή οι πωλητές επικράτησαν της συγκεκριμένη χρονική περίοδο, καθώς και για το αν επικράτησαν εύκολα ή δύσκολα. Τέλος, μεμονωμένα ή συνδυαστικά (2-5 *candlesticks*) δίνουν τη δυνατότητα να αντιληφθούμε με υψηλές πιθανότητες επιτυχίας ποια κατεύθυνση θα ακολουθήσει η μετοχή, όμως δεν ενδείκνυται ως μοναδικό στοιχείο λήψης αποφάσεων (Positron Investments).

Ο λόγος για τον οποίο λήφθηκε υπόψη ο όγκος είναι γιατί αποτελεί σημαντικό κριτήριο λήψης αποφάσεων. Ειδικότερα, όταν ένα *candlestick* συνοδεύεται από μεγάλο όγκο τιμών, τότε είναι πολύ πιθανό ότι η τιμή βραχυχρόνια θα ακολουθήσει την ίδια τάση. Ακόμη, αν μια μετοχή ακολουθεί ανοδική τάση με υψηλό όγκο και οι καθοδικές διορθώσεις γίνονται με χαμηλό όγκο, τότε βάσει πιθανοτήτων θα συνεχίσει να έχει ανοδική τάση. Ομοίως, όταν μια μετοχή που σε καθοδική τάση κάνει υψηλό όγκο και αντίστοιχα στις ανοδικές διορθώσεις χαμηλό, θα συνεχίσει να κινείται καθοδικά. Όμως, δε μπορεί να χρησιμοποιηθεί σαν βασικός κριτήριο λήψης αποφάσεων καθώς είναι σχετικός και εξαρτάται από το μέσο όρο όγκου των προηγούμενων ημερών (Positron Investments). Τέλος, τα δελτία που συλλέχθηκαν είναι εκείνα που κατατίθενται σε sites (ΝΑΥΤΕΜΠΟΡΙΚΗ) και διατίθενται ελεύθερα στους άμεσα ενδιαφερόμενους, δηλαδή στους επενδυτές.

Το γεγονός ότι δεν υπήρχε η δυνατότητα ελέγχου των *intraday* τιμών για κάθε μετοχή, υποδηλώνει την ύπαρξη μερικής αβεβαιότητας για την ορθή κατηγοριοποίηση των δεδομένων. Αξίζει να σημειωθεί ότι η ύπαρξη αβεβαιότητας είναι ο βασικός λόγος για τον οποίο διαχωρίστηκαν τα δεδομένα εκπαίδευσης που θα χρησιμοποιήσουμε και για τον οποίο εφαρμόζουμε τους ανάλογους αλγόριθμους κατηγοριοποίησης και συσταδοποίησης ξεχωριστά. Απώτερος σκοπός είναι πέρα της εύρεσης του καταλληλότερου αλγόριθμου, η εύρεση και των δεδομένων που οδηγούν σε ακριβέστερα αποτελέσματα.

4.1.1. ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ ΑΡΙΘΜΗΤΙΚΩΝ ΔΕΔΟΜΕΝΩΝ

Η κατηγοριοποίηση αριθμητικών δεδομένων έγινε χρησιμοποιώντας τις τιμές κλεισίματος των τραπεζικών μετοχών για το έτος 2014 καθώς και τις τιμές του όγκου για τη κάθε μια από αυτές (δύο μεταβλητές εισόδου). Τονίζουμε ότι τα δεδομένα εκπαίδευσης αφορούν **μόνο** τις ημέρες κατά τις οποίες εκδόθηκε δελτίο τύπου ή άρθρο ενώ ο διαχωρισμός των δεδομένων αφορά τη δημιουργία δύο κατηγοριών ακολουθώντας το κανόνα:

1. Όταν η τιμή κλεισίματος είναι μεγαλύτερη από τη προηγούμενη και ο όγκος παρουσιάζει αυξητική τάση, τότε ορίζουμε τη κατηγορία 'Good'. Ακόμη, όταν η τιμή κλεισίματος είναι μικρότερη από τη προηγούμενη και ο όγκος παρουσιάζει πτωτική τάση, τότε ορίζουμε την ίδια κατηγορία.
2. Όταν η τάση της τιμής κλεισίματος δε συνάδει με τη τάση του όγκου, ορίζουμε τη κατηγορία 'Bad'. Δηλαδή, αν η τιμή κλεισίματος παρουσιάζει ανοδική τάση και ο όγκος καθοδική και αν η τιμή κλεισίματος παρουσιάζει καθοδική τάση και ο όγκος ανοδική, η κατηγορία είναι 'Bad'.

Παρόλο που στη περίπτωση 'Good' δε μας 'συμφέρει' η μείωση της τιμής κλεισίματος, ορίζουμε τη κατηγορία 'Good' καθώς μας ενημερώνει για τη κίνηση της μετοχής οπότε μπορούμε να αποφανθούμε για τη καταλληλότερη σύσταση που θα προτείναμε στους επενδυτές.

4.1.2. ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ ΚΕΙΜΕΝΙΚΩΝ ΔΕΔΟΜΕΝΩΝ

Για τη κατηγοριοποίηση των κειμενικών δεδομένων βασιστήκαμε, ως επί το πλείστον, στις ρυθμίσεις του Mittermayer. Η κατηγοριοποίηση έγινε λαμβάνοντας υπόψη το δελτίο ή τα δελτία τύπου που εκδόθηκαν ανά ημέρα για το έτος 2014 και τη τιμή της εκάστοτε τραπεζικής μετοχής μόνο για τις ημέρες τις οποίες υπήρχε δημοσίευση αρχειου (δύο μεταβλητές εισόδου). Ακολούθως, παρουσιάζονται τα φίλτρα που ορίστηκαν:

1. Στη κατηγορία 'Good News', ομοίως με το NewsCATS, ανήκουν τα δελτία τύπου που οδηγούν σε μέση αύξηση της τιμής της μετοχής τουλάχιστον 1% από τη τιμή που είχε όταν δημοσιεύτηκε το δελτίο ενώ παράλληλα οδηγούν σε αύξηση της τιμής τουλάχιστον κατά 3% κατά τη διάρκεια 60 λεπτών από τη δημοσίευσή τους.
2. Στη κατηγορία 'Bad News', εμπεριέχονται τα δελτία που οδηγούν σε μέγιστη πτώση της τιμής 3% κατά τη διάρκεια μιας ώρας από τη δημοσίευσή τους και σε μια μέση τιμή μείωσης 1% από την τιμή που είχε η μετοχή όταν δημοσιεύτηκε το άρθρο.
3. Η κατηγορία 'No Movers', όπου ανήκουν τα αρχεία που δεν ανήκουν στις άλλες δύο.

Η κατηγοριοποίηση αφορά κειμενικά δεδομένα εκπαίδευσης που προέκυψαν κατά τις ώρες και ημέρες λειτουργίας του Χρηματιστηρίου Αθηνών, δηλαδή καθημερινές από τις 10^{30} πμ έως τις 5^{30} μμ.

4.1.3. ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ ΥΒΡΙΔΙΚΩΝ ΔΕΔΟΜΕΝΩΝ

Η κατηγοριοποίηση υβριδικών δεδομένων συνεπάγεται τη δημιουργία κανόνων που συνδυάζουν τα δεδομένα που περιγράφηκαν στις παραπάνω υποενότητες, δηλαδή αριθμητικά και κειμενικά. Στόχος είναι να δημιουργηθεί ένας τρόπος κατηγοριοποίησης που θα συνδυάζει τη πληροφορία που μπορεί να παραχθεί από τη κατηγοριοποίηση των αριθμητικών και των κειμενικών δεδομένων, ενώ παράλληλα θα δίνει τη δυνατότητα δημιουργίας ενός

δυναμικού συστήματος πρόβλεψης, το οποίο θα μπορεί προσαρμόζεται και να υποδεικνύει πότε ο χρήστης πρέπει να χρησιμοποιήσει το απότέλεσμα της αριθμητικής κατηγοριοποίησης και πότε της κειμενικής ή πότε το συνδυασμό τους προκειμένου για εγκυρότερες συστάσεις.

Η κατηγοριοποίηση υβριδικών δεδομένων προϋποθέτει περαιτέρω επεξεργασία των δεδομένων. Ειδικότερα, χρησιμοποιούμε τις τιμές κλεισίματος και τη τιμή του όγκου μόνο για τις ημέρες κατά τις οποίες εκδόθηκε δελτίο τύπου για την εκάστοτε μετοχή καθώς και το σύνολο δελτίων τύπου που δημοσιεύτηκαν (τρεις μεταβλητές εισόδου) ενώ δημιουργούμε μια νέα μεταβλητή απόκρισης που προκύπτει από τη μεταβλητή απόκρισης των αριθμητικών δεδομένων και τη μεταβλητή απόκρισης των κειμενικών, βάσει του ακόλουθου κανόνα:

- Στις περιπτώσεις όπου η απόφαση κατηγοριοποίησης των δύο μεταβλητών συμπίπτουν, ορίζουμε τη τιμή 0.
- Στις περιπτώσεις όπου η απόφαση κατηγοριοποίησης των δύο μεταβλητών διαφέρουν, ορίζουμε τη τιμή 1.

Δηλαδή, αν για παράδειγμα η κατηγοριοποίηση βάσει όγκου και τιμής κλεισίματος δίνει ‘Good’ και η κατηγοριοποίηση βάσει δελτίου τύπου δίνει ‘Bad News’ τότε ορίζουμε τη τιμή 1. Αλλιώς ορίζουμε τη 0.

Δεδομένου ότι για τα αριθμητικά δεδομένα είχαμε 2 κατηγορίες (0: ‘Good’, 1: ‘Bad’) και για τα κειμενικά τρεις (‘Good’, ‘Bad’ και ‘No Mover’), θα χρησιμοποιήσουμε 2 επιπλέον μεταβλητές, που ορίζονται με τον εξής τρόπο:

- Όταν η μεταβλητή κειμενικών κατηγοριοποιεί ως ‘No Mover’ και η μεταβλητή αριθμητικών ως ‘Good’, ορίζουμε τη τιμή 2.
- Όταν η μεταβλητή κειμενικών κατηγοριοποιεί ως ‘No Mover’ και η μεταβλητή αριθμητικών ως ‘Bad’, ορίζουμε τη τιμή 3.

Δηλαδή, αν για παράδειγμα η κατηγοριοποίηση αριθμητικών δεδομένων έδωσε ‘Good’ και η καηγοριοποίηση κειμενικών έδωσε ‘No Mover’ τότε η νέα μεταβλητή παίρνει τη τιμή 2. Αν, όμως η κατηγοριοποίηση των αριθμητικών έδινε ‘Bad’ η νέα μεταβλητή ορίζεται με τη τιμή 3.

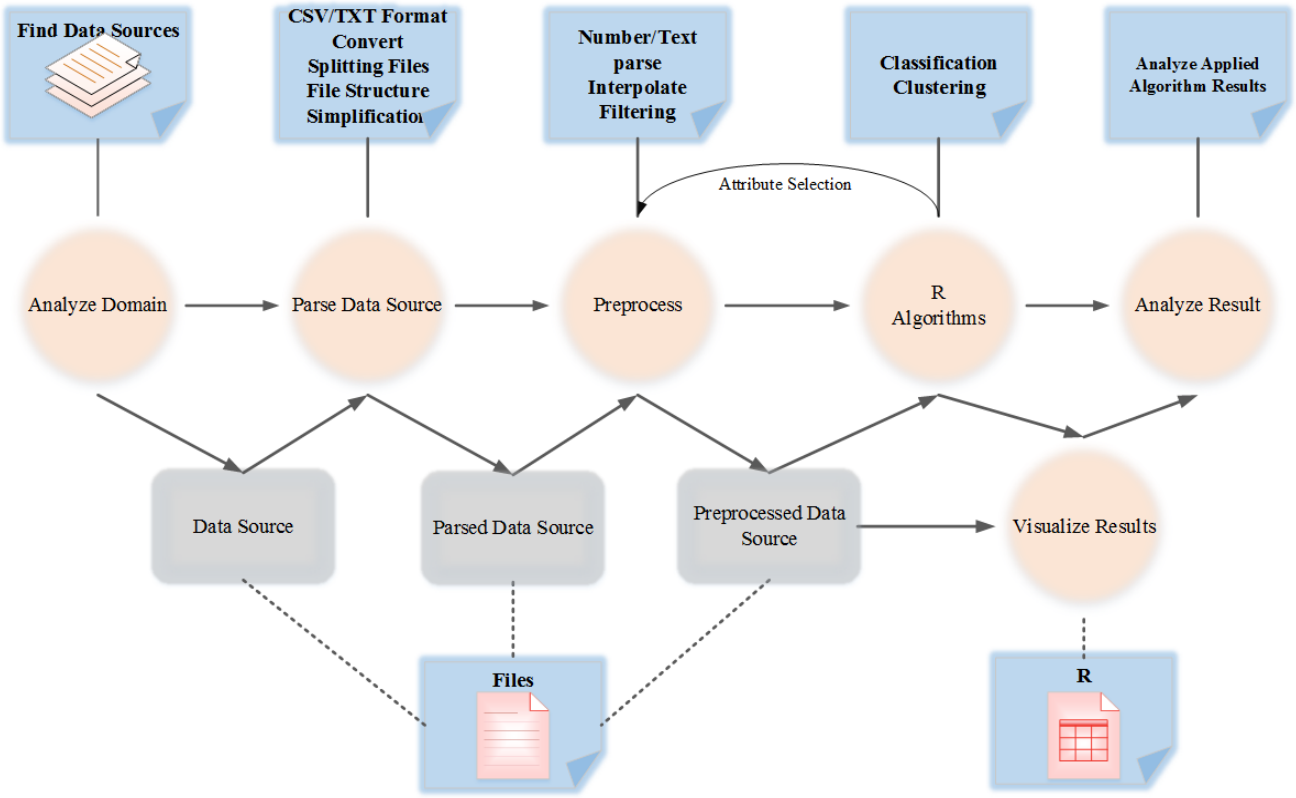
4.2.ΠΕΡΙΓΡΑΦΗ ΤΟΥ ΔΙΑΓΡΑΜΜΑΤΟΣ ΡΟΗΣ ΤΟΥ ΣΥΣΤΗΜΑΤΟΣ ΜΑΣ

Στην ενότητα αυτή θα γίνει μια συνοπτική περιγραφή του συστήματος που θα εφαρμόσουμε για κάθε μια από τις τρεις κατηγορίες δεδομένων που καθορίσαμε προηγουμένως. Συγκεκριμένα, πρώτο μέλημά μας είναι η εύρεση των κατάλληλων πηγών δεδομένων και ανάλυση των *domain* των πηγών μας. Ακολουθεί η επεξεργασία τους προκειμένου να μπορούν να εφαρμοστούν αλγόριθμοι. Ουσιαστικά, απαιτείται η μετατροπή των διαθέσιμων αρχείων σε αρχεία .cave και .txt, ενώ ταυτόχρονα ενδέχεται να χρειαστεί ο διαχωρισμός τους ή η απλοποίησή τους προκειμένου για συνέχιση της διεργασίας.

Επόμενο βήμα είναι η προ-επεξεργασία των δεδομένων, δηλαδή απαιτούνται διεργασίες φιλτραρίσματος, ανάλυσης αριθμών, κοινωνικοποίησης κλπ, προκειμένου να μπορέσουν να χρησιμοποιηθούν οι αλγόριθμοι κατηγοριοποίησης και συστηματοποίησης που επιθυμούμε. Στο σημείο αυτό να σημειώσουμε ότι η διαδικασία αυτή επαναλαμβάνεται για κάθε διαφορετικό σύνολο εκπαίδευσης που χρησιμοποιούμε. Με άλλα λόγια, διαφορετικού είδους προ-επεξεργασία απαιτείται για τις τρεις κατηγορίες δεδομένων εκπαίδευσης που θα εισάγουμε στο σύστημα πρόβλεψής μας.

Τέταρτο βήμα είναι η εφαρμογή των απαραίτητων αλγόριθμων, προκειμένου για κατηγοριοποίηση και συστηματοποίηση. Εν συνεχεία, αστικοποιούμε τα αποτελέσματα των αλγόριθμων που εφαρμόσαμε εξάγοντάς τα, είτε μέσω αρχείων του στατιστικού πακέτου R που θα χρησιμοποιήσουμε είτε μέσω Excel. Τελικό βήμα του συστήματός μας είναι η ανάλυση και η αξιολόγηση των αποτελεσμάτων που προσέφερε ο κάθε αλγόριθμος, υπολογίζοντας την ακρίβεια κάθε τεχνικής.

Ακολούθως, φαίνεται μέσω ενός γραφήματος η διαδικασία εφαρμογής των αλγορίθμων κατηγοριοποίησης και συστηματοποίησης που θα ακολουθήσουμε στη παρούσα εργασία και που περιγράφεται αναλυτικά στις επόμενες ενότητες.



ΕΙΚΟΝΑ 3-1: ΔΙΑΓΡΑΜΜΑ ΡΟΗΣ

ΚΕΦΑΛΑΙΟ 5
ΘΕΩΡΗΤΙΚΗ ΠΕΡΙΓΡΑΦΗ ΑΛΓΟΡΙΘΜΩΝ ΣΥΣΤΗΜΑΤΟΣ ΚΑΙ ΕΙΣΑΓΩΓΙΚΑ ΘΕΜΕΛΙΑ ΤΗΣ R

Στα προηγούμενα κεφάλαια περιγράφηκε το θεωρητικό υπόβαθρο που αφορά την εξόρυξη δεδομένων καθώς και οι μέθοδοι κατηγοριοποίησης, ομαδοποίησης και πρόβλεψης. Ο σκοπός της παρούσας διπλωματικής εργασίας είναι η εφαρμογή των μεθόδων εξόρυξης για την ανάλυση των χρηματιστηριακών δεδομένων και συγκεκριμένα των τραπεζικών μετοχών. Εφόσον τα δεδομένα μας αφορούν χρονίσεις, οι τεχνικές εξόρυξης που υλοποιήθηκαν είναι η κατηγοριοποίηση και η συσταδοποίηση. Στο παρών κεφάλαιο, λοιπόν, παρουσιάζεται το θεωρητικό πλαίσιο των αλγορίθμων συσταδοποίησης και κατηγοριοποίησης που θα χρησιμοποιηθούν στο σύστημά μας, ενώ παράλληλα γίνεται και μια εισαγωγή στο στατιστικό πακέτο R που θα χρησιμοποιηθεί.

5.1. ΤΟ ΠΡΟΒΛΗΜΑ ΤΗΣ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗΣ

Η παρούσα υποενότητα εστιάζεται στην επίλυση ενός προβλήματος κατηγοριοποίησης (*classification*). Το πρόβλημα της κατηγοριοποίησης έγκειται στον προσδιορισμό της κατηγορίας στην οποία ανήκει ένα αντικείμενο, βάσει ενός συνόλου μετρήσεων. Οι μετρήσεις αυτές λαμβάνονται μέσω αισθητήρων, οι οποίοι συλλέγουν πληροφορίες για τα υπό μελέτη αντικείμενα. Τα φυσικά αυτά μεγέθη αποτελούν τις μεταβλητές εισόδου του προβλήματος και ονομάζονται χαρακτηριστικά (*features*). Στόχος, λοιπόν, είναι ο προσδιορισμός της κατηγορίας (κλάσης) στην οποία ανήκει ένα αντικείμενο, θεωρώντας ένα προκαθορισμένο σύνολο κατηγοριών. Επίσης, κάθε μεμονωμένη παρατήρηση, δηλαδή κάθε σύνολο μετρήσεων για όλα τα χαρακτηριστικά ονομάζεται δείγμα (*sample*). Ο προσδιορισμός της κλάσης κάθε αντικειμένου δίνει την έξοδο του προβλήματος. Η διαδικασία εύρεσης των κλάσεων διενεργείται υπολογιστικά μέσω ενός αλγοριθμικού μοντέλου, τον κατηγοριοποιητή (*classifier*).

Κάθε μοντέλο κατηγοριοποίησης περιλαμβάνει ελεύθερες παραμέτρους (δηλαδή αριθμούς ή και δομικά στοιχεία των μεθοδολογιών που απαιτούνται προκειμένου για καθορισμό του τελικού μοντέλου), ο κατάλληλος προσδιορισμός των οποίων επιτρέπει την επίλυση διαφορετικών προβλημάτων κατηγοριοποίησης. Οι ελεύθερες παράμετροι καθορίζονται μέσω του αλγορίθμου εκπαίδευσης (*training algorithm*) ή αλγορίθμου εκμάθησης (*learning algorithm*), ο οποίος αποσκοπεί στον βέλτιστο καθορισμό τους (όσο είναι δυνατόν) ούτως ώστε να ελαχιστοποιείται (ή και να μηδενίζεται) το σφάλμα κατηγοριοποίησης.

Για το μηδενισμό του σφάλματος, λοιπόν, είναι απαραίτητη η ύπαρξη ενός συνόλου δειγμάτων, το οποίο είναι ήδη ορθά κατηγοριοποιημένο. Πρόκειται για το σύνολο εκπαίδευσης (*training set*), το οποίο περιλαμβάνει Q ζεύγη εισόδου-εξόδου της μορφής $e^p = (x^p, c^p), p = 1, \dots, Q$, που ονομάζονται πρότυπα εκπαίδευσης (*training patterns*). Σε κάθε πρότυπο, η συνιστώσα $x^p = \left[x_1^p, \dots, x_N^p\right]$ είναι το διάνυσμα εισόδου και $c^p \in C$ είναι η κλάση του προτύπου. Ουσιαστικά, ο αλγόριθμος εκπαίδευσης αξιοποιεί το σύνολο εκπαίδευσης για τη κατάλληλη εκμάθηση του μοντέλου, ώστε ο κατηγοριοποιητής που προκύπτει να κατηγοριοποιεί ορθά τα νέα δείγματα (Δόβα, 2013).

5.1.1. ΣΦΑΛΜΑ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗΣ ΚΑΙ ΤΕΧΝΙΚΕΣ ΔΙΑΣΤΑΥΡΩΜΕΝΗΣ ΕΠΙΚΥΡΩΣΗΣ

Δεδομένου ότι είναι αδύνατη η παντελής ανυπαρξία σφαλμάτων κατά τη διαδικασία κατηγοριοποίησης, κρίνεται απαραίτητη η αναφορά στο σφάλμα κατηγοριοποίησης *Err*, το οποίο ορίζεται ως το ποσοστό των εσφαλμένων κατηγοριοποιήσεων προς το πλήθος των στοιχείων του συνόλου εκπαίδευσης. Με μαθηματικούς όρους, έστω ότι το σύνολο

εκπαίδευσης έχει Q πρότυπα αναφοράς και έστω Q_c τα πρότυπα που ανατέθηκαν σε λάθος κατηγορία. Τότε το σφάλμα κατηγοριοποίησης ορίζεται ως:

$$Err = \frac{Q_c}{Q}$$

Το σφάλμα εκφράζεται είτε ως απόλυτος αριθμός που ανήκει στο διάστημα $[0,1]$ είτε ως ποσοστό επί τοις εκατό. Εναλλακτικά, αντί για σφάλμα μπορεί να υπολογιστεί η ακρίβεια κατηγοριοποίησης (*classification error*), η οποία εκφράζει το ποσοστό των ορθά κατηγοριοποιημένων προτύπων βάσει της σχέσης:

$$Acc = 1 - Err = 1 - \frac{Q_c}{Q}$$

Η ακρίβεια κατηγοριοποίησης παρουσιάζεται ως ποσοστό επί τοις εκατό κυρίως και όπως είναι το προφανές, ο βέλτιστος κατηγοριοποιητής παρουσιάζει ακρίβεια ίση με 100%.

Εντρυφώντας στην ορολογία, το σφάλμα που εντοπίζεται στα πρότυπα αναφοράς του συνόλου εκπαίδευσης ονομάζεται σφάλμα αντικατάστασης (*resubstitution error*) ή σφάλμα εκπαίδευσης (*training error*), καθώς αναφέρεται στα δεδομένα που χρησιμοποιήθηκαν για την εκπαίδευση του κατηγοριοποιητή. Προσπαθώντας ο κατηγοριοποιητής να ελαττώσει το σφάλμα αντικατάστασης (βελτιστοποιώντας τις ελεύθερες παραμέτρους), ενδέχεται να ελαττώσει την ακρίβεια κατηγοριοποίησης στα δεδομένα που δε χρησιμοποιήθηκαν κατά την εκπαίδευση. Το φαινόμενο αυτό ονομάζεται υπερπροσαρμογή (*overfitting*) στα δεδομένα εκπαίδευσης και οδηγεί σε μείωση της δυνατότητας γενίκευσης (*generalization*) του κατηγοριοποιητή στην ορθή κατηγοριοποίηση δεδομένων που δεν ανήκουν στα πρότυπα αναφοράς. Δεδομένου αυτού, η απόδοση του μοντέλου κρίνεται βάσει ενός ξεχωριστού συνόλου δεδομένων, το οποίο ονομάζεται σύνολο δοκιμής (*testing set*). Το υπολογιζόμενο σφάλμα κατηγοριοποίησης στο σύνολο δοκιμής ονομάζεται σφάλμα δοκιμής (*testing error*) ή σφάλμα γενίκευσης (*generalization error*).

Στη περίπτωση που δεν είναι διαθέσιμο κάποιο ξεχωριστό σύνολο δοκιμής, εφαρμόζεται η διαδικασία της διασταυρωμένης επικύρωσης (*cross-validation*). Η πλέον διαδεδομένη επιλογή είναι η διασταυρωμένη επικύρωση *k-fold*, κατά την οποία τα διαθέσιμα δεδομένα διαχωρίζονται με τυχαίο τρόπο σε *k* μη επικαλυπτόμενα τεμάχια (*folds*), κάθε ένα από τα οποία έχει τον ίδιο αριθμό προτύπων ενώ διατηρούνται προσεγγιστικά και οι σχετικές αναλογίες των προτύπων κάθε κλάσης. Η διαδικασία επαναλαμβάνεται k φορές θεωρώντας κάθε φορά ένα τεμάχιο ως δεδομένο δοκιμής και τα υπόλοιπα *k-1* ως δεδομένα εκπαίδευσης. Τελικό στάδιο είναι ο υπολογισμός του συνολικού σφάλματος κατηγοριοποίησης *k* συνόλων δοκιμής, που ονομάζεται σφάλμα διασταυρωμένης επικύρωσης (*cross-validation error*).

Εναλλακτικός τρόπος αναπαράστασης των σφαλμάτων κατηγοριοποίησης είναι μέσω του πίνακα σύγχυσης (*confusion matrix*). Πρόκειται για έναν τετραγωνικό πίνακα, στον οποίο ο αριθμός γραμμών και στηλών ισούται με τον αριθμό τον κατηγοριών (S.V.Stehman, 1997). Εντρυφώντας, έστω ένα στοιχείο $x_{i,j}$ από τον πίνακα. Το στοιχείο αυτό αναπαριστά τον αριθμό των προτύπων που ανήκουν στη κατηγορία *i* και ανατέθηκαν στη κατηγορία *j*. Οπότε, οι γραμμές του πίνακα αντικατοπτρίζουν τη κατηγορία αναφοράς και οι στήλες τη κατηγορία ανάθεσης. Τα στοιχεία της κύριας διαγωνίου αναπαριστούν τα ορθώς κατηγοριοποιημένα πρότυπα ανά κατηγορία, ενώ τα υπόλοιπα τα λανθασμένα.

Έστω ο ακόλουθος πίνακας σύγχυσης, για ένα υποθετικό πρόβλημα κατηγοριοποίησης με δύο κλάσεις $\{C_1, C_2\}$ και με έναν οποιονδήποτε κατηγοριοποιητή.

Κατηγορία Αναφοράς	*Αποτέλεσμα κατηγοριοποίησης*	
	C_1	C_2
C_1	32	18
C_2	0	50

ΠΙΝΑΚΑΣ 4-1: ΠΙΝΑΚΑΣ ΣΥΓΧΥΣΗΣ

Τότε το στοιχείο $x_{1,1}$ υποδηλώνει ότι 32 πρότυπα που ανήκουν στη κατηγορία C_1, σύμφωνα με τα δεδομένα αναφοράς κατηγοριοποίηθηκαν ορθώς στη κατηγορία C_1, ενώ το στοιχείο $x_{1,2}$ δείχνει ότι 18 πρότυπα της C_1 κατηγοριοποιήθηκαν τελικά στη κατηγορία C_2. Διαπιστώνουμε, δηλαδή, ότι ο πίνακας σύγχυσης παρέχει συνοπτικά και εύκολα τη πληροφορία για τον διαχωρισμό μεταξύ των κατηγοριών του προβλήματος.

5.1.2. ΚΑΤΗΓΟΡΙΟΠΟΙΗΤΗΣ QDA

Η διαχωριστική ανάλυση χρησιμοποιεί τα δεδομένα εκπαίδευσης προκειμένου να υπολογίσει τις παραμέτρους μιας συνάρτησης διαχωρισμού (*discriminant function*) $f(x)$, η οποία ορίζεται στο χώρο εισόδων και καθορίζει τα σύνορα διαχωρισμού μεταξύ των κατηγοριών του προβλήματος (Hilbe, 2009; Krzanowski, 1988; Seber, 1984). Έστω ένα πρόβλημα κατηγοριοποίησης με *N* χαρακτηριστικά και *M* κατηγορίες. Επίσης, έστω ένα σύνολο εκπαίδευσης με *Q* κατηγοριοποιημένα πρότυπα αναφοράς και ένα δείγμα εισόδου το αναπαρίσταται ως ένα διάνυσμα $x = [x_1, x_2, \dots, x_N]$.

Η διαχωριστική ανάλυση υποθέτει ότι όλα τα χαρακτηριστικά του προβλήματος ακολουθούν τη κανονική κατανομή. Σε πρώτη φάση υποθέτουμε ότι το πρόβλημα έχει μόνο δύο κατηγορίες (*M* =2) και οπότε χρησιμοποιείται η πιο απλή μορφή κατηγοριοποιητή η οποία υποθέτει ότι όλες οι επιμέρους τιμές συνδιακύμανσης (*covariance*) μεταξύ των χαρακτηριστικών είναι ίσες. Τότε η ανάλυση ονομάζεται Γραμμική Διαχωριστική Ανάλυση (*Linear Discriminant Analysis, LDA*) και έχει ως αποτέλεσμα τη δημιουργία ενός γραμμικού κατηγοριοποιητή.

Τα περισσότερα προβλήματα κατηγοριοποίησης δεν είναι γραμμικά διαχωρίσιμα, όπως και στη περίπτωσή μας. Στις περιπτώσεις αυτές μπορεί να χρησιμοποιηθεί η Τετραγωνική Γραμμική Ανάλυση (*Quadratic Discriminant Analysis, QDA*) που αποτελεί γενίκευση της *LDA*. Η *QDA* επίσης υποθέτει ότι τα χαρακτηριστικά ακολουθούν τη κανονική κατανομή αλλά δεν κάνει υπόθεση ως προς τις τιμές της συνδιακύμανσης των επιμέρους μεταβλητών. Η συνάρτηση διαχωρισμού του κατηγοριοποιητή QDA δίνεται ως εξής:

$$B_{QDA}(x) = K + x \cdot L + x \cdot Q \cdot x^T$$

Όπου K μια σταθερά, $L = [l_1, \dots, l_N]^T$ το διάνυσμα-στήλη των γραμμικών παραμέτρων διάστασης $N \times 1$ και όπου Q ένας τετραγωνικός πίνακας, διάστασης $N \times N$, στον οποίο εμπεριέχονται οι παράμετροι των τετραγωνικών όρων. Ένα δείγμα κατηγοριοποιείται σε μία από τις δύο κατηγορίες του προβλήματος ανάλογα το πρόσημο της συνάρτησης. Δηλαδή, το διάνυσμα x κατηγοριοποιείται στη πρώτη κλάση αν $B_{QDA}(x) < 0$ και στη δεύτερη αν $B_{QDA}(x) > 0$. Ακόμη, η εξίσωση $B_{QDA}(x) = 0$ ορίζει το σύνορο διαχωρισμού στο χώρο διαστάσεων. Η δική μας περίπτωση αφορά ένα πρόβλημα με δύο χαρακτηριστικά εισόδου, όπου ο πίνακας έχει τη μορφή:

$$Q = \begin{bmatrix} q_{11} & q_{12} \\ q_{21} & q_{22} \end{bmatrix}$$

Και οπότε η αρχική σχέση παίρνει τη μορφή:

$$B_{QDA}(x_1, x_2) = K + x_1 \cdot l_1 + x_2 \cdot l_2 + x_1^2 \cdot q_{11} + x_{21}^2 \cdot q_{22} + x_1 x_2 \cdot (q_{12} + q_{21})$$

Όπως είναι εμφανές, οι τετραγωνικοί όροι που εμφανίζονται στη παραπάνω σχέση υποδεικνύουν ότι πρόκειται για έναν σύνθετο κατηγοριοποιητή, ο οποίος συνίσταται για την επίλυση δύσκολων προβλημάτων κατηγοριοποίησης, αυξάνοντας ταυτόχρονα τη πιθανότητα μείωσης των δυνατοτήτων γενίκευσης του κατηγοριοποιητή.

Για ένα δεδομένο πρόβλημα κατηγοριοποίησης, σκοπός του αλγορίθμου εκπαίδευσης είναι η ανεύρεση βέλτιστων τιμών για τις παραμέτρους *K, L* και *Q*, βάσει του συνόλου εκπαίδευσης. Αρχικά, υπολογίζονται ο μέσος όρος κάθε χαρακτηριστικού και ο πίνακας συνδιακύμανσης (*covariance matrix*), όπως προκύπτουν από τα δεδομένα εκπαίδευσης. Στη συνέχεια, οι τιμές αυτές χρησιμοποιούνται για των υπολογισμό των προαναφερθέντων παραμέτρων μέσω κατάλληλων μαθηματικών σχέσεων.

5.1.3. ΚΑΤΗΓΟΡΙΟΠΟΙΗΤΗΣ ΕΓΓΥΤΑΤΟΥ ΓΕΙΤΟΝΑ

Ο κατηγοριοποιητής εγγύτατου γείτονα (KNN) βασίζεται στην απλή ιδέα ότι ένα δείγμα μπορεί να κατηγοριοποιηθεί βάσει της απόστασής του από αντιπροσωπευτικά πρότυπα εκπαίδευσης (Mitchell, 1997; Cover & P.E.Hart, 1967; Shakhnarovich, Darell, & Indyk, 2005). Έστω ένα πρόβλημα κατηγοριοποίησης που ορίζεται σε έναν χώρο *N* μεταβλητών εισόδου $X = \{X_1, \dots, X_N\} \subseteq R^N$ και έστω $C = \{C_1, \dots, C_M\}$ το σύνολο των *M* κατηγοριών (κλάσεων). Επίσης, έστω ένα σύνολο εκπαίδευσης E_{trn} με Q κατηγοριοποιημένα πρότυπα αναφοράς $E_{trn} = \{e^p = (x^p, c^p), p = 1, \dots, Q\}$, όπου $x^p \in N$ και $c^p \in C$. Δεδομένου ακόμη ενός μέτρου απόστασης *D*, ο KNN κατηγοριοποιεί τα δείγματα εισόδου βάσει των k κοντινότερων προτύπων αναφοράς, όπου k το πλήθος των γειτόνων που λαμβάνονται υπόψη (ο χρήστης καθορίζει τη τιμή του).

Η λειτουργία του περιγράφεται ακολούθως: Για ένα οποιαδήποτε δείγμα $y \in X$ που καλούμαστε να κατηγοριοποιήσουμε, υπολογίζονται οι αποστάσεις του δείγματος με όλα τα σημεία του συνόλου εκπαίδευσης $D(y, x^p), p = 1, \dots, Q$, και κρατούνται τα k σημεία με τη μικρότερη απόσταση. Το δείγμα y κατηγοριοποιείται τελικά στην κατηγορία που εμφανίζεται με τη μεγαλύτερη συχνότητα στα k πρότυπα εκπαίδευσης, μια διαδικασία που ονομάζεται κανόνας πλειοψηφίας (*majority rule*). Όταν *k=1*, το δείγμα κατηγοριοποιείται στη κλάση του πλησιέστερου προτύπου εκπαίδευσης.

Ως μέτρο απόστασης θεωρείται η ευκλείδεια απόσταση, αλλά κατά καιρούς έχουν προταθεί διάφορα μέτρα απόφασης. Αξίζει να σημειωθεί, δε, ότι σε κάθε περίπτωση κατηγοριοποίησης απαιτείται ένα διαφορετικό μέτρο απόστασης, προκειμένου να βελτιστοποιηθεί η ακρίβεια του κατηγοριοποιητή. Έστω x και y δύο σημεία του χώρου εισόδου που εκφράζονται ως διανύσματα γραμμή διάστασης *N*. Τότε τα μέτρα υπολογίζονται βάσει των σχέσεων:

1. Ευκλείδεια απόσταση:

$$D_{euc}(x, y) = (x - y)(x - y)^T$$

2. Απόσταση οικοδομικού τετραγώνου:

$$D_{cb}(x, y) = \sum_{i=1}^{N} |x_i - y_i|$$

3. Απόσταση συνημιτόνου:

$$D_{cos}(x,y) = 1 - \frac{x \cdot y^T}{\sqrt{(x \cdot x^T)(y \cdot y^T)}}$$

4. Απόσταση συνέλιξης:

$$D_{cor}(x,y) = 1 - \frac{(x-\bar{x})(x-\bar{y})^T}{\sqrt{(x-\bar{x})(x-\bar{x})^T \cdot (y-\bar{y})(y-\bar{y})^T}}$$

Όπου, $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$ και $\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$

5.1.4. ΔΕΝΤΡΑ ΑΠΟΦΑΣΕΩΝ

Τα δέντρα αποφάσεων (*decision trees*) είναι μια οικογένεια μη παραμετρικών κατηγοριοποιητών, που συνήθως καταφέρνουν υψηλά ποσοστά ακρίβειας και χρησιμοποιούνται σε ένα ευρύ φάσμα προβλημάτων κατηγοριοποίησης (Breiman, Friedman, Olsen, & Stone, 1984). Πρόκειται για μια εύληπτη διαδικασία κατά την οποία το μοντέλο κατηγοριοποίησης αποτελείται από ένα δυαδικό δέντρο, όπου κάθε κόμβος λαμβάνει μια διχοτομική απόφαση σε ένα χαρακτηριστικό του προβλήματος ενώ κάθε φύλλο αναπαριστά μία κλάση. Υπάρχουν διάφορες μορφές δέντρων αποφάσεων, η πιο διαδεδομένη εκ των οποίων και αυτή που θα ασχοληθούμε στη παρούσα εργασία είναι ο αλγόριθμος που ονομάζεται δέντρο κατηγοριοποίησης και παλινδρόμησης (*Classification and Regression Tree, CART*) .

Η περίπτωση κατηγοριοποίησης αριθμητικών δεδομένων αφορά ένα πρόβλημα που περιλαμβάνει τρία χαρακτηριστικά {Close Value, Volume} και δύο κατηγορίες {Good, Bad}, ενώ το σύνολο εκπαίδευσης περιέχει 248 πρότυπα εκπαίδευσης. Ένα μοντέλο CART δομείται από ένα σύνολο κανόνων της μορφής «ΑΝ … , ΤΟΤΕ κατηγοριοποίησε το δείγμα στη κλάση …, ΑΛΛΙΩΣ…». Το τμήμα υπόθεσης (το μέρος ΑΝ) εμπεριέχει μια ανισότητα για ένα χαρακτηριστικό του προβλήματος ενώ το τμήμα συμπεράσματος (το μέρος ΤΟΤΕ) περιγράφει την απόφαση που λαμβάνεται όταν ισχύει η ανισότητα της υπόθεσης. Τώρα, αν η απόφαση δεν αναθέτει το δείγμα σε μια κατηγορία του προβλήματος, οδηγούμαστε σε νέο κανόνα που έχει ως αποτέλεσμα τη δημιουργία πιο σύνθετων δέντρων. Το μέρος ΑΛΛΙΩΣ αφορά την απόφαση που λαμβάνεται όταν δεν ισχύει η υπόθεση. Έτσι, δημιουργείται το δυαδικό δέντρο αποφάσεων.

Ένα δέντρο CART δημιουργείται με τον ακόλουθο τρόπο: Η ρίζα (*root*) του δέντρου, δηλαδή ο αρχικός κόμβος, θεωρείται ότι αναπαριστά ένα φύλλο (τερματικοί κόμβοι, *leafs*), η κλάση του οποίου καθορίζεται ως αυτή που περιέχει το μεγαλύτερο αριθμό προτύπων αναφοράς στο σύνολο εκπαίδευσης. Σε αυτό το σημείο, το μοναδικό φύλλο αναθέτει όλα τα πρότυπα σε μία κλάση και κατηγοριοποιεί εσφαλμένως τα πρότυπα των υπόλοιπων κλάσεων. Το φύλλο που οδηγεί σε εσφαλμένες κατηγοριοποιήσεις ονομάζεται ακάθαρτο (*impure*) και πρέπει να μετατραπεί σε κόμβο (*node*), ο οποίος αντιστοιχεί σε μια νέα υπόθεση, δηλαδή σε έναν νέο υποθετικό κανόνα.

Η διαρκής επανάληψη αυτού του βήματος, οδηγεί σε μια αρκετά εκτεταμένη μορφή δέντρου, όπου τα περισσότερα φύλλα θα κατηγοριοποιούν ένα μόνο πρότυπο του συνόλου εκπαίδευσης. Για να μη συμβαίνει αυτό, η μετατροπή ενός φύλλου σε κόμβο διενεργείται εφόσον πληρούνται δύο κριτήρια:

- Ένα ακάθαρτο φύλλο μετατρέπεται σε κόμβο αν περιέχει τουλάχιστον *MinParent* πρότυπα αναφοράς
- Τα δύο νέα φύλλα που δημιουργούνται πρέπει να περιέχουν τουλάχιστον *MinLeaf* πρότυπα αναφοράς το καθένα.

Οι παράμετροι *MinParent* και *MinLeaf* εισάγονται από τον χρήστη και η επιλογή τους επηρεάζουν ποικιλοτρόπως τη διαδικασία. Δηλαδή, δύναται περιορισμός της δομής του μοντέλου αλλά ο καθορισμός της βέλτιστης τιμής τους είναι πρακτικά δύσκολος, καθώς μπορεί εύκολα να οδηγήσει σε υπεραπλουστευμένες δομές. Αν τελικά ένα φύλλο μετατραπεί σε κόμβο, επιλέγεται ένα χαρακτηριστικό και μια τιμή για το δεξιό μέλος της ανισότητας, η βέλτιστη τιμή των οποίων διενεργείται βάσει ενός στατιστικού κριτηρίου διχοτόμησης (*split criterion*). Η διαδικασία επαναλαμβάνεται μέχρι να σταθεροποιηθεί η τελική δομή του δέντρου, ή μέχρι να μη δύναται η διάσπαση κάποιου νέου φύλλου.

Προκειμένου να αποφευχθεί η χρήση των *MinParent* και *MinLeaf* χρησιμοποιείται μια πιο συστηματική διαδικασία που ονομάζεται κλάδεμα (*pruning*) του δέντρου. Ένα μοντέλο CART δημιουργείται προοδευτικά και παράλληλα ο αλγόριθμος εκπαίδευσης αναγνωρίζει κάποιους κόμβους οι οποίοι θα μπορούσαν να μετατραπούν σε φύλλα χωρίς να μειώνεται σημαντικά η ακρίβεια κατηγοριοποίησης. Η πληροφορία αυτή αποθηκεύεται κατά τη δημιουργία του μοντέλου και μπορεί να χρησιμοποιηθεί για την απλοποίηση της δομής σε δεύτερη φάση. Η διαδικασία κλαδέματος γίνεται σε επίπεδα (*levels)*, που αποτελούν τα κομβικά σημεία της εκπαίδευσης που έχουν αποθηκευτεί. Σημειώνουμε, τέλος, ότι το μέγιστο επίπεδο, το οποίο εξαρτάται από το πρόβλημα κατηγοριοποίησης και από το βάθος του αρχικού δέντρου, εξαρτάται από το ελάχιστο δυνατό δέντρο, δηλαδή αυτό που περιέχει μόνο ένα φύλλο και αποτελεί τη ρίζα του δέντρου.

Ο μαθηματικός τύπος που χρησιμοποιεί ο αλγόριθμος για την επιλογή σημείου διάσπασης *s* για τον κόμβο *t* είναι ο ακόλουθος (Πελέκης,2015):

$$\Phi(s/t) = 2P_L P_R \sum_{j=1}^{m} \left| P\left(C_j \middle| t_L\right) - P\left(C_j \middle| t_R\right) \right|$$

Όπου οι πιθανότητες P_L, P_R αντιστοιχούν στη πιθανότητα μια εγγραφή να βρεθεί αντίστοιχα στην αριστερή ή στη δεξιά πλευρά του δέντρου, ενώ οι πιθανότητες $P\left(C_j \middle| t_L\right), P\left(C_j \middle| t_R\right)$ αντιστοιχούν στην εξαρτημένη πιθανότητα μια εγγραφή να ανήκει στη κλάση C_j εφόσον βρίσκεται στην αριστερή ή δεξιά, αντίστοιχα, πλευρά του δέντρου.

5.2. ΤΟ ΠΡΟΒΛΗΜΑ ΤΗΣ ΣΥΣΤΑΔΟΠΟΙΗΣΗΣ

Πρόκειται για μια μεθοδολογία ανακάλυψης συστάδων και κατανομών ή προτύπων (*patterns*) που παρουσιάζουν ενδιαφέρον για τα δεδομένα που εξετάζουμε. Ως συστάδα (*cluster*) ορίζουμε μια συλλογή αντικειμένων (*objects*) από τα δεδομένα, βάσει της ομοιότητας τους. Δηλαδή, τα αντικείμενα μιας συστάδας αναμένεται να έχουν όμοια συμπεριφορά μεταξύ τους και ανόμοια με τα αντικείμενα άλλων συστάδων.

Σε αντίθεση με την ταξινόμηση (*classification*), που αναφέρθηκε σε προηγούμενη ενότητα, η συσταδοποίηση δε στηρίζεται σε εκ των προτέρων ορισμένες κλάσεις και εκπαιδευτικά παραδείγματα με διαμορφωμένα χαρακτηριστικά ανά κλάση (*class-labeled training examples*), αλλά επιτρέπει στα δεδομένα να αυτοπροσδιοριστούν βάσει των ομοιοτήτων τους ή των διαφορών τους και να καθοριστούν έτσι το πλήθος κλάσεων και τα ποιοτικά τους χαρακτηριστικά.

Ένα πλεονέκτημα της διαδικασίας αυτής, είναι η ικανότητά της να δημιουργεί τις αρχικές κατηγορίες στις οποίες οι τιμές ενός συνόλου μπορούν να κατηγοριοποιηθούν κατά την εφαρμογή της ταξινόμησης. Όσον αφορά τη μηχανική εκμάθηση (*machine learning*) και την αναγνώριση προτύπων, η ανάλυση συστάδων αναφέρεται συχνά ως μη εποπτευόμενη μάθηση (*unsupervised learning*).

5.2.1. ΔΙΑΔΙΚΑΣΙΑ ΣΥΣΤΑΔΟΠΟΙΗΣΗΣ

Αρχικά, είναι αναγκαίο να τονίσουμε ότι η προ-επεξεργασία των δεδομένων πριν εφαρμοστεί η συσταδοποίηση σε ένα σύνολο δεδομένων είναι αναγκαία καθώς από αυτή εξαρτάται ο τρόπος τμηματοποίησης της μεθόδου. Τα βήματα προκειμένου για επιτυχημένη συσταδοποίηση είναι (Fayyad et al, 1996):

1. **Επιλογή κατάλληλων γνωρισμάτων** (*attributes*) στα οποία θα εφαρμοστεί η μέθοδος έτσι ώστε να κωδικοποιηθεί όσο το δυνατόν περισσότερη πληροφορία.
2. **Επιλογή του κατάλληλου αλγορίθμου συσταδοποίησης** που θα οδηγήσει στο καθορισμό ενός καλού σχήματος συσταδοποίησης (*clustering scheme*) για το δοθέν σύνολο δεδομένων. Η επιλογή του αλγορίθμου καθορίζεται από το μέτρο εγγύτητας (*proximity measure*), το οποίο καθορίζει πόσο όμοια είναι τα δύο αντικείμενα, και από το κριτήριο συσταδοποίησης. Πρόκειται για ένα κριτήριο που εκφράζεται μέσω μιας συνάρτησης κόστους ή κάποιου άλλου τύπου κανόνων. Η κατάλληλη επιλογή του κριτηρίου οδηγεί σε τμηματοποίηση που ταιριάζει με το υπό μελέτη σύνολο δεδομένων.
3. **Επικύρωση αποτελεσμάτων:** Η κατάλληλη χρήση κριτηρίων και τεχνικών δίνει τη δυνατότητα προσδιορισμού της ακρίβειας των αποτελεσμάτων που δίνει ο αλγόριθμος. Δεδομένου, ακόμη, ότι οι αλγόριθμοι συσταδοποίησης καθορίζουν ομάδες που δεν είναι εκ των προτέρων γνωστές, είναι συχνό η τελική τμηματοποίηση να απαιτεί κάποιου είδους αξιολόγησης.
4. **Ερμηνεία αποτελεσμάτων**: Στόχος είναι να προκύψει ορθό αποτέλεσμα. Για το λόγο αυτό συνίσταται η ένωση των αποτελεσμάτων της διαδικασίας με άλλα πειραματικά στοιχεία και αποτελέσματα προηγούμενης ανάλυσης.

5.2.2. ΔΙΑΚΡΙΣΗ ΜΕΘΟΔΩΝ ΚΑΙ ΑΛΓΟΡΙΘΜΩΝ

Στα πλαίσια της βιβλιογραφικής ανασκόπησης, διαπιστώσαμε ότι υπάρχουν αρκετοί αλγόριθμοι συσταδοποίησης. Για το λόγο αυτό κρίνεται απαραίτητη η κατηγοριοποίησης τους σε συγκεκριμένες ομάδες προκειμένου για αποσαφήνιση. Ο διαχωρισμός έγινε βάσει της μεθόδου συσταδοποίησης που θέλουμε να εφαρμόσουμε στα δεδομένα μας. Οι κατηγορίες είναι (Han & M.Kamber, 2001):

- Διαιρετική συσταδοποίηση (*Partitional clustering*)

Έστω ότι έχουμε μια βάση δεδομένων *n* αντικειμένων. Τότε μπορούμε να δημιουργήσουμε *k* διαμερίσεις (*partitions*) των δεδομένων, όπου κάθε διαμέριση παριστάνει μια συστάδα και ο αριθμός των διαμερίσεων είναι μικρότερος του συνόλου δεδομένων ($k \leq n$). Τα βασικά χαρακτηριστικά των ομάδων που δημιουργούνται είναι ότι κάθε ομάδα πρέπει να έχει τουλάχιστον ένα αντικείμενο, ότι κάθε αντικείμενο πρέπει να ανήκει σε μια ομάδα και ότι οι δημιουργούμενες ομάδες είναι ασυσχέτιστες.

Βασικότερος στόχος των αλγορίθμων της κατηγορίας αυτής είναι η ελαχιστοποίηση των μέτρων ανομοιότητας μεταξύ των δειγμάτων εντός κάθε συστάδας και η μεγιστοποίηση της ανομοιότητας μεταξύ των διαφορετικών συστάδων και το επιτυγχάνουν ελαχιστοποιώντας ή

μεγιστοποιώντας (ανάλογα τη περίσταση) τη κατάλληλη συνάρτηση. Ως ανοιμοιότητα θεωρούμε τη διαφορά μεταξύ δύο υποκειμένων *i* και *j*, ενώ σημειώνουμε ότι υπάρχει πιθανότητα να απαιτηθεί επανάληψη της διαδικασίας, μετά τη δημιουργία της αρχικής διαμέρισης, προκειμένου να υπάρξει βελτίωση της δομής των συστάδων. Η επανάληψη σημαίνει τη μετακίνηση ενός αντικειμένου από τη μια συστάδα στην άλλη.

Οι κλασσικοί αλγόριθμοι που ανήκουν στη διαδικασία αυτή είναι οι k-means και k-medoid, ενώ παραλλαγές του k-means αποτελούν οι αλγόριθμοι k-modes και k-prototypes (Σταυλιώτης, 2008).

- Ιεραρχική συσταδοποίηση (*Hierarchical clustering*)

Η ιεραρχική μέθοδος προκαλεί μια ιεραρχική αποσύνθεση του υπό μελέτη συνόλου δεδομένων. Συγκεκριμένα, η ιεραρχική συσταδοποίηση χωρίζεται σε δύο μέρη, ανάλογα με τον τρόπου που επιθυμεί ο χρήστης να γίνει ο διαχωρισμός. Το πρώτο μέρος αποτελεί την συσσωρευτική προσέγγιση (*agglomerative approach*) ή 'από κάτω προς τα πάνω' προσέγγιση (*"bottom-up" approach*). Η διαδικασία ξεκινά θεωρώντας κάθε αντικείμενο ως μεμονωμένη συστάδα και προχωρά 'προς τα πάνω' δημιουργώντας ομάδες. Το δεύτερο μέρος είναι η διαχωριστική προσέγγιση (*divisive approach*) ή αλλιώς 'από πάνω προς τα κάτω' προσέγγιση (*"top-down" approach*). Η διαδικασία εδώ ξεκινά θεωρώντας όλα τα αντικείμενα σε μια συστάδα, τα οποία στη συνέχεια διαχωρίζονται σε ομοιογενείς ομάδες. Στο τέλος της διαδικασίας, κάθε αντικείμενο ανήκει σε συστάδα ή σε καμία, ανάλογα τις συνθήκες που πληρούνται.

Το ελάττωμα των ιεραρχικών μεθόδων συσταδοποίησης είναι ότι κάθε συσσώρευση ή διαχωρισμός που πραγματοποιείται δε μπορεί να ανακληθεί. Για τη βελτίωση της μεθόδου, δύναται η δυνατότητα συνδυασμού ιδιοτήτων μέσω της ιεραρχικής συσσώρευσης (*hierarchical agglomeration*). Δηλαδή, εφαρμόζοντας αρχικά έναν συσσωρευτικό αλγόριθμο και στη συνέχεια να γίνει εκλέπτυνση του αποτελέσματος μέσω μιας επαναληπτικής τοποθέτησης. Μερικοί αλγόριθμοι αυτής της κατηγορίας είναι οι κλιμακούμενοι αλγόριθμοι συσταδοποίησης BIRCH (*Balanced Iterative Reducing and Clustering using Hierarchies*), ο οποίος είναι διαχωριστικός, και CURE (*Clustering Using Representatives*), ο οποίος είναι συσσωρευτικός. Επίσης, στην ίδια κατηγορία ανήκει ο ROCK (*Robust for Categorical Attributes*) που εφαρμόζεται σε κατηγορικά δεδομένα.

Η μέθοδος δίνει ως έξοδο ένα δέντρο από συστάδες, το δενδρογράφημα. Πρόκειται για ένα δέντρο το οποίο δίνει μια απεικόνιση της σχέσης μεταξύ των τελικών νεοδημιουργηθέντων συστάδων και δε πρέπει να συγχέεται με τα δέντρα ταξινόμησης ή παλινδρόμησης τα οποία χρησιμοποιούνται για τη διενέργεια πρόβλεψης.

- Μέθοδοι βασισμένες στη πυκνότητα (*Density-based methods*)

Η διαφορά των αλγορίθμων αυτής της κατηγορίας από τους διαιρετικούς αλγόριθμους είναι ότι καταλήγουν στη δημιουργία αλγορίθμων αυθαίρετου (*arbitrary*) σχήματος ενώ σκοπός τους είναι η ένωση των περιοχών όμοιας πυκνότητας σε συστάδες ή αλλιώς, αποσκοπούν στην ένωση των αντικειμένων των συστάδων με βάση τη συνάρτηση κατανομής πυκνότητας (*density function distribution*). Έχουν την ικανότητα να απομακρύνουν το θόρυβο από τα δεδομένα.

Χαρακτηριστικοί αλγόριθμοι που βασίζονται στη πυκνότητα είναι ο DBSCAN (*Density-Based Spatial Clustering of Applications with Noise*) και ο DENCLUE (*DENsity-based CLUstering*).

- Μέθοδοι βασισμένες στο πλέγμα(*Grids-based methods*)

Οι αλγόριθμοι που βασίζονται στο πλέγμα μετατρέπουν σε κβάντα (*quantizes*) το χώρο του αντικειμένου σε ένα περιορισμένο αριθμό κελιών, τα οποία δημιουργούν μια δομή πλέγματος. Έπειτα, στη δομή που δημιουργήθηκε εφαρμόζονται λειτουργίες συσταδοποίησης. Το θετικό της κατηγορίας είναι ότι έχουν γρήγορο χρόνο εκτέλεσης αφού ο χρόνος εκτέλεσης εξαρτάται μόνο από τον αριθμό των κελιών που δημιουργούνται σε κάθε διάσταση του κβαντωμένου χώρου. Ενδεικτικά αναφέρονται οι αλγόριθμοι STING (*Statistical Information Grid*) και CLIQUE (*Clustering In Quest*).

- Μέθοδοι βασισμένες σε μοντέλο (*Model-based methods*)

Οι συγκεκριμένοι μέθοδοι υποθέτουν ένα συγκεκριμένο μοντέλο για κάθε μια συστάδα και στη συνέχεια επιχειρούν να βρουν τη βέλτιστη λύση μεταξύ του υποθετικού μοντέλου και των δεδομένων. Στηρίζονται στην υπόθεση ότι τα δεδομένα παράγονται από μια ανάμειξη υποκείμενων κατανομών πιθανότητας και υπάρχουν δύο τρόποι προσέγγισης τέτοιων μεθόδων: η στατιστική προσέγγιση (στηρίζεται στη θεωρία πιθανοτήτων) και η προσέγγιση μέσω νευρωνικών δικτύων (*neural networks*). Ένας αλγόριθμος στατιστικής προσέγγισης είναι ο COBWEB ενώ από την ομάδα νευρωνικών δικτύων ο πιο δημοφιλής είναι ο SOM (*Self-Organized feature Map*).

5.2.3. ΑΛΓΟΡΙΘΜΟΣ K-MEANS

Όπως αναφέρθηκε σε προηγούμενη ενότητα η μέθοδος k-means είναι μια από τις πιο συχνές μεθόδους συσταδοποίησης που ανήκει ση κατηγορία της διαιρετικής συσταδοποίησης. Πρόκειται για αλγόριθμο που δε χρειάζεται να κρατά στη μνήμη πολλά στοιχεία, είναι αρκετά γρήγορος και απαιτεί λίγες επαναλήψεις (Κούτρας, 2007). Η αδυναμία του εντοπίζεται στο ότι οι συστάδες που δημιουργεί έχουν διαφορετικό μέγεθος ή πυκνότητα και ότι δεν εξαλείφει τα outliers από τα δεδομένα. Είναι ένας αλγόριθμος που είναι κατάλληλος για συνεχή δεδομένα και αποσκοπεί στην άμεση αποσύνθεση του συνόλου δεδομένων σε ένα σύνολο ασυσχέτιστων συστάδων. Αυτό επιτυγχάνεται ελαχιστοποιώντας τη μέση τετραγωνική απόσταση των δεδομένων από τα πλησιέστερα κέντρα των συστάδων. Η συνάρτηση που χρησιμοποιείται δίνεται από τη σχέση:

$$E = \sum_{i=1}^{k} \sum_{x \in C_i} d(x, m_i)$$

Όπου m_i το κέντρο της συστάδας C_i και $d(x, m_i)$ η ευκλείδεια απόσταση μεταξύ ενός στοιχείου x και του κέντρου m_i.

Ουσιαστικά, ο αλγόριθμος χρησιμοποιώντας την E , αρχικά, ελαχιστοποιεί την απόσταση κάθε σημείου από το κέντρο της συστάδας όπου ανήκει το σημείο και στη συνέχεια αναθέτει κάθε στοιχείο του συνόλου δεδομένων στη συστάδα της οποίας το κέντρο είναι πιο κοντά και ξανά-υπολογίζει τα κέντρα, χρησιμοποιώντας το μέσο όρο των σημείων τους. Για να ελαχιστοποιηθεί η E θεωρούνται k σημεία ως τα κέντρα k συστάδων και επιλέγονται οι k πρώτες εγγραφές ή οι k αντιπροσωπευτικές εγγραφές από κάθε συστάδα. Αυτό συνεχίζεται μέχρι να σταματήσουν να αλλάζουν τα κέντρα των συστάδων. Η τελική λύση σχετίζεται με τον ορισμό των αρχικών κέντρων και με τον τρόπο διάταξης των αντικειμένων στο σύνολο δεδομένων οπότε αν n το σύνολο των αντικειμένων, τότε πρέπει $k \leq n$.

5.3. ΣΤΑΤΙΣΤΙΚΟ ΠΑΚΕΤΟ R

Πρόκειται για μια γλώσσα προγραμματισμού που χρησιμοποιείται κυρίως για ανάλυση δεδομένων και εφαρμογή διαφόρων 'κλασικών' και 'σύγχρονων' στατιστικών τεχνικών. Είναι διαθέσιμη προς όλους και μπορεί να αποκτηθεί δωρεάν από την ιστοσελίδα http://www.r-foge.r-projects.org ή από τα πρότυπα (*mirror*) του CRAN (*Comprehensive R Archive*) από την ιστοσελίδα http://cran.r-project.org. Το CRAN είναι ένα δίκτυο διανομής της R παγκοσμίως, μέσω διαδικτύου.

Αξίζει να σημειωθεί, ότι η R μπορεί να χρησιμοποιηθεί είτε κατευθείαν με εντολές που υπάρχουν είτε με προγράμματα που δημιουργεί ο χρήστης ή παρέχονται σε πακέτα. Οι δυνατότητές της επεκτείνονται ταχύτατα, δίνοντας τη δυνατότητα στο χρήστη να χειρίζεται εξειδικευμένες τεχνικές, γραφήματα, εργαλεία αναφοράς κλπ.

Στη δική μας περίπτωση, θα χρησιμοποιήσουμε πακέτα που απαιτούνται για τους κατηγοριοποιητές QDA, KNN και CART ενώ όσον αφορά τα κειμενικά, θα ασχοληθούμε με το πακέτο tm (text mining), το οποίο προσφέρει τεράστιες δυνατότητες στην εξόρυξη δεδομένων. Η λειτουργία του αφορά τη διαχείριση εγγράφων κειμένου ενώ αφαιρεί τη διαδικασία χειραγώγησης του χειραγώγησης κειμένου και διευκολύνει τη χρήση ετερογενών μορφών κειμένου. Το πακέτο έχει ολοκληρωμένη βάση back-end η οποία υποστηρίζει την ελαχιστοποίηση των απαιτήσεων της μνήμης. Παρέχει ακόμα την δυνατότητα ανάγνωσης διαφόρων μορφών κειμένων όπως PDF, XML, TXT, EXCEL, WORD και άλλων αρχείων (Μακρής 2015).

Στη παρούσα εργασία χρησιμοποιείται η έκδοση 3.1.1. Για κάθε υποενότητα που ακολουθεί περιγράφονται τα βήματα για τη διαδικασία κατηγοριοποίησης αρχικά των αριθμητικών δεδομένων και στη συνέχεια των κειμενικών.

ΚΕΦΑΛΑΙΟ 6
ΕΦΑΡΜΟΓΗ ΑΛΓΟΡΙΘΜΩΝ DATA MINING ΓΙΑ ΚΑΘΕ ΣΥΝΟΛΟ ΔΕΔΟΜΕΝΩΝ

Σκοπός του κεφαλαίου, είναι να παρουσιαστεί ο τρόπος επεξεργασίας ενός όγκου αδόμητων δεδομένων εφαρμόζοντας τους αλγορίθμους που περιγράφηκαν στη προηγούμενη ενότητα. Συγκεκριμένα, ο αδόμητος όγκος που συλλέχθηκε αφορά αριθμητικά και κειμενικά δεδομένα τραπεζικών μετοχών για το έτος 2014. Η διαδικασία υλοποίησης περιλαμβάνει τρία επίπεδα εφαρμογής. Χωρίσαμε τα δεδομένα σε τρεις κατηγορίες, όπου πρώτη περιλαμβάνει αριθμητικά δεδομένα, δηλαδή τις τιμές κλεισίματος των τραπεζικών μετοχών για το έτος 2014, η δεύτερη τα κειμενικά δεδομένα για τις ίδιες μετοχές, δηλαδή άρθρα, δελτία τύπου ή/και ανακοινώσεις που εκδόθηκαν το έτος 2014 και αφορούσαν τις συγκεκριμένες μετοχές και τέλος η τρίτη κατηγορία περιλαμβάνει αριθμητικά και κειμενικά δεδομένα των επικείμενων μετοχών.

6.1 ΚΑΤΗΓΟΡΙΟΠΟΙΗΤΗΣ QDA ΣΕ ΑΡΙΘΜΗΤΙΚΑ ΔΕΔΟΜΕΝΑ

1ο Βήμα: Create a Text File & Read Data

Αρχικό βήμα είναι αποθηκεύσουμε τα αριθμητικά μας δεδομένα σε ένα αρχείο .txt και να το αποθηκεύσουμε. Για κάθε μετοχή εφαρμόζουμε την ίδια διαδικασία και συνεπώς ο κώδικας θα παρουσιαστεί για τη μία. Έπειτα, με τις ακόλουθες εντολές διαβάζουμε το αρχείο στην R:

```
sc=read.table("C:/Users/user/Desktop/Tot_Num_R/AlfaTXTNum.txt",header=T)
attach(sc)
print(sc)
str(sc)
```

Η εντολή attach() χρησιμοποιείται προκειμένου το αρχείο να είναι προσπελάσιμο στην R με λιγότερες πληκτρολογήσεις. Η εντολή print() μας δείχνει το αρχείο που εισήχθηκε και η εντολή str() δίνει μια εποπτεία στα δεδομένα μας. Δηλαδή είδος αρχείου (data frame), σύνολο μεταβλητών και κατηγορία αυτών. Στη προκειμένη έχουμε στη πρώτη στήλη τις τιμές κλεισίματος της μετοχής, στη δεύτερη τη τιμή του όγκου για κάθε μετοχή ημερησίως και στη τρίτη τη κατηγορία (0:" Good", 1: "Bad"). Επίσης, συνολικά έχουμε 248 δεδομένα.

Το αποτέλεσμα της str() φαίνεται στη παρακάτω εικόνα:

```
> str(sc)
'data.frame':   248 obs. of  3 variables:
 $ Close : num  0.652 0.641 0.669 0.695 0.709 0.737 0.73 0.706 0.727 0.705 ...
 $ Volume: int  6291033 3986210 12529355 14008718 22434800 22187437 12251459 10
 $ Class : int  0 0 0 0 0 1 0 0 0 1 ...
>
```

2ο Βήμα: Loading Data into R

Επόμενο βήμα είναι να 'φορτώσουμε' τα πακέτα που απαιτούνται αφού δεν υπάρχουν από προεπιλογή. Εγκαθίστανται και φορτώνονται χρησιμοποιώντας τον ακόλουθο κώδικα:

```
install.packages("knitr")  # Install Packages
install.package("caret")
install.packages("Mass")
library(knitr) # Load Packages
library(caret)
library(MASS)# For QDA
```

3ο Βήμα: Partitioning Data

`

Ουσιαστικά, χωρίζουμε ισομερώς τα δεδομένα των στηλών σε δεδομένα εκπαίδευσης και δεδομένα δοκιμής. Η εντολή sc_train_labels() παίρνει τια 124 πρώτες τιμές κατηγοριοποίησης της μεταβλητής απόκρισης και επιστρέφει ένα data frame με τις ετικέτες για τα πρώτα 124. Σκοπός είναι να εφαρμόσουμε τον κατηγοριοποιητή για τα δεδομένα εκπαίδευσης και έπειτα να εφαρμόσουμε τα δεδομένα δοκιμής προκειμένου να διαπιστώσουμε την εγκυρότητα του QDA. Οι εντολές φαίνονται ως εξής:

```
sc_train <- sc[1:124,]
sc_test <- sc[125:248,]
sc_train_labels <- sc[1:124, 3]
sc_test_labels <- sc[125:248, 3]
```

4ο Βήμα: Fit QDA Model in Training Data, Predict on Test Data & Accuracy

Χρησιμοποιούμε τη συνάρτηση qda() χρησιμοποιώντας τις εντολές του διπλανού σχήματος:

```
model.qda <- qda(Class~ .,sc_train)
qda.fit <- qda(Class ~ Close + Volume,
data =sc_train)
qda.fit
```

Το output είναι:

```
> qda.fit
Call:
qda(Class ~ Close + Volume, data = sc_train)

Prior probabilities of groups:
        0         1
0.5564516 0.4435484

Group means:
      Close   Volume
0 0.6950725 24917504
1 0.6963273 24142485
```

Αφού δημιουργήθηκε το μοντέλο βάσει του συνόλου εκπαίδευσης, το χρησιμοποιούμε προκειμένου να προβλέψουμε τη κατηγοριοποίηση του συνόλου δοκιμής και να επικυρώσουμε την ακρίβεια του κατηγοριοποιητή. Χρησιμοποιούμε τη συνάρτηση predict() για τη πρόβλεψη και τη συνάρτηση ConfusionMatrix() για να δημιουργήσουμε το πίνακα σύγχυσης δηλαδή το πίνακα που υποδεικνύει το αποτέλεσμα της κατηγοριοποίησης. Τέλος, υπολογίζοντας το μέσο (mean) υπολογίζουμε την εγκυρότητα ή το σφάλμα διασταυρωμένης επικύρωσης του μοντέλου. Οι εντολές φαίνονται ακολούθως:

```
qda.class <- predict(qda.fit, sc_test)$class
confusionMatrix(sc_test$Class,
predict(model,sc_test))
mean(qda.class == sc_test$Class)
#Accuracy
mean(qda.class != sc_test$Class)
# Misclassification Error
```

Το output της ακρίβειας και του πίνακα διασταυρωμένης επικύρωσης δίνεται ως εξής:

```
> mean(qda.class == sc_test$Class)##Accuracy
[1] 0.516129
> mean(qda.class != sc_test$Class) ## Miss-classification Error
[1] 0.483871
>
```

`

```
> confusionMatrix(sc_test$Class, predict(model,sc_test))
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0  8 50
         1 10 56

               Accuracy : 0.5161
                 95% CI : (0.4247, 0.6068)
    No Information Rate : 0.8548
    P-Value [Acc > NIR] : 1

                  Kappa : -0.0142
 Mcnemar's Test P-Value : 4.782e-07

            Sensitivity : 0.44444
            Specificity : 0.52830
         Pos Pred Value : 0.13793
         Neg Pred Value : 0.84848
             Prevalence : 0.14516
         Detection Rate : 0.06452
   Detection Prevalence : 0.46774
      Balanced Accuracy : 0.48637

       'Positive' Class : 0
```

Η μήτρα σύγχυσης υποδεικνύει ότι από τα 58 πραγματικά (actual) δεδομένα που κατηγοριοποιούνται ως 0 (“Good”) ο κατηγοριοποιητής κατηγοριοποίησε σωστά τα 8 και τα υπόλοιπα 50 λάθος. Αντίστοιχα, από τα 66 πραγματικά δεδομένα που ανήκουν στη κατηγορία 1 (“Bad”) ο αλγόριθμος κατηγοριοποίησε τα 10 λάθος και τα 56 ορθά.

Παρατηρούμε ότι η Negative Prediction (Πρόβλεψη “Bad”) είναι ~85% που σημαίνει ότι όταν η τιμή της μετοχής μειώνεται μπορούμε με ακρίβεια ~85% να συμπεράνουμε ότι θα συνεχίσει να μειώνεται (R Statistics.net|TheNo.1 Educational Reference For Statistical Computing With R~ Seek.Learn.Grow, 2015; R.Kelly, 2014).

Τα αποτελέσματα εγκυρότητας και λάθους κατηγοριοποίησης υποδηλώνουν ότι οι προβλέψεις του μοντέλου είναι ακριβείς κατά 51%, που θεωρείται καλό ποσοστό όταν πρόκειται για μετοχές.

Alternative 4ο Βήμα: Fit Random Forest Model for QDA

Μέσω του πακέτου care
προσαρμόζουμε το μοντέλο QDA στ
δεδομένα, αφού μετατρέψουμε τ
μεταβλητή απόκρισης (Class) σ
παράγοντα. Με τη συνάρτηση set.seed(
επιβεβαιώνουμε ότι θα παίρνουμε για κάθ
επανάληψη το ίδιο αποτέλεσμα. Τέλο
μέσω του αλγορίθμου Random Forest κα
χρησιμοποιώντας τη μέθοδ
διασταυρωμένης επικύρωσ
υπολογίζουμε την εγκυρότητα. Τέλο
υπολογίζουμε τη μήτρα σύγχυσης όπω
και πριν. Οι εντολές παρουσιάζονται στο
διπλανό πίνακα:

```
# Convert Class to Factor
sc_train$Class <- factor(sc_train$Class)
set.seed(42) # Set a random seed
# Train the model using a "random forest" algorithm
model <- train(Class ~ Close + Volume, #Class is a function of the variables we decided to include
               data = sc_train, # Use the trainSet dataframe as the training data
               method = "qda",# Use the "random forest" algorithm
               trControl = trainControl(method = "cv", # Use cross-validation
                                        number = 5)
) # Use 5 folds for cross-validation
model
confusionMatrix(sc_test$Class, predict(model,sc_test))
```

Τα αποτελέσματα δίνονται από τη παρακάτω εικόνα:

`

```
> model
Quadratic Discriminant Analysis

124 samples
  2 predictor
  2 classes: '0', '1'

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 99, 99, 100, 99, 99
Resampling results

  Accuracy  Kappa       Accuracy SD  Kappa SD
  0.427     -0.1078367  0.06360031   0.1366611
```

Βλέπουμε τις λεπτομέρειες του δείγματος εκπαίδευσης καθώς και την ακρίβεια του μοντέλου που ισούται με 42.7%.

6.1.1 ΚΑΤΗΓΟΡΙΟΠΟΙΗΤΗΣ KNN ΣΕ ΑΡΙΘΜΗΤΙΚΑ ΔΕΔΟΜΕΝΑ

1ο Βήμα: Install and Load Required Packages

Φορτώνουμε τα πακέτα που απαιτούνται για το κατηγοριοποιητή KNN με τις εντολές που φαίνονται στο διπλανό πίνακα:

```
install.packages("class")  # Install Packages
install.packages("gmodels")
library(class) # Load Packages
library(gmodels)
```

2ο Βήμα & 3ο Βήμα: Read Data & Splitting
Η διαδικασία όπως και πριν.

4ο Βήμα: Fit KNN Model

Εφαρμόζουμε το μοντέλο βάσει της συνάρτησης που παρέχεται στο πακέτο. Οι εντολές είναι:

```
sc$Class <-factor(sc$Class)
table(sc$Class)
model.knn <- knn(train=sc_train, test=sc_test,
cl=sc_train_labels, k=2, prob=T)
```

Χρησιμοποιούμε τη συνάρτηση knn() προκειμένου να κατηγοριοποιήσουμε τα δεδομένα εκπαίδευσης. Η επιλογή για το k είναι 2 γιατί θέλουμε να δημιουργήσουμε δύο ομάδες κατηγοριοποίησης.

5ο Βήμα: Evaluate the model performance

Χρησιμοποιώντας το μοντέλο που δημιουργήσαμε, θα ελέγξουμε την ακρίβεια των προβλεπόμενων τιμών προκειμένου να αποφανθούμε για το αν 'ταιριάζουν' με τις πραγματικές. Για το λόγο αυτό θα κάνουμε χρήση της συνάρτησης CrossTbale(). Η εντολή δίνεται ακολούθως:

```
set.seed(25)
CrossTable(x=sc_test_labels,y=model.knn,prop.chisq=F) # Accuracy
```

`

Το output είναι:

```
> set.seed(25)
> CrossTable(x=sc_test_labels,y=model.knn,prop.chisq=F)# Accuracy

   Cell Contents
|-------------------------|
|                       N |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|

Total Observations in Table:  124

                | model.knn
sc_test_labels |         0 |         1 | Row Total |
---------------|-----------|-----------|-----------|
             0 |        38 |        20 |        58 |
               |     0.655 |     0.345 |     0.468 |
               |     0.535 |     0.377 |           |
               |     0.306 |     0.161 |           |
---------------|-----------|-----------|-----------|
             1 |        33 |        33 |        66 |
               |     0.500 |     0.500 |     0.532 |
               |     0.465 |     0.623 |           |
               |     0.266 |     0.266 |           |
---------------|-----------|-----------|-----------|
  Column Total |        71 |        53 |       124 |
               |     0.573 |     0.427 |           |
---------------|-----------|-----------|-----------|
```

Αρχικά χρησιμοποιούμε την εντολή set.seed() για να έχουμε το ίδιο αποτέλεσμα για κάθε επανάληψη. Το σύνολο δοκιμών αποτελείται από 124 παρατηρήσεις εκ των οποίων 38 έχουν προβλεφθεί σωστά (*True Negatives*) ως "Good" το οποίο συνιστά το 30.6%. Επίσης, 33 από τις 124 παρατηρήσεις προβλέφθηκαν ορθά (*True Positives*) ως "Bad", το οποίο συνιστά το 26.6%.. Υπάρχουν 33παρατηρήσεις που ενώ ήταν "Good" προβλέφθηκαν λάθος (*False Negatives*) δηλαδή το 26.6% και ομοίως 20 παρατηρήσεις που ενώ ήταν "Bad" προβλέφθηκαν ως "Good" (*False Positives*), δηλαδή το 16.%.

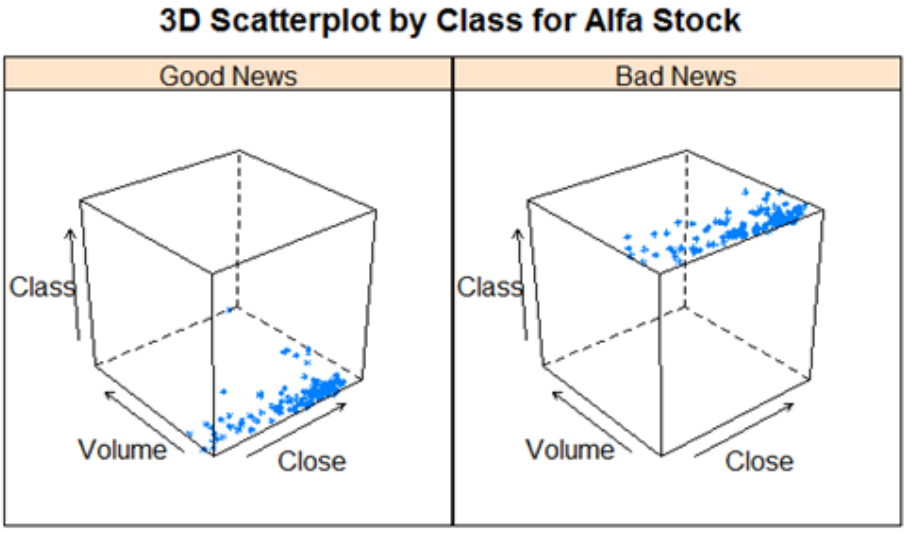
Η συνολική ακρίβεια του μοντέλου ισούται με 57% (38+33/124). Το μοντέλο βελτιώνει την απόδοση με διαδοχικές επαναλήψεις (αλλαγή της τιμής της set.seed) (Choudhury, 2015).

Το γράφημα από το αποτέλεσμα της κατηγοριοποίησης μπορεί να αποκτηθεί από τις εντολές που παρουσιάζονται στο διπλανό πίνακα (R. I. Kabacoff, 2014):

```
library(lattice)
Classification<- factor(sc$Class,levels=c(0,1),
  labels=c("Good News","Bad News"))
cloud(Class~Close*Volume|Classification,
  main="3D Scatterplot by Class for Alfa
Stock")
```

Το γράφημα έχει τη μορφή:

`



ΕΙΚΟΝΑ 4-1: 3D ΑΠΕΙΚΟΝΙΣΗ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗΣ ΤΩΝ ΤΙΜΩΝ ΤΗΣ ALFA ΜΕΤΟΧΗΣ

6.1.2 ΚΑΤΗΓΟΡΙΟΠΟΙΗΤΗΣ CART ΣΕ ΑΡΙΘΜΗΤΙΚΑ ΔΕΔΟΜΕΝΑ

1ο Βήμα: Install and Load Required Packages

Όπως προηγήθηκε, φορτώνουμε στην R τα κατάλληλα πακέτα για την εφαρμογή του αλγορίθμου CART:

```
install.packages("rpart")  # Install Packages
install.packages("rpart.plot")
 library(rpart) # Load Packages
 library(rpart.plot)
```

2ο Βήμα: Read Data & Splitting Generating a Function

Όμοια με πριν διαβάζουμε τα δεδομένα και στη συνέχεια χωρίζουμε τα δεδομένα δημιουργώντας μια συνάρτηση, εν αντιθέσει με τις προηγούμενες τεχνικές. Αφού δημιουργηθεί η συνάρτηση, είναι απαραίτητο τα νέα σύνολα δεδομένων να μετατραπούν σε data frames. Ακολούθως, παρουσιάζονται οι αντίστοιχες εντολές (DnI Institute|Build Data And Decision Science Experience, 2014; R Statistics.net|TheNo.1 Educational Reference For Statistical Computing With R~ Seek.Learn.Grow, 2015):

```
sc=read.table("C:/Users/user/Desktop/Tot_Num_R/AlfaTXTNum.txt",header=T)
attach(sc)
# splitdf function will return a list of training and testing sets
sc$Class<-as.factor(sc$Class)
sc$Class
splitdf <- function(dataframe, seed=NULL) {
        if (!is.null(seed)) set.seed(seed)
        index <- 1:nrow(dataframe)
        trainindex <- sample(index, trunc(length(index)/2))
        trainset <- dataframe[trainindex, ]
        testset <- dataframe[-trainindex, ]
        list(trainset=trainset,testset=testset)
}
#Apply the function
splits <- splitdf(sc, seed=808)
#Use str() to return a list - two data frames called trainset and testset
str(splits)
# Save the training and testing sets as data frames
training <- splits$trainset
testing <- splits$testset
```

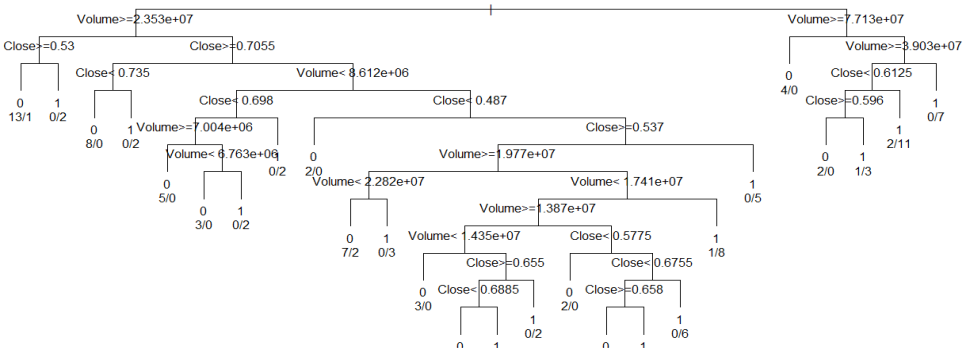
`

3ο Βήμα: Fit CART Model

Εφαρμόζουμε τον Αλγόριθμο CART στα δεδομένα εκπαίδευσης, σιγουρεύοντας ότι οι μεταβλητές είναι όλες κατηγορικές προκειμένου να δημιουργηθεί το δέντρο. Οι εντολές έπονται (Ma, 2014):

```
set.seed(19)
model.tree <- rpart(Class ~ Close + Volume ,
training, control = rpart.control(minsplit=5))
plot(model.tree, uniform=T)
text(model.tree, use.n=T)
```

Το δέντρο που δημιουργεί ο κατηγοριοποιητής παρουσιάζεται ως εξής:



ΕΙΚΟΝΑ 4-2: ΑΠΕΙΚΟΝΙΣΗ ΔΕΝΔΡΟΓΡΑΜΜΑΤΟΣ ΑΠΟ ΤΟ ΚΑΤΗΓΟΡΙΟΠΟΙΗΤΗ CART

4ο Βήμα: Prune Tree

Προκειμένου να 'κλαδέψουμε' το δέντρο που δημιουργήσαμε, θα υπολογίσουμε το μέγεθος που πρέπει να έχει το δέντρο ώστε να ελαχιστοποιεί το σφάλμα διασταυρωμένης επικύρωσης (*Misclassification rate*). Άρα, ψάχνουμε τη τιμή cp που ελαχιστοποιεί το xerror. Οπότε υπολογίζουμε τη ποσότητα με τις εντολές:

```
printcp(model.tree)
bestcp <-
model.tree$cptable[which.min(model.tree$cpt
able[,"xerror"]),"CP"]
```

Η βέλτιστη τιμή φαίνεται στην επόμενη εικόνα:

```
> printcp(model.tree)

Classification tree:
rpart(formula = Class ~ Close + Volume, data = training, control = rpart.contro$

Variables actually used in tree construction:
[1] Close  Volume

Root node error: 59/124 = 0.47581

n= 124

        CP nsplit rel error  xerror     xstd
1 0.101695      0   1.00000 1.10169 0.094258
2 0.084746      1   0.89831 1.13559 0.094062
3 0.067797      3   0.72881 1.08475 0.094320
4 0.033898      5   0.59322 0.79661 0.091565
5 0.028249      9   0.45763 0.83051 0.092271
6 0.022599     12   0.37288 0.84746 0.092584
7 0.016949     15   0.30508 0.83051 0.092271
8 0.011299     22   0.18644 0.93220 0.093765
9 0.010000     25   0.15254 0.94915 0.093926
```

`

Αφού υπολογίστηκε το βέλτιστο cp μπορούμε να υπολογίσουμε την ακρίβεια της πρόβλεψης που θα ισούται με:

$$Pred.Error\ in\ CV = Root\ node\ error \cdot xerror \cdot 100\% = 0.47581 \cdot 0.59322 \cdot 100\% = 28.22\%$$

Συνεπώς, η ακρίβεια της πρόβλεψης ισούται με 71.77%. Δημιουργούμε το δέντρο που απαιτείται με τις εντολές:

```
tree.pruned <- prune(model.tree, cp = bestcp)
predict(tree.pruned)
nrow(predict(tree.pruned))
```

5ο Βήμα: Confusion Matrix

Για να δημιουργήσουμε το πίνακα σύγχυσης θα χρησιμοποιήσουμε τις ακόλουθες εντολές:

```
conf.matrix <- table(training$Class,
predict(tree.pruned,type="class"))
rownames(conf.matrix) <- paste("Actual",
rownames(conf.matrix), sep = ":")
colnames(conf.matrix) <- paste("Pred",
colnames(conf.matrix), sep = ":")
print(conf.matrix)
```

Η μήτρα σύγχυσης είναι η εξής:

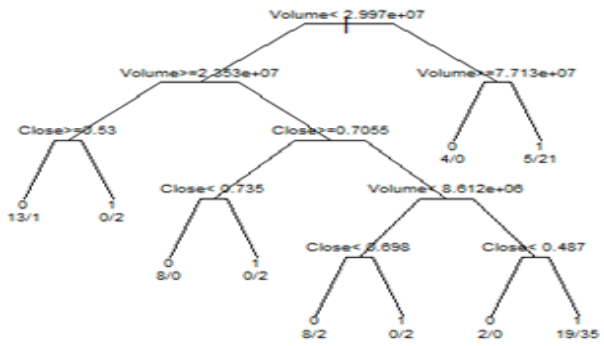
	Pred:0	Pred:1
Actual:0	33	26
Actual:1	9	56

Δηλαδή, 33 παρατηρήσεις από τις 124 κατηγοριοποιήθηκαν ορθά στη κατηγορία 0 και 56 από τις 124 κατηγοριοποιήθηκαν επίσης ορθά στη κατηγορία 1.

Το νέο δέντρο απόφασης δημιουργείται από τις εντολές:

```
# Decision Tree Dendogram
plot(tree.pruned,   uniform = T, compress = T, margin = 0.2, branch = 0.3)
# Label on Decision Tree
text(tree.pruned,  use.n = T,  digits = 3,  cex = 0.6)
```

Οπότε θα έχουμε:



ΕΙΚΟΝΑ 4-3: ΑΠΕΙΚΟΝΙΣΗ ΠΕΡΙΟΡΙΣΜΕΝΟΥ ΔΕΝΔΡΟΓΡΑΜΜΑΤΟΣ ΑΠΟ ΤΟΝ ΚΑΤΗΓΟΡΙΟΠΟΙΗΤΗ CART

`

Ο συγκεντρωτικός πίνακας αποτελεσμάτων κατηγοριοποίησης αριθμητικών δεδομένων για κάθε μετοχή παρουσιάζεται στον επόμενο πίνακα:

Όνομα Μετοχής	Accuracy			
	QDA_1	**QDA_2**	**KNN**	**CART**
Alfa	51.61%	42.70%	57.25%	71.77%
Attica	53.22%	62.83%	49.19%	71.77%
ETE	61.29%	58.13%	53.62%	78.22%
ΤτΕ	45.96%	55.66%	51.61%	8548%
Kyp*	60%	-	40.00%	60.00%
Peir	44.35%	49.13%	50.80%	92.74%
Eur	54.03%	48.33%	52.41%	85.48%

ΠΙΝΑΚΑΣ 4-2:ΣΥΓΚΕΝΤΡΩΤΙΚΟΣ ΠΙΝΑΚΑΣ ΑΠΟΔΟΣΗΣ ΚΑΤΗΓΟΡΙΟΠΟΙΗΤΩΝ.

Παρατηρούμε ότι για τη πρώτη εφαρμογή του QDA καλύτερη κατηγοριοποίηση επιτυγχάνεται στη μετοχή ΕΤΕ και για τη δεύτερη εφαρμογή του κατηγοριοποιητή καλύτερη κατηγοριοποίηση γίνεται για τη μετοχή της Attica. Ο αλγόριθμος KNN παρουσιάζει καλύτερα αποτελέσματα για την Alfa ενώ ο CART κατηγοριοποιεί καλύτερα τη μετοχή της Κύπρου. Παράλληλα, διαπιστώνουμε ότι ο πιο αποδοτικός κατηγοριοποιητής είναι ο CART καθώς σε κάθε περίπτωση παρουσιάζει τη μεγαλύτερη ακρίβεια στα δεδομένα.

6.1.3 ΙΕΡΑΡΧΙΚΗ ΣΥΣΤΑΔΟΠΟΙΗΣΗ ΚΑΙ K-MEANS ΣΕ ΚΕΙΜΕΝΙΚΑ ΔΕΔΟΜΕΝΑ

Η διαδικασία κατηγοριοποίησης και ομαδοποίησης που εφαρμόζεται στα κειμενικά δεδομένα, απαιτεί τη δημιουργία και εισαγωγή ελληνικών *stopwords*. Πρόκειται για λέξεις που είναι κοινές στο λεξιλόγιο οπότε δεν έχουν σημασία για τη μέθοδο που ακολουθείται και οπότε αφαιρούνται. Επίσης, η λίστα για τις κοινές λέξεις δεν είναι καθορισμένη αλλά μπορεί να μεταβληθεί (να προστεθούν ή να αφαιρεθούν λέξεις), ανάλογα με τους σκοπούς του χρήστη. Δεδομένου ότι δεν υπάρχει έτοιμη εντολή στο στατιστικό πακέτο που χρησιμοποιούμε, για την ελληνική λίστα, δημιουργούμε τη δική μας λίστα και την εισάγουμε στο πακέτο, όπως θα δειχθεί στη διαδικασία προ- επεξεργασίας που θα ακολουθήσει (TRANSLATUM| The Greek Translation Vortal, 2006).

Εν αντιθέσει με πριν, για τα κειμενικά δεδομένα δημιουργούμε τρεις κατηγορίες "Good", "Bad" και "No Movers" βάσει του κανόνα που αναφέρθηκε σε προηγούμενη ενότητα (βλ. ενότητα **Error! Reference source not found.**). Όπως και πριν, περιγράφουμε τα βήματα για κάθε τεχνική που χρησιμοποιείται. Για τα κειμενικά δεδομένα παρουσιάζουμε, επίσης, τα αποτελέσματα της μετοχής της Alpha Bank.

Υπενθυμίζουμε ότι η ιεραρχική ομαδοποίηση καθώς και η μέθοδος K-Means δεν απαιτούν προγενέστερη κατηγοριοποίηση δεδομένων και για το λόγο αυτό χρησιμοποιούμε μόνο τα δελτία τύπου που εκδόθηκαν για τη κάθε μετοχή. Για τη μετοχή της Κύπρου έχουμε συνολικά τέσσερα δεδομένα.

1^0 Βήμα: Install & Load Required Packages

* ΓΙΑ ΤΗ ΜΕΤΟΧΗ ΤΗΣ ΤΡΑΠΕΖΑΣ ΚΥΠΡΟΥ ΧΡΗΣΙΜΟΠΟΙΗΘΗΚΑΝ 10 ΔΕΔΟΜΕΝΑ.

`

```
#Install
install.packages("tm")  # For text mining
install.package("SnowballC") # For text stemming
install.packages("wordcloud") # Word-cloud generator
install.packages("RColorBrewer") # Color palettes
#Load
library("tm")
library("SnowballC")
library("wordcloud")
library("RColorBrewer")
```

2^0 Βήμα: Create Corpus

Μετά τη φόρτωση των κατάλληλων πακέτων, είμαστε σε θέση να δημιουργήσουμε μια συλλογή εγγράφων, το Corpus, στο περιβάλλον της R. Η διαδικασία αυτή περιλαμβάνει τη φόρτωση αρχείων (τα δεδομένα μας) σε ένα κείμενο Corpus. Το πακέτο tm παρέχει τη συνάρτηση Corpus προκειμένου να γίνει αυτό. Υπάρχουν αρκετοί τρόποι για να δημιουργηθεί το Corpus και παρακάτω θα παραταχθεί ο κώδικας που χρησιμοποιήθηκε για τα δεδομένα μας. Με λίγα λόγια, ένα Corpus είναι μια λίστα αρχείων (στη προκειμένη ενός αρχείου.

Αφού εισάγουμε το αρχείο στη πλατφόρμα και δημιουργήσουμε το Corpus μπορούμε να 'δούμε' αν το αρχείο εισήχθηκε σωστά με την εντολή *inspect()* ή μπορούμε να δούμε συγκεκριμένη σειρά ή σειρές με την εντολή *writeLines().*

Η διαδικασία προ-επεξεργασίας επιτρέπει την αφαίρεση αριθμών, των κοινών λέξεων και των σημείων στίξης και μετατροπή των κεφαλαίων γραμμάτων σε πεζά, ώστε το κείμενό μας να είναι διαθέσιμο για ανάλυση. Η εντολή *removePunctuation* περιλαμβάνει την αφαίρεση σημείων στίξης και άλλων ειδικών χαρακτήρων, τα οποία διαβάζονται ως λέξεις από την R. Επίσης, δίνεται η δυνατότητα προσαρμογής της εντολής ώστε να αφαιρέσει συγκεκριμένα σημεία στίξης που επιθυμεί ο χρήστης. Η εντολή *removeNumbers* αφαιρεί τα αριθμητικά δεδομένα του κειμένου δεδομένου ότι συνήθως οι αριθμοί δεν επηρεάζουν το νόημα του κειμένου, αν και ίσως να μην αληθεύει πάντα αυτό. Η εντολή *content_transformer(tolower)* είναι απαραίτητη για τη σωστή λειτουργία του Corpus γιατί το πακέτο αντιλαμβάνεται για παράδειγμα τις λέξεις 'Κείμενο' και 'κείμενο' ως δύο διαφορετικές. Εμείς θέλουμε κάθε λέξη να εμφανίζεται ως η ίδια κάθε φορά που εντοπίζεται στο κείμενο και για αυτό αλλάζουμε τα πάντα σε πεζά. Η εντολή *removewords* χρησιμοποιείται για την αφαίρεση κοινών λέξεων 'η/και για την αφαίρεση λέξεων που επιθυμεί ο χρήστης. Στη προκειμένη, δημιουργούμε τη δική μας λίστα κοινών λέξεων, την οποία επεξεργαζόμαστε κατάλληλα (εισαγωγή στο πακέτο και εισαγωγή στο Corpus) προκειμένου να αφαιρέσουμε τις ανούσιες λέξεις. Η εντολή *stemDocument* χρησιμοποιείται για την αφαίρεση των καταλήξεων των λέξεων. Ουσιαστικά, με την εντολή αυτή αποκόπτουμε τιε καταλήξεις ώστε οι λέξεις με κοινή ρίζα να εμφανίζονται βάσει της κοινής τους ρίζας (για παράδειγμα 'χρηματιστήριο' και 'χρηματιστηρίων'). Η εντολή *stripWhitespace* μας διαβεβαιώνει ότι θα αφαιρεθούν από το κείμενό μας τα 'κενά' που δημιουργούνται κατά τη προηγούμενη επεξεργασία που υπέστη (με την αφαίρεση λέξεων κλπ) (Ma M. , 2014; Wordpress.com, 2015; STHDA| Statistical Tools For High-Throuput Data Analysis, 2015; Khalifa, 2015).

Είναι σημαντικό μετά από κάθε εντολή επεξεργασίας να επαναλαμβάνουμε τις εντολές *inspect()* και *writeLines()* ώστε να ελέγχουμε τα επίπεδα επεξεργασίας. Ο κώδικας παρουσιάζεται ακολούθως:

`

```
Al <- read.csv ("C:/Users/user/Desktop/Tot_Text_R/Clust_Text_R/AlfaCSVPlain.csv",sep
=";", header = TRUE,stringsAsFactors=F) # Insert Data
str(Al)
txt <- Corpus(DataframeSource(Al)) # Create Corpus
summary(txt); writeLines(as.character(txt[[1]]))
#Remove punctuation – replace punctuation marks with " "
    replacePunctuation <- content_transformer(function(txt) {return (gsub("[[:punct:]]"," ",
    txt,perl = T))})
    txt<- tm_map(txt, replacePunctuation ); writeLines(as.character(txt[[2]]))
#Transform to lower case (need to wrap in content_transformer)
    txt <- tm_map(txt,content_transformer(tolower)); writeLines(as.character(txt[[1]]))
#Strip digits (std transformation, so no need for content_transformer)
     txt <- tm_map(txt, removeNumbers) ; writeLines(as.character(txt[[2]]))
#Creates stopwords in Greek and other custom words
     StopWords=read.table("C:/Users/user/Desktop/Tot_Text_R/Gr_Stopwords.txt",header=T)
     attach(StopWords)
     StopWords_vec = as.vector(StopWords$Gr_St)
     txt<-tm_map(txt,removeWords,StopWords_vec); writeLines(as.character(txt[[2]]))
# Text stemming
     txt<- tm_map(txt, stemDocument); writeLines(as.character(txt[[1]]))
# Eliminate extra white spaces
    txt <- tm_map(txt, stripWhitespace); writeLines(as.character(txt[[1]]))
```

3^{0} Βήμα: Create Document Term Matrix (TDM)

Πρόκειται για έναν πίνακα που παραθέτει όλες τις εμφανίσεις των λέξεων στο Corpus, ανά έγγραφο. Στο TDM, τα έγγραφα παρουσιάζονται ανά γραμμές και οι όροι (*terms*) ανά στήλες. Αν μια λέξη εμφανίζεται σε ένα συγκεκριμένο έγγραφο, τότε ο πίνακας εισάγει στην αντίστοιχη σειρά και στήλη το 1, αλλιώς το 0. Επίσης, πολλαπλές εμφανίσεις σε ένα έγγραφο σημαίνει ότι στην ανάλογη θέση του πίνακα θα εμφανίζεται ο αριθμός εμφανίσεων την εκάστοτε λέξης. Η εντολή δημιουργίας του TDM δίνεται ακολούθως:

```
dtm <- TermDocumentMatrix(txt)
 m <- as.matrix(dtm)
 v<-sort(rowSums(m),decreasing=TRUE)
 d <- data.frame(word = names(v),freq=v)
 head(d, 10) # Show first 10
```

4^{0} Βήμα: Mining the Corpus

Η κατασκευή του πίνακα TDM σημαίνει ότι το Corpus μετατράπηκε σε μαθηματικό αντικείμενο, το οποίο μπορεί να αναλυθεί χρησιμοποιώντας ποσοτικές τεχνικές άλγεβρας. Αρχικά, παίρνουμε τη λίστα με τους όρους που εμφανίζονται τουλάχιστον 4 φορές σε όλο το Corpus, με την εντολή *findFreqTerms()*. Έπειτα, βάσει των συχνότερα εμφανιζόμενων λέξεων μπορούμε να ελέγξουμε τις συσχετίσεις (*correlations*) μεταξύ αυτών και κάποιων λέξεων που εμφανίζονται στο σώμα.

Η συσχέτιση εδώ, αποτελεί ένα ποσοτικό μέτρο των λέξεων που συνυπάρχουν. Το πακέτο tm επιτρέπει τον υπολογισμό της συσχέτισης με τη συνάρτηση *findAsssocs(),* που απαιτεί και την εισαγωγή του βαθμό συσχέτισης που επιθυμούμε. Ο βαθμός είναι ένας αριθμός μεταξύ

του 0 και του 1 και ορίζει ένα κατώτερο φράγμα της ισχύος της συσχέτισης μεταξύ των όρων αναζήτησης και των τελικών όρων. Εμείς θέτουμε ως *correlation limit* το 0.5 που σημαίνει ότι θα παρατεθούν οι λέξεις που ο όρος συνύπαρξης με τις *frequent* είναι το ελάχιστο 50%.

```
#Explore frequent terms and their associations
findFreqTerms(dtm, lowfreq = 4)
findAssocs(dtm, terms = "υψηλής", corlimit = 0.5)
head(d, 10) #The frequency table of first 10 words
```

Παρακάτω απεικονίζονται οι 10 πιο συχνές λέξεις του κειμένου:

```
> head(d, 10)
                word freq
alpha          alpha  188
bank            bank  188
ευρώ            ευρώ  143
ψήφου          ψήφου   75
δισ              δισ   72
εκατ            εκατ   66
αύξηση        αύξηση   63
οικονομίας οικονομίας   63
τράπεζας    τράπεζας   63
μετοχών      μετοχών   61
```

Η σημαντικότητα των λέξεων αναπαρίσταται γραφικά μέσω της εντολής *wordcloud* ενώ οι πιο συχνές λέξεις μπορούν να αναπαρασταθούν γραφικά μέσω της εντολής *barplot*. Οι αντίστοιχοι κώδικες παρουσιάζονται ακολούθως:

```
#Create Word Cloud
set.seed(1234) # So we get the same look each time
wordcloud(words = d$word, freq = d$freq, min.freq = 1, max.words=200, random.order=FALSE,
colors=brewer.pal(8, "Dark2"))
barplot(d[1:10,]$freq, las = 2, names.arg = d[1:10,]$word, col ="pink", main ="Most frequent words",
ylab = "Word frequencies") #Plot the frequencies
```

Το Word Cloud που δημιουργήθηκε απεικονίζεται ακολούθως:

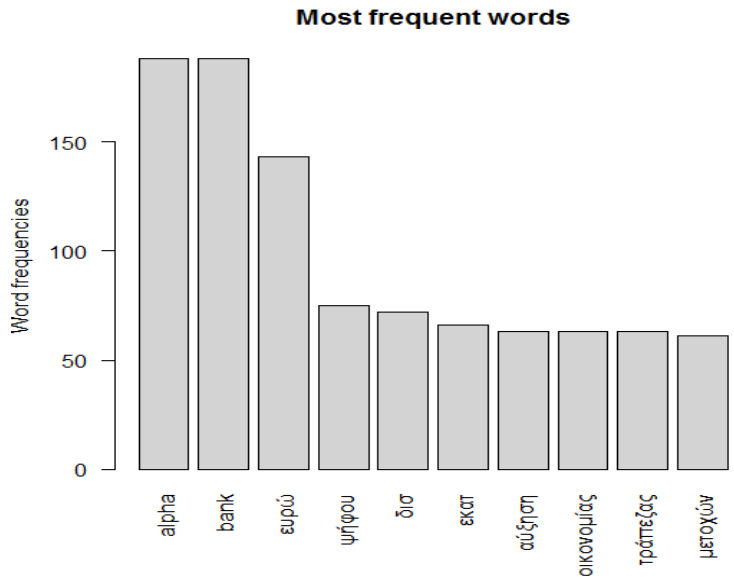
ΕΙΚΟΝΑ 4-4: ΑΠΕΙΚΟΝΙΣΗ WORD CLOUD ΓΙΑ ΤΑ ΚΕΙΜΕΝΙΚΑ ΔΕΔΟΜΕΝΑ ΤΗΣ ALFA

`

Οι λέξεις που σχετίζονται το ελάχιστο 50% με τη λέξη ‘θετικές’ απεικονίζονται ακολούθως:

```
> findAssocs(dtm, terms = "θετικές", corlimit = 0.5)
$θετικές
       fitch      βελτίωσή διαφαινόμενη       ζήτησης αβεβαιότητες    προσδοκιών
        0.70         0.70         0.70         0.70         0.66         0.59
 οικονομικού      ομόλογα     προϊόντα   συστημικών      κλίματος
        0.56         0.56         0.56         0.56         0.51
```

Το γράφημα των συχνότερα εμφανιζόμενων λέξεων παρουσιάζεται παρακάτω:



ΕΙΚΟΝΑ 4-5: ΓΡΑΦΙΚΗ ΑΠΕΙΚΟΝΙΣΗ ΤΩΝ ΣΥΧΝΟΤΕΡΑ ΕΜΦΑΝΙΖΟΜΕΝΩΝ ΛΕΞΕΩΝ ΓΙΑ ΤΗΝ ALFA

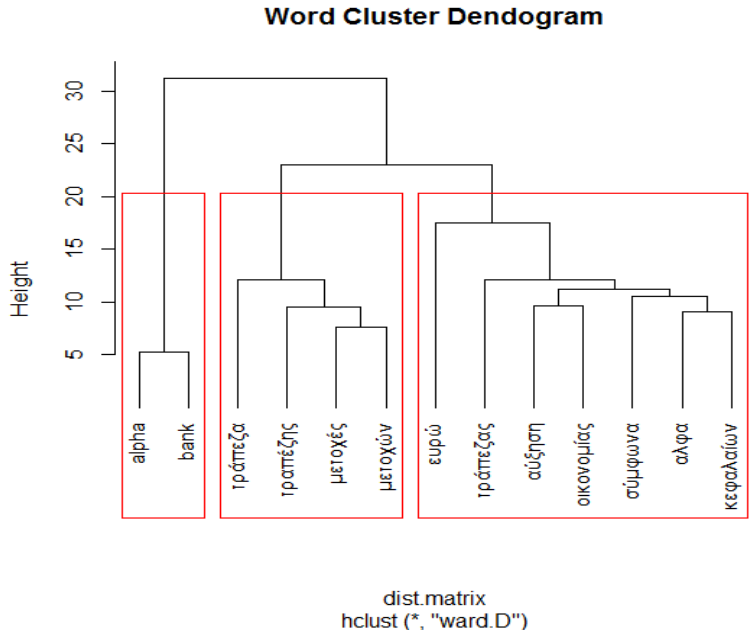
5^{0} Βήμα: Hierarchical Clustering

Αρχικά αφαιρούμε τους όρους που η συχνότητα εμφάνισής τους είναι μικρότερη από το κατώφλι που θεωρήσαμε (0.75), προκειμένου να αποφευχθεί η υπερπροσαρμογή. Ουσιαστικά, οι όροι αποτελούν *outliers* για τα δεδομένα μας (Janson & Gaur, 2015). Στη συνέχεια, υπολογίζουμε την απόσταση μεταξύ των λέξεων και στη συνέχεια ομαδοποιούμε βάσει ομοιότητας. Για καλύτερη κατανόηση των ομάδων, δημιουργούμε το δενδρόγραμμα ενώ επιλέγουμε τη δημιουργία τριών ομάδων (k=3) και την αναπαράσταση αυτών σε κόκκινα κουτιά. Τελικό βήμα είναι η εκτίμηση των όρων που εμπεριέχονται στις συστάδες χρησιμοποιώντας τη συνάρτηση *cutree()*. Ο κώδικας παρουσιάζεται ακολούθως (Liu, 2014):

```
#Hierachical Clustering
dtm2<-removeSparseTerms(dtm,sparse=0.75) #Remove Sparse Terms
m2<-as.matrix(dtm2)
dist.matrix<-dist(scale(m2)) # First calculate distance between words
fit<-hclust(dist.matrix,method="ward.D")
plot(fit, cex=0.9,hang=-1,main="Word Cluster Dendogram") #Plot the Dendrogram
rect.hclust(fit,k=3)  #Cut Tree
(groups<-cutree(fit,k=3)) # Draw dendogram with red borders around the 3 clusters
 (fit.groups<-cutree(fit,k=3)) #Evaluate the terms in clusters
```

`

Το δενδρόγραμμα που δημιουργήθηκε παρουσιάζεται ως εξής:



ΕΙΚΟΝΑ 4-6: ΑΠΕΙΚΟΝΙΣΗ ΔΕΝΔΡΟΓΡΑΜΜΑΤΟΣ ΧΡΗΣΙΜΟΠΟΙΩΝΤΑΣ ΙΕΡΑΡΧΙΚΗ ΣΥΣΤΑΔΟΠΟΙΗΣΗ

Παρακάτω παρουσιάζονται οι όροι που εμφανίζονται σε κάθε συστάδα:

```
> (fit.groups<-cutree(fit,k=3))
     alpha       bank       αλφα     αύξηση       ευρώ  κεφαλαίων     μετοχές
         1          1          2          2          2          2          3
  μετοχών οικονομίας    σύμφωνα    τράπεζα   τράπεζας   τραπέζης
         3          2          2          3          2          3
```

6⁰ Βήμα: K-means

Πρώτο βήμα είναι η μετάθεση του πίνακα μέσω της εντολής t() και στη συνέχεια η εφαρμογή της μεθόδου μέσω της εντολής *kmeans()*. Απαραίτητο στοιχείο είναι ο ορισμός του αριθμού συστάδων που θα δημιουργηθεί (στη προκειμένη *k*=3) καθώς και η επανάληψη της διαδικασίας προκειμένου να βρεθεί η βέλτιστη λύση που στη περίπτωσή μας ισούται με 10 (*nstarts*=10). Έπειτα στρογγυλοποιούμε τις αποστάσεις στα τρία δεκαδικά ψηφία χρησιμοποιώντας την εντολή *Round()* και δημιουργούμε τις τρεις συστάδες. Τέλος, μέσω ενός γραφήματος παρουσιάζουμε τον τρόπο κατανομής των δεδομένων στις συστάδες (Hahsler, 2016; Basic Text Mining in R).

```
#K-means
#Compute k-means
m3<-t(m2) #Transpose
k<-3
Res<-kmeans(m3,centers=3,nstart=10)
print(Res)
round(Res$centers,digits=3)[1:3,1:10]
for(i in 1:k) {
 cat(paste("cluster",i,":",seq=""))
 s=sort(Res$centers[i,],decreasing=TRUE)
 cat(names(s)[1:3],"\n")
}# Generate clusters
def.par <- par(no.readonly = TRUE) #Save default for resetting
layout(t(1:4)) # 3 plots in one
for(i in 1:4) barplot(kfit$centers[i,], ylim=c(0,40),
main=paste("Cluster", i))
Con.Matrix<-table(df$Category,Res$cluster)
sum(diag(Con.Matrix))/sum(Con.Matrix)
#Accuracy
```

`

Το αποτέλεσμα της εντολής kmeans() φαίνεται ακολούθως:

```
K-means clustering with 3 clusters of sizes 1, 60, 30

Cluster means:
  alpha     bank      αλφα     αύξηση       ευρώ κεφαλαίων   μετοχές    μετοχών
1  8.00 8.000000 0.000000 6.00000000 32.0000000 4.0000000 0.0000000 12.0000000
2  1.25 1.166667 0.400000 0.08333333  0.5333333 0.2166667 0.5833333  0.5333333
3  3.50 3.666667 1.033333 1.73333333  2.6333333 0.9000000 0.2666667  0.5666667
  οικονομίας   σύμφωνα   τράπεζα τράπεζας  τραπέζης
1        0.0 4.0000000 0.0000000     0.00 6.0000000
2        0.2 0.2166667 0.6666667     0.55 0.6500000
3        1.7 0.7000000 0.7000000     1.00 0.2666667
```

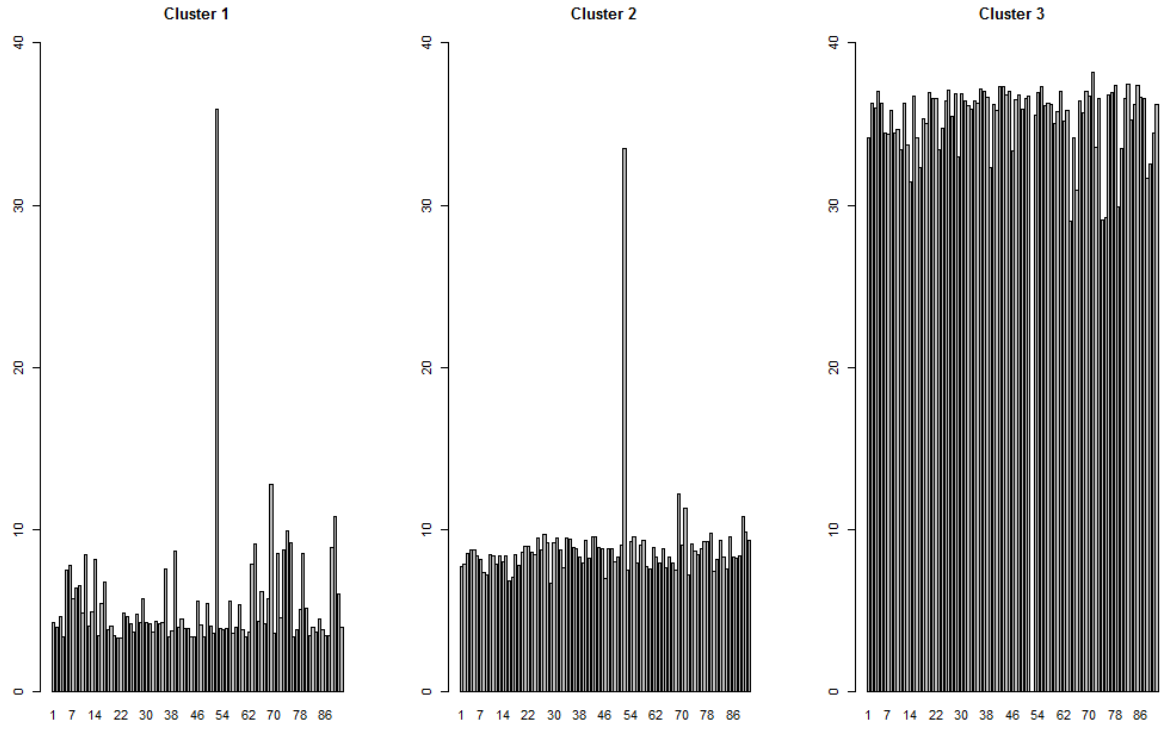
Το αποτέλεσμα της στρογγυλοποίησης είναι:

```
> round(Res$centers,digits=3)[1:3,1:10]
  alpha  bank  αλφα αύξηση   ευρώ κεφαλαίων μετοχές μετοχών οικονομίας σύμφωνα
1  8.00 8.000 0.000  6.000 32.000      4.000   0.000  12.000        0.0   4.000
2  1.25 1.167 0.400  0.083  0.533      0.217   0.583   0.533        0.2   0.217
3  3.50 3.667 1.033  1.733  2.633      0.900   0.267   0.567        1.7   0.700
```

Ακολούθως παρουσιάζονται οι συστάδες που δημιουργήθηκαν:

```
cluster 1 : ευρώ μετοχών alpha
cluster 2 : alpha bank τράπεζα
cluster 3 : bank alpha ευρώ
```

Ο τρόπος με τον οποίο ομαδοποιήσαμε τα δεδομένα μας παρουσιάζεται γραφικά:



ΕΙΚΟΝΑ 4-7: ΑΠΕΙΚΟΝΙΣΗ ΟΜΑΔΟΠΟΙΗΣΗΣ ΤΩΝ ΚΕΙΜΕΝΙΚΩΝ ΔΕΔΟΜΕΝΩΝ ΜΕΣΩ K-MEANS

`

Τελικό βήμα είναι ο υπολογισμός της εγκυρότητας του μοντέλου. Υπολογίζουμε το πίνακα σύγχυσης και την ακρίβεια του μοντέλου. Τα αποτελέσματα παρουσιάζονται ως εξής:

```
> print(Con.Matrix)

            1  2  3
  Bad       3 19  0
  Good     14 21  0
  No Mover 13 20  1
> sum(diag(Con.Matrix))/sum(Con.Matrix) #Accuracy
[1] 0.2747253
```

Ουσιαστικά, παρατηρούμε ότι 25 παρατηρήσεις από τις 91 ομαδοποιήθηκαν σωστά και 66 λανθασμένα ενώ η ακρίβεια είναι μόλις 27%.

Ένας τρόπος σύγκρισης των ανωτέρω αλγορίθμων συσταδοποίησης είναι η χρήση του πακέτου *clValid* που προσφέρεται από την R. Το πακέτο μας επιτρέπει να αξιολογήσουμε ταυτόχρονα τις μεθόδους συσταδοποίησης και παράλληλα να προσδιορίσουμε τη κατάλληλη μέθοδο αλλά και τον βέλτιστο αριθμό των συστάδων για το σύνολο των δεδομένων Το πακέτο επίσης προσφέρει τρεις τύπους επικύρωσης ("*internal*", "*stability*" και "*biological*") προκειμένου να αξιολογήσει τις συστάδες που δημιούργησαν οι αλγόριθμοι hierarchical clustering και k-means. Εμείς θα χρησιμοποιήσουμε τον τύπο "*internal*", ο οποίος υπολογίζει βάσει του συνόλου δεδομένων και τον διαχωρισμό των συστάδων που δημιουργήθηκαν και χρησιμοποιεί εσωτερικές πληροφορίες στα δεδομένα ώστε να αποδώσει τη ποιότητα της συσταδοποίησης.

Ο τύπος "*internal*" χρησιμοποιεί μέτρα που αφορούν τη πυκνότητα, τη συνεκτικότητα και το διαχωρισμό των διακριθέντων συστάδων. Ειδικότερα, η συνεκτικότητα αφορά το κατά πόσο οι παρατηρήσεις τοποθετούνται στην συστάδα που τοποθετούνται οι γείτονές τους, η πυκνότητα αφορά την ομοιογένεια της συστάδας παρατηρώντας τη διασπορά εντός των συστάδων ενώ ο διαχωρισμός ποσοτικοποιεί το βαθμό διαχωρισμού μεταξύ των συστάδων, υπολογίζοντας την απόσταση. Αφού η πυκνότητα παρουσιάζει αντίθετες τάσεις από το διαχωρισμό (όσο ο αριθμός των συστάδων αυξάνει, η πυκνότητα αυξάνεται και ο διαχωρισμός μειώνεται) η τελική απόφαση λαμβάνεται βάσει ενός score που αποτελεί των συνδυασμό και των δύο. Τέλος, για τη πυκνότητα και το διαχωρισμό χρησιμοποιούνται τα μέτρα *Dunn Index* και το *Silhuette Width* αντίστοιχα (Brock, Pihur, Datta, & Datta, 2008). Ο κώδικας που χρησιμοποιείται παρατίθεται ακολούθως (Validate cluster analysis in R, 2013):

```
#Required Package
library(clValid)
intern<- clValid(m3,3:10, clMethods = c("hierarchical", "kmeans"),  validation = "internal")
summary(intern)
# Show Only Optimal Solutions
optimalScores(intern)
```

Το output που παίρνουμε είναι:

```
> summary(intern)

Clustering Methods:
 hierarchical kmeans

Cluster sizes:
 3 4 5 6 7 8 9 10

Validation Measures:
                                  3       4       5       6       7       8       9      10

hierarchical Connectivity    5.8579  8.7869 11.7159 22.9952 29.7460 30.4611 33.5218 36.4508
             Dunn            0.7016  0.5885  0.5143  0.4137  0.3797  0.4020  0.4020  0.4020
             Silhouette      0.5328  0.4611  0.3625  0.3535  0.3424  0.3330  0.2944  0.2754
kmeans       Connectivity   31.8131 36.7532 50.5226 51.8643 55.4750 55.6750 63.5758 75.5377
             Dunn            0.0851  0.1034  0.1034  0.1355  0.1374  0.1543  0.1667  0.1667
             Silhouette      0.3142  0.3139  0.2917  0.3078  0.3079  0.3082  0.2519  0.1772

Optimal Scores:

             Score  Method       Clusters
Connectivity 5.8579 hierarchical 3
Dunn         0.7016 hierarchical 3
Silhouette   0.5328 hierarchical 3

> optimalScores(intern)
                 Score       Method Clusters
Connectivity 5.8579365 hierarchical        3
Dunn         0.7015896 hierarchical        3
Silhouette   0.5328499 hierarchical        3
```

Υπενθυμίζουμε ότι η συνδεσιμότητα θα έπρεπε να ελαχιστοποιηθεί ενώ η πυκνότητα (*Dunn*) και ο διαχωρισμός (*Silhuette*) θα έπρεπε να μεγιστοποιηθούν προκειμένου για βέλτιστη λύση. Συνεπώς, παρατηρώντας το output, καταλήγουμε ότι η ιεραρχική συσταδοποίηση με τρεις συστάδες δίνει το καλύτερο αποτέλεσμα. Συμπεραίνουμε, επίσης, ότι ο k-means είναι χειρότερος για το συγκεκριμένο σύνολο δεδομένων και ότι η επιλογή τριών συστάδων που χρησιμοποιήσαμε εξ' αρχής είναι η καλύτερη δυνατή.

6.1.4 ΚΑΤΗΓΟΡΙΟΠΟΙΗΤΗΣ KNN ΣΕ ΚΕΙΜΕΝΙΚΑ ΔΕΔΟΜΕΝΑ

1^0 Βήμα: Install & Load Required Packages

Πρώτο βήμα είναι η εγκατάσταση των κατάλληλων πακέτων για τη χρήση του κατηγοριοποιητή.

```
install.packages("tm")  # For text mining
install.package("SnowballC") # For text stemming
install.packages("class") # KNN model
library(tm) # Load
library(class)
library(SnowballC)
```

2^0 Βήμα: Create Corpus

Η διαδικασία δημιουργίας σώματος και DTM είναι η ίδια με τις προηγούμενες μεθόδους που αφορούν κειμενικά δεδομένα (βλ. ενότητα 6.1.3).

3^0 Βήμα: Create KNN model

Προκειμένου να χρησιμοποιηθεί σωστά ο κατηγοριοποιητής, δημιουργούμε ένα αρχείο που περιλαμβάνει δύο στήλες. Η μία είναι η στήλη που περιέχει το κείμενο και ονομάζεται Text και η δεύτερη τη κατηγορία του κειμένου και ονομάζεται Category.

Επαναλαμβάνοντας τη διαδικασία καθαρισμού του κειμένου και τη δημιουργία DTM όπως και στις προηγούμενες μεθόδους, δημιουργούμε ένα data frame και μια νέα στήλη με τη

`

γνωστή εκ των προτέρων κατηγοριοποίηση, μέσω της εντολής *cbind()* και μετονομάζουμε τη τελευταία στήλη σε 'category'. Το μοντέλο KNN απαιτεί τρία σύνολα δεδομένων: Τα δεδομένα εκπαίδευσης, ελέγχου και τον κατηγοριοποιητή. Για τη δημιουργία τους χρησιμοποιούμε ένα τυχαίο δείγμα με μέγεθος ίσο με το 75% του συνόλου δεδομένων και το ονομάζουμε train, ενώ test ονομάζουμε το υπόλοιπο. Για τη δημιουργία του κατηγοριοποιητή απομονώνουμε τη τελευταία στήλη και την ονομάζουμε *cl.* Αφού καθορίσαμε τα δεδομένα μας, δημιουργούμε ένα *data frame* χωρίς τη στήλη category και τρέχουμε το μοντέλο (Garonfolo, 2015). Ο κώδικας που χρησιμοποιήθηκε παρατίθενται ακολούθως:

```
set.seed(100) # Set seed for reproducible results
# Transform dtm to matrix to data frame - df is easier to work with
    mat.df <- as.data.frame(data.matrix(dtm), stringsAsfactors = FALSE)
# Column bind category (known classification)
    mat.df <- cbind(mat.df, df$Category)
# Change name of new column to "category"
    colnames(mat.df)[ncol(mat.df)] <- "category"
# Split data by rownumber into portions
    train <- sample(nrow(mat.df), ceiling(nrow(mat.df) * .75))
    test <- (1:nrow(mat.df))
# Isolate classifier
    cl <- mat.df[, "category"]
# Create model data and remove "category"
    modeldata <- mat.df[,!colnames(mat.df) %in% "category"]
# Create model: training set, test set, training set classifier
    knn.pred <- knn(modeldata[train, ], modeldata[test, ], cl[train])
```

4^{0} Βήμα: Compute Accuracy

Προκειμένου να ελέγξουμε την εγκυρότητα του κατηγοριοποιητή χρησιμοποιούμε τη μήτρα σύγχυσης και υπολογίζουμε τη ποσότητα μέσω των εντολών:

```
conf.mat <- table("Predictions" = knn.pred, Actual = cl[test]) # Confusion matrix
(accuracy <- sum(diag(conf.mat))/length(test) * 100) # Accuracy
```

Το αποτέλεσμα της μήτρας σύγχυσης καθώς και η εγκυρότητα του μοντέλου παρουσιάζονται ως εξής:

```
> conf.mat
            Actual
Predictions Bad Good No Mover
   Bad       17    3        2
   Good       1   32        2
   No Mover   4    0       30
>
> (accuracy <- sum(diag(conf.mat))/length(test) * 100)
[1] 86.81319
```

Δηλαδή, 79 (=17+32+30) παρατηρήσεις από τις 91 κατηγοριοποιήθηκαν ορθά στις κατηγορίες Bad, Good και No Mover αντίστοιχα, ενώ 12 παρατηρήσεις από τις 91

`

κατηγοριοποιήθηκαν λανθασμένα. Επιπρόσθετα, η ακρίβεια της πρόβλεψης ισούται με 86.81%.

Ο συγκεντρωτικός πίνακας αποτελεσμάτων κατηγοριοποίησης και συσταδοποίηση κειμενικών δεδομένων για κάθε μετοχή παρουσιάζεται στον επόμενο πίνακα:

Όνομα Μετοχής	Optimal Solution	Accuracy
	Hier. Clustering / K-means	KNN
Alfa	Hier. Clustering	86.81%
Attica	Hier. Clustering	83.33%
ETE	Hier. Clustering	89.41%
ΤτΕ	Hier. Clustering	83.05%
Kyp	Hier. Clustering	100%
Peir	Hier. Clustering	87.80%
Eur	Hier. Clustering	81.63%

ΠΙΝΑΚΑΣ 4-3: ΣΥΓΚΕΝΤΡΩΤΙΚΟΣ ΠΙΝΑΚΑΣ ΑΠΟΤΕΛΕΣΜΑΤΩΝ ΚΕΙΜΕΝΙΚΩΝ ΔΕΔΟΜΕΝΩΝ

Δεδομένου ότι χρησιμοποιούνται δύο διαφορετικές τεχνικές εξόρυξης δεδομένων, δεν είναι δυνατή η σύγκριση μεταξύ των αλγορίθμων συσταδοποίησης και κατηγοριοποίησης. Συνεπώς, παρουσιάζεται η καλύτερη απόδοση για την επιλογή συσταδοποίησης και η εγκυρότητα για τον αλγόριθμο KNN.

Διαπιστώνουμε ότι για όλες τις περιπτώσεις καλύτερη απόδοση χρησιμοποιώντας τεχνικές συσταδοποίησης επιτυγχάνεται με την ιεραρχική συσταδοποίηση. Ακόμη, Για τον αλγόριθμο KNN παρατηρούμε ότι η μέγιστη ακρίβεια επιτυγχάνεται για τη μετοχή της Κύπρου, η οποία περιέχει τέσσερα αρχεία. Συνεπώς, ο αλγόριθμος είναι πιο αποδοτικός όσο το μικρότερο είναι το σύνολο δεδομένων, ενώ αποδίδει εξαιρετικά αποτελέσματα για όλες τις μετοχές, αφού το ποσοστό ακρίβειας είναι άνω των 80% για κάθε περίπτωση.

6.1.5 ΚΑΤΗΓΟΡΙΟΠΟΙΗΤΗΣ KNN ΣΕ ΥΒΡΙΔΙΚΑ ΔΕΔΟΜΕΝΑ

Η κατηγοριοποίηση υβριδικών δεδομένων απαιτεί τη δημιουργία μιας νέας κατηγορικής μεταβλητής με τέσσερα επίπεδα, βάσει ενός κανόνα που περιγράφηκε λεπτομερώς σε προηγούμενη ενότητα (βλ. ενότητα 4.1.3):

Για τη κατηγοριοποίηση των υβριδικών δεδομένων χρησιμοποιούμε τον αλγόριθμο KNN, ακολουθώντας την ίδια διαδικασία που χρησιμοποιήσαμε σε προηγούμενη ενότητα (βλ. ενότητα 6.1.4). Συνεπώς, τα αποτελέσματα του κατηγοριοποιητή για το συγκεκριμένο σύνολο δεδομένων έπονται ακολούθως:

```
> # Confusion matrix
> conf.mat <- table("Predictions" = knn.pred, Actual = cl[test])
> conf.mat
           Actual
Predictions  0  1  2  3
          0 22  1  1  0
          1  2 26  4  2
          2  2  2  8  0
          3  2  0  0 19
>
> # Accuracy
> (accuracy <- sum(diag(conf.mat))/length(test) * 100)
[1] 82.41758
```

Ο πίνακας σύγχυσης υποδεικνύει ότι 22 από τις 91 παρατηρήσεις κατηγοριοποιήθηκαν και στις δύο περιπτώσεις κατηγοριοποίησης που προηγήθηκαν στη κατηγορία 'Good'. Ομοίως, 26 από τις 91 κατηγοριοποιήθηκαν στη κατηγορία 'Bad', 8 δεδομένα κατηγορίας 'No Mover' κατηγοριοποιήθηκαν στη κατηγορία 'Good' και 19 στη κατηγορία 'Bad'. Επιπλέον, η εγκυρότητα του αλγορίθμου ισούται με 82.41%.

Ο συγκεντρωτικός πίνακας για τη κάθε μετοχή παρουσιάζεται ακολούθως:

Όνομα Μετοχής	**Accuracy**
	KNN
Alfa	82.41%
Attica	94.44%
ETE	89.41%
ΤτΕ	83.05%
Kyp	100%
Peir	87.80%
Eur	81.60%

ΠΙΝΑΚΑΣ 4-4: ΣΥΓΚΕΝΤΡΩΤΙΚΟΣ ΠΙΝΑΚΑΣ ΑΠΟΤΕΛΕΣΜΑΤΩΝ ΚΕΙΜΕΝΙΚΩΝ ΔΕΔΟΜΕΝΩΝ

Παρατηρούμε ότι η μέθοδος παρουσιάζει αρκετά καλή επίδοση για κάθε μετοχή και ότι για τη μετοχή της Κύπρου επιτυγχάνει 100% ακρίβεια.

6.1.6 ΣΥΓΚΡΙΣΗ ΜΕΘΟΔΩΝ ΚΑΙ ΑΠΟΤΕΛΕΣΜΑΤΩΝ

Όσον αφορά τα αριθμητικά δεδομένα, η καλύτερη μέθοδος κατηγοριοποίησης για όλες τις μετοχές είναι η χρήση του αλγορίθμου CART, που παρουσιάζει ποσοστό εγκυρότητας μεταξύ 60-93% για το σύνολο μετοχών. Επίσης, η πρώτη περίπτωση QDA που εφαρμόστηκε έχει ποσοστό εγκυρότητας μεταξύ 44-62% ενώ η δεύτερη μεταξύ 42-63%. Η μέθοδος KNN, τέλος, υποδείχθηκε η λιγότερο αποδοτική για τα αριθμητικά δεδομένα, αφού το ποσοστό εγκυρότητας κυμαίνεται μεταξύ 40-50% για κάθε μετοχή.

Τα κειμενικά δεδομένα απαιτούν ιδιαίτερη προσέγγιση όσον αφορά την επεξεργασία τους για την χρήση του κατάλληλου αλγορίθμου. Γίνεται εύκολα κατανοητό ότι λανθασμένη προ-επεξεργασία μπορεί να οδηγήσει σε λανθασμένη κατηγοριοποίηση και συνεπώς σε λανθασμένη κατηγοριοποίηση. Δεδομένου, λοιπόν, ότι τα δεδομένα εκπαίδευσης που χρησιμοποιήθηκαν είχαν την κατάλληλη μορφή για την R, διαπιστώνουμε ότι η καλύτερη μέθοδος συσταδοποίησης είναι η ιεραρχική συσταδοποίηση ενώ παράλληλα η μέθοδος κατηγοριοποίησης KNN που χρησιμοποιήθηκε παρουσίασε αρκετά υψηλή απόδοση που κυμαίνεται από 81 έως 100%.

`

Υπενθυμίζουμε ότι οι μέθοδοι συσταδοποίησης δε μπορούν να εφαρμοστούν σε κατηγορικές μεταβλητές απόκρισης και συνεπώς, ούτε στα υβριδικά δεδομένα. Για τα υβριδικά δεδομένα χρησιμοποιήθηκε ο κατηγοριοποιητής KNN, του οποίου η εγκυρότητα ήταν επίσης σε πολύ καλά επίπεδα. Συγκεκριμένα, η επίδοση κυμαινόταν μεταξύ του 81 έως 100%.

Προκειμένου για πρόβλεψη των χρηματιστηριακών δεικτών, διαπιστώνουμε ότι καλύτερη μέθοδος για τη κατηγοριοποίηση των κειμενικών δεδομένων αποδοτικότερη μέθοδος κατηγοριοποίησης υποδεικνύεται η χρήση αλγορίθμου KNN ενώ για τη μέθοδο συσταδοποίησης ο αλγόριθμος ιεραρχικής συσταδοποίησης. Στον ίδιο βαθμό απόδοσης συγκαταλέγεται και η κατηγοριοποίηση υβριδικών δεδομένων με τον αλγόριθμο KNN. Συνεπώς, συνίσταται η αξιολόγηση των κειμενικών δεδομένων μεμονωμένα ή ο συνδυασμός τους με τα αριθμητικά προκειμένου για καλύτερα αποτελέσματα.

6.2 ΜΕΛΛΟΝΤΙΚΕΣ ΕΠΕΚΤΑΣΕΙΣ

Το διευρυμένο σύστημα που παρουσιάστηκε στις προηγούμενες ενότητες είναι πρωτότυπο και αποσκοπεί στη προσφορά συγκεκριμένης λειτουργικότητας στους χρήστες. Σαφώς, λοιπόν, κατά τη δημιουργία του να υλοποιήθηκαν και διαμορφώθηκαν ιδέες επέκτασης του συστήματος (π.χ. χρήση ελληνικής γλώσσας, χρήση διαφορετικών ειδών δεδομένων).

Μια πρώτη επέκταση του συστήματός μας θα μπορούσε να θεωρηθεί η προσαρμογή του συστήματος και σε άλλες γλώσσες ή ακόμη και σε υβριδικές γλώσσες όπως τα *greeklish*, χωρίς να μειώνεται η απόδοσή του. Αυτό θα μπορούσε να επιτευχθεί με τη δημιουργία μιας νέας λίστας λέξεων που θα περιλαμβάνει τις κοινές μεταξύ ελληνικών και αγγλικών και με τη δημιουργία ενός υποπρογράμματος, το οποίο θα μετατρέπει τα *greeklish* σε ελληνικά.

Μια δεύτερη προοπτική εξέλιξης θα μπορούσε να θεωρηθεί η χρήση λέξεων ή φράσεων ευρέως χρησιμοποιούμενων που είναι ικανές να επηρεάσουν τις τιμές των μετοχών. Ουσιαστικά, προτείνεται η χρήση των κοινών λέξεων προκειμένου για εξαγωγή αποτελέσματος, υπό την προϋπόθεση ότι οι λέξεις αυτές θα μπορούσαν να οριστούν από ειδήμονες ώστε να μη μειωθεί η ευελιξία του συστήματός μας.

Τέλος, η διεύρυνση του συστήματος μας δίνει τη δυνατότητα μετατροπής του σε δυναμικό. Αυτό συνεπάγεται ότι του δίνει τη δυνατότητα να εναλλάσσεται μεταξύ των διαφορετικών δεδομένων που εισάγονται και να επιλέγει ανά περίσταση τη καταλληλότερη κατηγορία δεδομένων ή το συνδυασμό τους, ελαχιστοποιώντας το σφάλμα της κατηγοριοποίησης. Αυτό θα μπορούσε να επιτευχθεί με την εισαγωγή ενός αλγορίθμου που με τα κατάλληλα *vibes* υποδυκνείει τα βέλτιστα δεδομένα εκπαίδευσης που πρέπει να ληφθούν υπόψη.

`

ΠΑΡΑΡΤΗΜΑ

ΔΕΔΟΜΕΝΑ ΕΚΠΑΙΔΕΥΣΗΣ

Π.1 ΠΙΝΑΚΕΣ ΚΕΙΜΕΝΙΚΩΝ ΔΕΔΟΜΕΝΩΝ ΕΚΠΑΙΔΕΥΣΗΣ

Στους παρακάτω πίνακες, φαίνονται οι ημέρες και ώρες έκδοσης των δελτίων τύπου καθώς και η κατηγοριοποίησή τους σε 'Good', 'Bad' και 'No Movers' για κάθε τραπεζική μετοχή:

- Η μετοχή της Alphabank:

	Ημερομηνία	***Ώρα***	***Δελτίο Τύπου/ Άρθρο***	***Κατηγορία***
1.	09/01/2014	13:58:37	**JPMorgan: Ξεκινά εκ νέου την κάλυψη τριών ελληνικών τραπεζών**	No Mover
2.	24/01/2014	12:08:42	**Alpha User: Η καθυστέρηση των δόσεων επηρεάζει την ανάκαμψη**	No Mover
3.	24/01/2014	15:51:31	**ALPHA ΤΡΑΠΕΖΑ Α.Ε. : Νέα επιτόκια από την Alpha Bank**	Bad
4.	24/01/2014	16:09:00	**Μείωση επιτοκίων από την Alpha Bank**	Bad
5.	31/01/2014	10:55:49	**Αντικατάσταση μέλους στο δ.σ. της Alpha Bank**	No Mover
6.	31/01/2014	12:42:08	**Alpha Bank: Χώρα προσέλκυσης κεφαλαίων η Ελλάδα**	Good
7.	03/02/2014	10:32:49	**ALPHA ΤΡΑΠΕΖΑ Α.Ε. : Διαχείριση των Αμοιβαίων Κεφαλαίων "ΕΡΜΗΣ" από την Alpha Asset ManageMent ΑΕΔΑΚ**	Good
8.	07/02/2014	14:14:54	**Alpha Bank: Η βελτίωση του κλίματος επιβεβαιώνει την ανάκαμψη**	No Mover
9.	14/02/2014	13:27:04	**Alpha Bank: Με φειδώ και περίσκεψη να διατεθεί το πρωτογενές πλεόνασμα**	No Mover
10.	21/02/2014	11:04:00	**Reuters: Νέα κεφάλαια 5 δισ. χρειάζονται οι συστημικές τράπεζες**	No Mover
11.	21/02/2014	14:05:03	**Alpha Bank: Αλήθειες και ψέματα για την υπερφορολόγηση**	Good
12.	06/03/2014	13:34:34	**Alpha Bank: Πληθαίνουν οι ενδείξεις ανάκαμψης της ελληνικής οικονομίας**	Good
13.	11/03/2014	11:51:23	**Ξεκινά κάλυψη για τις ελληνικές τράπεζες η Wood**	Good
14.	13/03/2014	13:11:23	**Alpha Bank: Οι συνεχείς αναθεωρήσεις του ΑΕΠ εντείνουν την αβεβαιότητα**	No Mover
15.	18/03/2014	10:46:27	**S&P: Επιβεβαιώνει το «CCC/C» για Πειραιώς, Alpha**	No Mover
16.	21/03/2014	15:13:00	**Alpha Bank: Ανατρέπεται το μεταπολιτευτικό οικονομικό μοντέλο των ελλειμμάτων**	No Mover
17.	26/03/2014	11:14:00	**Με άνοδο και εντυπωσιακό τζίρο ξεκίνησε το Χ.Α.**	No Mover
18.	26/03/2014	13:04:37	**Πήγασος: Σύναψη ενυπόθηκου ομολογιακού 80 εκατ. ευρώ**	Bad
19.	27/03/2014	12:54:37	**BlooMberg: Ανακάμπτουν οι**	No Mover

`

			ελληνικές τράπεζες	
20.	28/03/2014	13:14:33	**ΤΧΣ: «Πράσινο» για τις ΑΜΚ σε Alpha και Πειραιώς**	Good
21.	28/03/2014	15:30:32	**ALPHA ΤΡΑΠΕΖΑ Α.Ε. : ΑΝΑΚΟΙΝΩΣΗ ΓΙΑ ΤΙΣ ΑΠΟΦΑΣΕΙΣ ΓΕΝΙΚΗΣ ΣΥΝΕΛΕΥΣΗΣ**	Bad
22.	28/03/2014	15:31:15	**ALPHA ΤΡΑΠΕΖΑ Α.Ε. : ΑΝΑΚΟΙΝΩΣΗ ΓΙΑ ΤΙΣ ΑΠΟΦΑΣΕΙΣ ΓΕΝΙΚΗΣ ΣΥΝΕΛΕΥΣΗΣ**	Bad
23.	28/03/2014	17:00:12	**Alpha Bank: «Πράσινο» από Γ.Σ. για την ΑΜΚ**	No Mover
24.	31/03/2014	17:19:00	**Γ. Προβόπουλος: Οι αγορές ξανανοίγουν για την Ελλάδα**	No Mover
25.	01/04/2014	14:03:59	**ALPHA ΤΡΑΠΕΖΑ Α.Ε. : Γνωστοποίηση σημαντικών αλλαγών στα δικαιώματα ψήφου σύμφωνα με τον Ν. 3556/2007**	Bad
26.	02/04/2014	10:40:01	**Fitch: Αναβάθμιση προοπτικών για ελληνικά καλυμμένα ομόλογα**	No Mover
27.	02/04/2014	10:40:01	**ALPHA ΤΡΑΠΕΖΑ Α.Ε. : Εισαγωγή προς διαπραγμάτευση των νέων μετοχών που προέκυψαν από την ολοκληρωθείσα Αύξηση Μετοχικού Κεφαλαίου με καταβολ**	Bad
28.	03/04/2014	11:23:00	**Εισαγωγή μετοχών από ΑΜΚ με ιδιωτική τοποθέτηση (ΑΛΦΑ)**	Good
29.	04/04/2014	16:30:00	**Γ. Κωστόπουλος: Επωφελείς για την οικονομία οι ΑΜΚ της Alpha Bank**	Good
30.	04/04/2014	17:01:00	**Εισαγωγή μετοχών από ΑΜΚ με ιδιωτική τοποθέτηση (ΑΛΦΑ)**	No Mover
31.	08/04/2014	13:35:41	**ALPHA ΤΡΑΠΕΖΑ Α.Ε. : Γνωστοποίηση σημαντικών αλλαγών στα δικαιώματα ψήφου σύμφωνα με τον Ν. 3556/2007 [8.4.2014]**	Good
32.	08/04/2014	16:16:24	**ALPHA ΤΡΑΠΕΖΑ Α.Ε. : Γνωστοποίηση σημαντικών αλλαγών στα δικαιώματα ψήφου σύμφωνα με τον Ν. 3556/2007 [8.4.2014]**	Good
33.	10/04/2014	13:03:00	**Alpha Bank: Κλείνει ένας επώδυνος κύκλος προσαρμογής**	Bad
34.	10/04/2014	16:30:39	**ALPHA ΤΡΑΠΕΖΑ Α.Ε. : ΑΝΑΚΟΙΝΩΣΗ ΡΥΘΜΙΖΟΜΕΝΗΣ ΠΛΗΡΟΦΟΡΙΑΣ ΤΟΥ Ν. 3556/2007: Γνωστοποίηση για μεταβολή ποσοστού μετόχων σε επίπεδο δικαιωμάτων**	Good
35.	10/04/2014	16:31:29	**ALPHA ΤΡΑΠΕΖΑ Α.Ε. : ΑΝΑΚΟΙΝΩΣΗ ΡΥΘΜΙΖΟΜΕΝΗΣ ΠΛΗΡΟΦΟΡΙΑΣ ΤΟΥ Ν. 3556/2007: Γνωστοποίηση για μεταβολή ποσοστού μετόχων σε επίπεδο δικαιωμάτων**	Good
36.	17/04/2014	13:02:00	**Alpha Bank: Η επάνοδος στις αγορές πιστοποιεί την ανάκαμψη της Ελλάδος**	No Mover

`

37.	24/04/2014	14:31:12	**Alpha Bank: Άκαρπη η πρώτη συνάντηση εργαζομένων - διοίκησης**	Bad
38.	25/04/2014	11:00:00	**Alpha Bank: Διευκρινίσεις για το ενδεχόμενο συμφωνίας με Citi**	Good
39.	25/04/2014	11:41:00	**Alpha Bank: Διαψεύστηκαν οι προφήτες της κατάρρευσης της Ελλάδας**	Good
40.	30/04/2014	12:50:51	**ALPHA ΤΡΑΠΕΖΑ Α.Ε. : ΑΝΑΚΟΙΝΩΣΗ ΟΙΚΟΝΟΜΙΚΟΥ ΗΜΕΡΟΛΟΓΙΟΥ**	No Mover
41.	30/04/2014	16:16:17	**Με 66,9% το ΤΧΣ στην Alpha Bank**	Good
42.	06/05/2014	15:46:26	**ALPHA ΤΡΑΠΕΖΑ Α.Ε. : ΑΝΑΚΟΙΝΩΣΗ ΓΙΑ ΤΗΝ ΠΡΟΑΝΑΓΓΕΛΙΑ ΓΕΝΙΚΗΣ ΣΥΝΕΛΕΥΣΗΣ**	No Mover
43.	06/05/2014	15:47:34	**ALPHA ΤΡΑΠΕΖΑ Α.Ε. : ΑΝΑΚΟΙΝΩΣΗ ΓΙΑ ΤΗΝ ΠΡΟΑΝΑΓΓΕΛΙΑ ΓΕΝΙΚΗΣ ΣΥΝΕΛΕΥΣΗΣ**	Bad
44.	08/05/2014	15:28:33	**ALPHA ΤΡΑΠΕΖΑ Α.Ε. : Ανακοίνωση Αποτελεσμάτων α΄ τριμήνου 2014, την 29 Μαΐου 2014**	No Mover
45.	08/05/2014	15:32:00	**Alpha Bank: Στις 29/5 τα αποτελέσματα α' τριμήνου**	No Mover
46.	15/05/2014	14:22:14	**Alpha Bank: Συνεχίζεται η εντυπωσιακή πρόσβαση στις αγορές**	No Mover
47.	16/05/2014	16:33:07	**ALPHA ΤΡΑΠΕΖΑ Α.Ε. : Ορθή επανάληψη: Γνωστοποίηση σημαντικών αλλαγών στα δικαιώματα ψήφου σύμφωνα με τον Ν.3864/2010 [16.5.2014]**	Good
48.	16/05/2014	16:39:01	**ALPHA ΤΡΑΠΕΖΑ Α.Ε. : Ορθή επανάληψη: Γνωστοποίηση σημαντικών αλλαγών στα δικαιώματα ψήφου σύμφωνα με τον Ν.3864/2010 [16.5.2014]**	Good
49.	22/05/2014	16:25:15	**Ανάπτυξη 0,6% στο β' τρίμηνο βλέπει η Alpha Bank**	Good
50.	29/05/2014	13:24:00	**Παραιτείται ο Γ. Κωστόπουλος από την Alpha Bank**	No Mover
51.	29/05/2014	17:21:51	**ALPHA ΤΡΑΠΕΖΑ Α.Ε. : ΑΝΑΚΟΙΝΩΣΗ ΓΙΑ ΤΙΣ ΑΠΟΦΑΣΕΙΣ ΓΕΝΙΚΗΣ ΣΥΝΕΛΕΥΣΗΣ**	No Mover
52.	29/05/2014	17:28:52	**ALPHA ΤΡΑΠΕΖΑ Α.Ε. : ΑΝΑΚΟΙΝΩΣΗ ΠΕΡΙ ΣΧΟΛΙΑΣΜΟΥ ΟΙΚΟΝΟΜΙΚΩΝ ΚΑΤΑΣΤΑΣΕΩΝ/ ΕΚΘΕΣΕΩΝ**	No Mover
53.	02/06/2014	12:29:00	**Υπεγράφη νέα επιχειρησιακή σύμβαση στην Alpha Bank**	Good
54.	06/06/2014	16:53:08	**ALPHA ΤΡΑΠΕΖΑ Α.Ε. : ΠΡΟΣΚΛΗΣΗ ΣΕ ΤΑΚΤΙΚΗ ΓΕΝΙΚΗ ΣΥΝΕΛΕΥΣΗ ΤΩΝ ΜΕΤΟΧΩΝ**	Good
55.	06/06/2014	17:24:43	**ALPHA ΤΡΑΠΕΖΑ Α.Ε. : ΑΝΑΚΟΙΝΩΣΗ ΓΙΑ ΤΗΝ ΠΡΟΑΝΑΓΓΕΛΙΑ ΓΕΝΙΚΗΣ ΣΥΝΕΛΕΥΣΗΣ**	Good
56.	13/06/2014	15:01:56	**Alpha Bank: Σαφείς ενδείξεις**	No Mover

`

			ότι η οικονομία ανακάμπτει	
57.	18/06/2014	16:11:00	ALPHA ΤΡΑΠΕΖΑ Α.Ε. : Γνωστοποίηση σημαντικών αλλαγών στα δικαιώματα ψήφου σύμφωνα με τον Ν. 3556/2007 [18.6.2014]	Bad
58.	30/06/2014	15:59:56	ALPHA ΤΡΑΠΕΖΑ Α.Ε. : ΓΝΩΣΤΟΠΟΙΗΣΗ ΣΗΜΑΝΤΙΚΩΝ ΑΛΛΑΓΩΝ ΣΤΑ ΔΙΚΑΙΩΜΑΤΑ ΨΗΦΟΥ ΣΥΜΦΩΝΑ ΜΕ ΤΟΝ Ν. 3864/2010	No Mover
59.	30/06/2014	16:48:53	Με 66,36% το ΤΧΣ στην Alpha Bank	No Mover
60.	03/07/2014	15:00:40	Alpha Bank: Στα 162,1 δισ. οι καταθέσεις τον Μάιο	Good
61.	09/07/2014	14:32:46	ALPHA ΤΡΑΠΕΖΑ Α.Ε. : Ανακοίνωση	Bad
62.	11/07/2014	14:58:40	Alpha Bank: Έτος καμπής για τα στεγαστικά δάνεια το 2015	Bad
63.	21/08/2014	12:50:00	Alpha Bank: Αύξηση του ΑΕΠ 1,5% - 2% στο β' εξάμηνο	Good
64.	26/08/2014	13:36:44	Θετικές συστάσεις Deutsche Bank για τις ελληνικές τράπεζες	Good
65.	27/08/2014	11:49:41	Συστάσεις από Citi για τις τραπεζικές μετοχές	No Mover
66.	28/08/2014	17:27:52	ALPHA ΤΡΑΠΕΖΑ Α.Ε. : Δελτίου Τύπου Αποτελεσμάτων Εξαμήνου 2014	No Mover
67.	29/08/2014	16:38:00	Alpha Bank: Απαιτείται επαναφορά του ΕΝΦΙΑ σε σωστή βάση	Good
68.	12/09/2014	14:06:05	Alpha Bank: Ανάκαμψη της αγοράς ακινήτων από το 2015	Bad
69.	15/09/2014	12:25:23	Alpha Bank: Συνεργασία με την China UnionPay	Good
70.	19/09/2014	13:46:31	ALPHA ΤΡΑΠΕΖΑ Α.Ε. : Ανακοίνωση	No Mover
71.	19/09/2014	14:26:44	Alpha Bank: Τι προβλέπει το σχέδιο αναδιάρθρωσης	Good
72.	26/09/2014	13:50:00	Alpha Bank: Η Ελλάδα μπορεί χωρίς το ΔΝΤ από το 2015	Good
73.	03/10/2014	16:42:40	Alpha Bank: Σε γερές βάσεις η ανάκαμψη της οικονομίας	Good
74.	06/10/2014	11:31:08	Ολοκληρώθηκε η εθελουσία στην Alpha Bank	Good
75.	09/10/2014	13:21:00	Αυξάνει τις τιμές - στόχους για Εθνική και Alpha Bank η BofA	Bad
76.	14/10/2014	17:29:35	ALPHA ΤΡΑΠΕΖΑ Α.Ε. : ΓΝΩΣΤΟΠΟΙΗΣΗ ΣΗΜΑΝΤΙΚΩΝ ΑΛΛΑΓΩΝ ΣΤΑ ΔΙΚΑΙΩΜΑΤΑ ΨΗΦΟΥ ΣΥΜΦΩΝΑ ΜΕ ΤΟΝ Ν. 3864/2010 [14.10.2014]	No Mover
77.	17/10/2014	12:26:33	ALPHA ΤΡΑΠΕΖΑ Α.Ε. : ΠΡΟΣΚΛΗΣΗ ΣΕ ΕΚΤΑΚΤΗ ΓΕΝΙΚΗ ΣΥΝΕΛΕΥΣΗ ΤΩΝ ΜΕΤΟΧΩΝ	Bad
78.	29/10/2014	11:04:00	Ορόσημο για τις τράπεζες η απεμπλοκή από τα κρατικά κεφάλαια	Bad
79.	04/11/2014	17:21:40	ALPHA ΤΡΑΠΕΖΑ Α.Ε. :	No Mover

`

			Αποτελέσματα Εννεαμήνου 2014: Κέρδη μετά από Φόρους Ευρώ 110,5 εκατ.	
80.	05/11/2014	11:05:16	Alpha Bank: Καμία συζήτηση για αύξηση κεφαλαίου	Bad
81.	07/11/2014	15:50:00	Reuters: Προς τιτλοποίηση ναυτιλιακών δανείων η Alpha Bank	Bad
82.	11/11/2014	20:11:00	Μιχ. Σάλλας: Έτοιμες οι τράπεζες για την εφαρμογή της ρύθμισης για τα «κόκκινα» δάνεια	Bad
83.	19/11/2014	13:26:42	ΥΠΟΙΚ: Εκδόθηκαν 1,2 εκατ. e-παράβολα μέσα σε 10 μήνες	Good
84.	19/11/2014	17:03:59	Alpha Bank: «Ασύμμετρη» η στάση της τρόικας	No Mover
85.	28/11/2014	11:54:23	ALPHA ΤΡΑΠΕΖΑ Α.Ε. : ΔΙΑΔΙΚΑΣΙΑ ΑΣΚΗΣΕΩΣ ΤΩΝ ΠΑΡΑΣΤΑΤΙΚΩΝ ΤΙΤΛΩΝ ΔΙΚΑΙΩΜΑΤΩΝ ΚΤΗΣΕΩΣ ΜΕΤΟΧΩΝ (WARRANTS) ΚΑΙ ΔΙΑΚΑΝΟΝΙΣΜΟΥ ΤΩΝ ΕΝΤΟΛΩΝ ΣΥΜΜΕΤ	Good
86.	03/12/2014	10:23:36	Alpha Bank: Νέα προγράμματα συμβολαιακής επιχειρηματικότητας	Bad
87.	04/12/2014	15:35:45	Alpha Bank: Αμφιβόλου σκοπιμότητας οι αιτιάσεις της τρόικας	Bad
88.	05/12/2014	10:52:03	Alpha Bank: Εξασφάλιση 500 εκατ. δολ. από τιτλοποίηση δανείων	Good
89.	12/12/2014	15:05:44	Alpha Bank: Σπασμωδική η αντίδραση των αγορών	Good
90.	15/12/2014	17:39:17	ALPHA ΤΡΑΠΕΖΑ Α.Ε. : Οριστικά αποτελέσματα ασκήσεως Παραστατικών Τίτλων Δικαιωμάτων Κτήσεως Μετοχών (Warrants) [15.12.2014]	No Mover
91.	31/12/2014	12:57:30	ALPHA ΤΡΑΠΕΖΑ Α.Ε. : Γνωστοποίηση σημαντικών αλλαγών στα δικαιώματα ψήφου σύμφωνα με τον Ν. 3864/2010 [31.12.2014]	Good

❖ Η μετοχή της Attica Bank:

	Ημερομηνία	***Ώρα***	***Δελτίο Τύπου/ Άρθρο***	***Κατηγορία***
1.	06/02/2014	13:46:29	ATTICA BANK ΑΝΩΝΥΜΗ ΤΡΑΠΕΖΙΚΗ ΕΤΑΙΡΕΙΑ : ΑΝΑΚΟΙΝΩΣΗ ΕΙΣΑΓΩΓΗΣ ΜΕΤΟΧΩΝ	Good
2.	06/02/2014	14:39:48	Attica Bank: Από 10/2 στο Χ.Α. οι νέες μετοχές	No Mover
3.	07/02/2014	17:21:02	ATTICA BANK ΑΝΩΝΥΜΗ ΤΡΑΠΕΖΙΚΗ ΕΤΑΙΡΕΙΑ : ΑΝΑΚΟΙΝΩΣΗ 07/02/2014	No Mover
4.	18/03/2014	12:18:00	Attica Bank: Προς υποβολή νέου επιχειρηματικού σχεδίου	Good
5.	24/03/2014	17:40:20	ATTICA BANK ΑΝΩΝΥΜΗ ΤΡΑΠΕΖΙΚΗ ΕΤΑΙΡΕΙΑ : ΑΝΑΚΟΙΝΩΣΗ ΡΥΘΜΙΖΟΜΕΝΗΣ ΠΛΗΡΟΦΟΡΙΑΣ ΤΟΥ Ν. 3556/2007:	Good

`

			Γνωστοποίηση συναλλαγών	
6.	19/05/2014	13:45:21	ATTICA BANK ΑΝΩΝΥΜΗ ΤΡΑΠΕΖΙΚΗ ΕΤΑΙΡΕΙΑ : ΤΡΟΠΟΠΟΙΗΣΗ ΟΙΚΟΝΟΜΙΚΟΥ ΗΜΕΡΟΛΟΓΙΟΥ 2014	Bad
7.	19/05/2014	13:56:00	Attica Bank: Στις 27/5 η ενημέρωση των αναλυτών	No Mover
8.	27/05/2014	16:02:01	ATTICA BANK ΑΝΩΝΥΜΗ ΤΡΑΠΕΖΙΚΗ ΕΤΑΙΡΕΙΑ : ΑΝΑΚΟΙΝΩΣΗ ΓΙΑ ΕΝΗΜΕΡΩΣΗ ΕΙΣΗΓΜΕΝΗΣ ΠΡΟΣ ΑΝΑΛΥΤΕΣ	Good
9.	18/06/2014	14:45:14	Attica Bank: Αισιοδοξία για συνέχιση της αυτόνομης πορείας	Good
10.	29/07/2014	17:16:39	ATTICA BANK ΑΝΩΝΥΜΗ ΤΡΑΠΕΖΙΚΗ ΕΤΑΙΡΕΙΑ : Ολοκλήρωση της διαδικασίας Εκποίησης Κλασματικών Υπολοίπων που προέκυψαν από Reverse Split των μετο	No Mover
11.	04/08/2014	17:51:10	ATTICA BANK ΑΝΩΝΥΜΗ ΤΡΑΠΕΖΙΚΗ ΕΤΑΙΡΕΙΑ : Ανακοίνωση για την εισαγωγή μετοχών από αύξηση μετοχικού κεφαλαίου μετά από μετατροπή ομολογιών σε	No Mover
12.	09/09/2014	13:00:19	Στην τελική ευθεία η αύξηση κεφαλαίου της Attica Bank	Good
13.	30/09/2014	16:20:29	Attica Bank: Αναβλήθηκε για τις 13/10 η Γ.Σ.	No Mover
14.	13/10/2014	12:53:26	Attica Bank: Μετατίθεται για τις 10 Νοεμβρίου η Γ.Σ.	No Mover
15.	14/10/2014	10:44:22	Attica Bank: Διευκρινίσεις για την αύξηση μετοχικού κεφαλαίου	Bad
16.	10/11/2014	13:17:43	Attica Bank: Αναβλήθηκε για τις 10 Δεκεμβρίου η Γ.Σ.	No Mover
17.	12/11/2014	10:39:29	Attica Bank: Συνεχίζονται οι επαφές με ξένους επενδυτές	Bad
18.	10/12/2014	13:36:28	Εγκρίθηκε η ΑΜΚ στην Attica Bank	Good

❖ Η μετοχή της Εθνικής Τράπεζας Ελλάδος:

	Ημερομηνία	***Ώρα***	***Δελτίο Τύπου/ Άρθρο***	***Κατηγορία***
1.	02/01/2014	14:37:09	ΕΘΝΙΚΗ ΤΡΑΠΕΖΑ ΤΗΣ ΕΛΛΑΔΟΣ Α.Ε. : Ολοκλήρωση μεταβίβασης ποσοστού της Εθνικής ΠΑΝΓΑΙΑ στην INVEL	Good
2.	02/01/2014	15:17:29	ΕΤΕ: Ολοκληρώθηκε η πώληση της Πανγαία	Good
3.	09/01/2014	13:58:37	JPMorgan: Ξεκινά εκ νέου την κάλυψη τριών ελληνικών τραπεζών	Bad
4.	29/01/2014	11:08:58	ΕΘΝΙΚΗ ΤΡΑΠΕΖΑ ΤΗΣ ΕΛΛΑΔΟΣ Α.Ε. : ΔΕΛΤΙΟ ΤΥΠΟΥ	No Mover
5.	29/01/2014	12:16:03	Νέα διευθυντικά στελέχη στην ΕΤΕ	Bad
6.	30/01/2014	11:11:54	Reuters: Πώληση της NBGI εξετάζει η ΕΤΕ	Bad
7.	10/02/2014	11:00:00	Με το «δεξί» ξεκινά την εβδομάδα το Χ.Α.	Good

`

8.	17/02/2014	15:56:42	**ΕΘΝΙΚΗ ΤΡΑΠΕΖΑ ΤΗΣ ΕΛΛΑΔΟΣ Α.Ε. : Ανακοίνωση ολοκλήρωσης της διαδικασίας εκποίησης κλασματικών υπολοίπων, που προέκυψαν από το reverse split**	Bad
9.	21/02/2014	11:04:00	**Reuters: Νέα κεφάλαια 5 δισ. χρειάζονται οι συστημικές τράπεζες**	Bad
10.	10/03/2014	15:57:51	**ΕΘΝΙΚΗ ΤΡΑΠΕΖΑ ΤΗΣ ΕΛΛΑΔΟΣ Α.Ε. : Δελτίο Τύπου-Ημερομηνία και ώρα ανακοίνωσης ετήσιων οικονομικών αποτελεσμάτων 2013**	No Mover
11.	11/03/2014	11:51:23	**Ξεκινά κάλυψη για τις ελληνικές τράπεζες η Wood**	Bad
12.	26/03/2014	13:04:37	**Πήγασος: Σύναψη ενυπόθηκου ομολογιακού 80 εκατ. ευρώ**	Bad
13.	14/04/2014	12:08:00	**Βγαίνει στις αγορές και η Εθνική Τράπεζα**	Bad
14.	15/04/2014	13:56:00	**Προβάδισμα της Fairfax για Eurobank**	Bad
15.	16/04/2014	11:01:29	**Σήμερα «κλειδώνει» η αύξηση κεφαλαίου της ΕΤΕ**	Bad
16.	22/04/2014	15:04:49	**ΕΘΝΙΚΗ ΤΡΑΠΕΖΑ ΤΗΣ ΕΛΛΑΔΟΣ Α.Ε. : ΕΚΤΑΚΤΗ ΓΕΝΙΚΗ ΣΥΝΕΛΕΥΣΗ ΤΗΣ 10ης Μαΐου 2014-Σχέδια Αποφάσεων/Σχόλια Διοικητικού Συμβουλίου επί θεμάτων ημ**	No Mover
17.	22/04/2014	16:19:00	**Εθνική Τράπεζα: Οι όροι της αύξησης του μετοχικού κεφαλαίου**	Bad
18.	24/04/2014	10:48:42	**ΕΘΝΙΚΗ ΤΡΑΠΕΖΑ ΤΗΣ ΕΛΛΑΔΟΣ Α.Ε. : ΑΝΑΚΟΙΝΩΣΗ**	Good
19.	24/04/2014	11:05:00	**ΕΤΕ: Έκδοση πενταετούς ομολόγου 750 εκατ. ευρώ**	Good
20.	24/04/2014	14:38:52	**ΕΘΝΙΚΗ ΤΡΑΠΕΖΑ ΤΗΣ ΕΛΛΑΔΟΣ Α.Ε. : ΟΡΘΗ ΕΠΑΝΑΛΗΨΗ-ΕΚΤΑΚΤΗ ΓΕΝΙΚΗ ΣΥΝΕΛΕΥΣΗ ΤΗΣ 10ης Μαΐου 2014-Σχέδια Αποφάσεων/Σχόλια Διοικητικού Συμβουλίου**	Bad
21.	06/05/2014	10:56:00	**ΕΤΕ: Άνοιξε το βιβλίο προσφορών για την ΑΜΚ**	Good
22.	06/05/2014	17:24:19	**Γαλλικός Τύπος: Ισχυρό ενδιαφέρον ξένων επενδυτών για ελληνικές τραπεζικές μετοχές**	No Mover
23.	08/05/2014	17:00:49	**ΕΘΝΙΚΗ ΤΡΑΠΕΖΑ ΤΗΣ ΕΛΛΑΔΟΣ Α.Ε. : ΑΝΑΚΟΙΝΩΣΗ**	Bad
24.	13/05/2014	10:58:47	**ΕΘΝΙΚΗ ΤΡΑΠΕΖΑ ΤΗΣ ΕΛΛΑΔΟΣ Α.Ε. : Δελτίο Τύπου**	No Mover
25.	13/05/2014	11:05:00	**ΕΤΕ: Στις 28/5 τα αποτελέσματα α' τριμήνου**	No Mover
26.	14/05/2014	14:50:21	**ΕΘΝΙΚΗ ΤΡΑΠΕΖΑ ΤΗΣ ΕΛΛΑΔΟΣ Α.Ε. : Αποτελέσματα Ψηφοφορίας Στα θέματα της Έκτακτης Γενικής Συνέλευσης των Μετόχων της Εθνικής Τράπεζας της Ελ**	Bad
27.	16/05/2014	11:58:00	**ΕΤΕ: Αναβαθμίζει σε «overweight» η JPMorgan**	Good
28.	20/05/2014	14:28:36	**Αλ. Τουρκολιάς: Στηρίζουμε**	No Mover

`

			δυναμικά την οικονομία	
29.	21/05/2014	16:21:36	ΕΤΕ: Δεύτερο έργο στην Αττική μέσω JESSICA	Good
30.	21/05/2014	16:47:46	ΕΘΝΙΚΗ ΤΡΑΠΕΖΑ ΤΗΣ ΕΛΛΑΔΟΣ Α.Ε. : Γνωστοποίηση σημαντικών αλλαγών στα δικαιώματα ψήφου σύμφωνα με το Ν.3556/2007 και το Ν.3864/2010	Good
31.	21/05/2014	16:55:25	ΕΘΝΙΚΗ ΤΡΑΠΕΖΑ ΤΗΣ ΕΛΛΑΔΟΣ Α.Ε. : Γνωστοποίηση σημαντικών αλλαγών στα δικαιώματα ψήφου σύμφωνα με το Ν.3556/2007	Good
32.	04/06/2014	15:30:28	ΕΘΝΙΚΗ ΤΡΑΠΕΖΑ ΤΗΣ ΕΛΛΑΔΟΣ Α.Ε. : ΠΡΟΣΚΛΗΣΗ ΤΑΚΤΙΚΗΣ ΓΕΝΙΚΗΣ ΣΥΝΕΛΕΥΣΗΣ ΤΩΝ ΜΕΤΟΧΩΝ ΤΗΣ 26ΗΣ ΙΟΥΝΙΟΥ 2014, ΗΜΕΡΑ ΠΕΜΠΤΗ ΚΑΙ ΩΡΑ 12:00	No Mover
33.	04/06/2014	16:00:36	ΕΤΕ: Στις 26 Ιουνίου η τακτική Γ.Σ.	No Mover
34.	05/06/2014	13:02:19	Αλ. Τουρκολιάς: Στηρίζει την οικονομία με 3 δισ. ευρώ η ΕΤΕ	Good
35.	06/06/2014	13:51:27	ΕΘΝΙΚΗ ΤΡΑΠΕΖΑ ΤΗΣ ΕΛΛΑΔΟΣ Α.Ε. : ΣΧΕΔΙΑ ΑΠΟΦΑΣΕΩΝ/ΣΧΟΛΙΑ ΔΙΟΙΚΗΤΙΚΟΥ ΣΥΜΒΟΥΛΙΟΥ ΕΠΙ ΘΕΜΑΤΩΝ ΗΜΕΡΗΣΙΑΣ ΔΙΑΤΑΞΗΣ ΤΗΣ ΤΑΚΤΙΚΗΣ ΓΕΝΙΚΗΣ ΣΥΝΕΛΕΥ	Good
36.	16/06/2014	17:09:48	ΕΘΝΙΚΗ ΤΡΑΠΕΖΑ ΤΗΣ ΕΛΛΑΔΟΣ Α.Ε. : ΑΝΑΚΟΙΝΩΣΗ ΡΥΘΜΙΖΟΜΕΝΗΣ ΠΛΗΡΟΦΟΡΙΑΣ ΤΟΥ Ν. 3556/2007: Γνωστοποίηση συναλλαγών	Good
37.	20/06/2014	16:51:50	ΕΘΝΙΚΗ ΤΡΑΠΕΖΑ ΤΗΣ ΕΛΛΑΔΟΣ Α.Ε. : ΔΕΛΤΙΟ ΤΥΠΟΥ	No Mover
38.	24/06/2014	12:17:57	Πώληση της NBGI μέχρι τον Σεπτέμβριο «βλέπει» η ΕΤΕ	Good
39.	24/06/2014	17:29:21	ΕΘΝΙΚΗ ΤΡΑΠΕΖΑ ΤΗΣ ΕΛΛΑΔΟΣ Α.Ε. : ΣΧΕΔΙΑ ΑΠΟΦΑΣΕΩΝ/ΣΧΟΛΙΑ ΔΙΟΙΚΗΤΙΚΟΥ ΣΥΜΒΟΥΛΙΟΥ ΕΠΙ ΘΕΜΑΤΩΝ ΗΜΕΡΗΣΙΑΣ ΔΙΑΤΑΞΗΣ ΤΗΣ ΤΑΚΤΙΚΗΣ ΓΕΝΙΚΗΣ ΣΥΝΕΛΕΥ	No Mover
40.	26/06/2014	16:12:40	Α. Τουρκολιάς: Σύνδεση φορολογικής πολιτικής με ανάγκη αύξησης επενδύσεων	No Mover
41.	26/06/2014	16:35:06	ΕΘΝΙΚΗ ΤΡΑΠΕΖΑ ΤΗΣ ΕΛΛΑΔΟΣ Α.Ε. : ΕΝΗΜΕΡΩΣΗ ΓΙΑ ΤΙΣ ΑΠΟΦΑΣΕΙΣ ΤΗΣ ΤΑΚΤΙΚΗΣ ΓΕΝΙΚΗΣ ΣΥΝΕΛΕΥΣΗΣ ΤΩΝ ΜΕΤΟΧΩΝ ΤΗΣ ΕΘΝΙΚΗΣ ΤΡΑΠΕΖΑΣ ΤΗΣ 26/6/2014	No Mover
42.	26/06/2014	16:43:14	Γ. Ζανιάς: Η οικονομία κερδίζει το στοίχημα της σταθεροποίησης	Good
43.	27/06/2014	17:21:20	ΕΘΝΙΚΗ ΤΡΑΠΕΖΑ ΤΗΣ ΕΛΛΑΔΟΣ Α.Ε. : Αποτελέσματα	No Mover

`

			άσκησης Παραστατικών Τίτλων Δικαιωμάτων Κτήσης Μετοχών (Warrants)–2Η Ασκηση (26/6/2014)	
44.	30/06/2014	15:59:15	ΕΘΝΙΚΗ ΤΡΑΠΕΖΑ ΤΗΣ ΕΛΛΑΔΟΣ Α.Ε. : Γνωστοποίηση σημαντικών αλλαγών στα δικαιώματα ψήφου σύμφωνα με το Ν.3556/2007	No Mover
45.	01/07/2014	16:13:26	ΕΘΝΙΚΗ ΤΡΑΠΕΖΑ ΤΗΣ ΕΛΛΑΔΟΣ Α.Ε. : Αποτελέσματα Ψηφοφορίας στα θέματα της Τακτικής Γενικής Συνέλευσης των Μετόχων της Εθνικής Τράπεζας της Ελ	Good
46.	04/07/2014	12:50:00	ΕΤΕ: Νέα σύμβαση για έργο στην Αττική	Good
47.	07/07/2014	10:40:00	Space Hellas: Επέκταση του συστήματος Cisco UCS στην ΕΤΕ	Good
48.	11/07/2014	16:38:11	Επιτροπή Κεφαλαιαγοράς: Ανανεώθηκε η θητεία του Κ. Μποτόπουλου	Bad
49.	18/07/2014	13:30:56	ΕΘΝΙΚΗ ΤΡΑΠΕΖΑ ΤΗΣ ΕΛΛΑΔΟΣ Α.Ε. : Ημερομηνία και ώρα ανακοίνωσης αποτελεσμάτων Α' εξαμήνου 2014	No Mover
50.	18/07/2014	15:26:31	ΕΘΝΙΚΗ ΤΡΑΠΕΖΑ ΤΗΣ ΕΛΛΑΔΟΣ Α.Ε. : ΔΕΛΤΙΟ ΤΥΠΟΥ	Good
51.	23/07/2014	13:42:40	ΕΘΝΙΚΗ ΤΡΑΠΕΖΑ ΤΗΣ ΕΛΛΑΔΟΣ Α.Ε. : Δελτίο Τύπου	No Mover
52.	24/07/2014	10:55:47	ΕΘΝΙΚΗ ΤΡΑΠΕΖΑ ΤΗΣ ΕΛΛΑΔΟΣ Α.Ε. : ΑΝΑΚΟΙΝΩΣΗ	No Mover
53.	28/07/2014	11:53:30	ΕΤΕ: Βελτιώνεται το επιχειρηματικό κλίμα	Bad
54.	20/08/2014	11:10:54	ΕΘΝΙΚΗ ΤΡΑΠΕΖΑ ΤΗΣ ΕΛΛΑΔΟΣ Α.Ε. : ΑΝΑΚΟΙΝΩΣΗ	No Mover
55.	20/08/2014	11:49:34	ΕΘΝΙΚΗ ΤΡΑΠΕΖΑ ΤΗΣ ΕΛΛΑΔΟΣ Α.Ε. : ΑΝΑΚΟΙΝΩΣΗ (ΟΡΘΗ ΕΠΑΝΑΛΗΨΗ)	No Mover
56.	26/08/2014	13:36:44	Θετικές συστάσεις Deutsche Bank για τις ελληνικές τράπεζες	Good
57.	27/08/2014	11:49:41	Συστάσεις από Citi για τις τραπεζικές μετοχές	Bad
58.	29/08/2014	16:00:05	ΕΘΝΙΚΗ ΤΡΑΠΕΖΑ ΤΗΣ ΕΛΛΑΔΟΣ Α.Ε. : ΔΕΛΤΙΟ ΤΥΠΟΥ	No Mover
59.	29/08/2014	16:18:09	ΕΤΕ: Ο Απ. Καζάκος βοηθός γενικός διευθυντής στρατηγικής	No Mover
60.	17/09/2014	15:33:53	Μείωση επιτοκίων από την Εθνική Τράπεζα	Bad
61.	17/09/2014	17:06:00	Την Κυριακή ο γραπτός διαγωνισμός για προσλήψεις στην Εθνική Τράπεζα	No Mover
62.	19/09/2014	11:08:03	ΕΘΝΙΚΗ ΤΡΑΠΕΖΑ ΤΗΣ ΕΛΛΑΔΟΣ Α.Ε. : ΔΕΛΤΙΟ ΤΥΠΟΥ	Good
63.	19/09/2014	13:21:00	ΕΤΕ: Υπεγράφη η σύμβαση για την πώληση του Αστέρα	Good
64.	29/09/2014	13:51:03	Συνολικά 2.652 υποψήφιοι συμμετείχαν στον διαγωνισμό για τις προσλήψεις στην ΕΤΕ	No Mover
65.	30/09/2014	15:27:00	Εθνική: Παρατείνεται το	No Mover

`

			ωράριο έως τις 18:00 σήμερα σε 70 καταστήματα	
66.	09/10/2014	13:21:00	Αυξάνει τις τιμές - στόχους για Εθνική και Alpha Bank η BofA	Good
67.	17/10/2014	11:04:27	ΕΘΝΙΚΗ ΤΡΑΠΕΖΑ ΤΗΣ ΕΛΛΑΔΟΣ Α.Ε. : ΠΡΟΣΚΛΗΣΗ ΕΚΤΑΚΤΗΣ ΓΕΝΙΚΗΣ ΣΥΝΕΛΕΥΣΗΣ ΤΩΝ ΜΕΤΟΧΩΝ της 7ης Νοεμβρίου 2014, ημέρα Παρασκευή και ώρα 13:0	No Mover
68.	17/10/2014	11:45:00	Συνελεύσεις στις τράπεζες για τον αναβαλλόμενο φόρο	Good
69.	17/10/2014	14:12:46	ΕΘΝΙΚΗ ΤΡΑΠΕΖΑ ΤΗΣ ΕΛΛΑΔΟΣ Α.Ε. : ΠΡΟΣΚΛΗΣΗ ΕΚΤΑΚΤΗΣ ΓΕΝΙΚΗΣ ΣΥΝΕΛΕΥΣΗΣ ΤΩΝ ΜΕΤΟΧΩΝ της 7ης Νοεμβρίου 2014, ημέρα Παρασκευή και ώρα 13:0	No Mover
70.	21/10/2014	12:05:21	Αισιόδοξος ο Αλ. Τουρκολιάς για τα stress tests	Good
71.	22/10/2014	10:36:30	ΕΘΝΙΚΗ ΤΡΑΠΕΖΑ ΤΗΣ ΕΛΛΑΔΟΣ Α.Ε. : ΑΝΑΚΟΙΝΩΣΗ	Good
72.	22/10/2014	11:19:05	ΕΤΕ: Προχωρά η ΑΜΚ στην Finansbank	Good
73.	24/10/2014	13:57:51	ΕΘΝΙΚΗ ΤΡΑΠΕΖΑ ΤΗΣ ΕΛΛΑΔΟΣ Α.Ε. : ΕΚΤΑΚΤΗ ΓΕΝΙΚΗ ΣΥΝΕΛΕΥΣΗ ΤΗΣ 7ης Νοεμβρίου 2014-Σχέδια Αποφάσεων/Σχόλια Διοικητικού Συμβουλίου επί θεμάτων	Good
74.	27/10/2014	17:42:52	ΕΘΝΙΚΗ ΤΡΑΠΕΖΑ ΤΗΣ ΕΛΛΑΔΟΣ Α.Ε. : ΑΝΑΚΟΙΝΩΣΗ	No Mover
75.	29/10/2014	11:04:00	Ορόσημο για τις τράπεζες η απεμπλοκή από τα κρατικά κεφάλαια	Good
76.	29/10/2014	14:16:33	Αλ. Τουρκολιάς: Δεν απαιτείται καμία περαιτέρω κεφαλαιακή ενέργεια	Bad
77.	07/11/2014	15:34:19	Αλ. Τουρκολιάς: Πλεόνασμα κεφαλαίων στην ΕΤΕ	Bad
78.	12/11/2014	15:56:18	Αστήρ Παλάς: «Πράσινο» από Γ.Σ. στην αύξηση κεφαλαίου	Bad
79.	12/11/2014	16:44:00	ΕΤΕ: Διευκρινίσεις για Finansbank	Good
80.	19/11/2014	13:26:42	ΥΠΟΙΚ: Εκδόθηκαν 1,2 εκατ. e-παράβολα μέσα σε 10 μήνες	Bad
81.	27/11/2014	17:10:44	Μνημόνιο συνεργασίας Εθνικής Τράπεζας με το Βιοτεχνικό Επιμελητήριο Αθηνών	Good
82.	16/12/2014	12:34:13	ΕΘΝΙΚΗ ΤΡΑΠΕΖΑ ΤΗΣ ΕΛΛΑΔΟΣ Α.Ε. : ΔΙΑΔΙΚΑΣΙΑ ΑΣΚΗΣΗΣ ΤΩΝ ΠΑΡΑΣΤΑΤΙΚΩΝ ΤΙΤΛΩΝ ΔΙΚΑΙΩΜΑΤΩΝ ΚΤΗΣΗΣ ΜΕΤΟΧΩΝ (WARRANTS) ΚΑΙ ΔΙΑΚΑΝΟΝΙΣΜΟΥ ΤΩΝ ΕΝΤ	Good
83.	19/12/2014	16:47:00	ΕΤΕ: Το i-bank «μπαίνει» στη μικρή λιανική	Bad
84.	23/12/2014	14:07:54	«Πράσινο» από ΥΠΑΝ για τη συγχώνευση ΕΤΕ - Εθνικής Κεφαλαίου	No Mover
85.	30/12/2014	15:01:47	ΕΘΝΙΚΗ ΤΡΑΠΕΖΑ ΤΗΣ	Good

`

			ΕΛΛΑΔΟΣ Α.Ε. : Αποτελέσματα άσκησης Παραστατικών Τίτλων Δικαιωμάτων Κτήσης Μετοχών (Warrants) – 3Η Άσκηση (29/12/20	

❖ Η μετοχή της Τράπεζας της Ελλάδος:

	Ημερομηνία	Ώρα	Δελτίο Τύπου/ Άρθρο	Κατηγορία
1.	03/01/2014	12:19:25	ΤτΕ: Στο -3,8% η πιστωτική επέκταση τον Νοέμβριο	Bad
2.	08/01/2014	13:18:00	ΤτΕ: Πτώση στα επιτόκια καταθέσεων και δανείων	Good
3.	13/01/2014	13:56:00	ΠΑΣΟΚ: Καμία συζήτηση για διαδοχή του Γ. Προβόπουλου στην ΤτΕ	No Mover
4.	14/01/2014	17:26:30	ΤΡΑΠΕΖΑ ΤΗΣ ΕΛΛΑΔΟΣ : ΑΝΑΚΟΙΝΩΣΗ ΟΙΚΟΝΟΜΙΚΟΥ ΗΜΕΡΟΛΟΓΙΟΥ	No Mover
5.	16/01/2014	11:01:42	Βουλή: Επίθεση Π. Καμμένου στον Γ. Προβόπουλο	Bad
6.	16/01/2014	11:33:00	Γ. Προβόπουλος: Σημαντικό πρόβλημα από το κοινωνικοπολιτικό κλίμα	Bad
7.	16/01/2014	12:12:00	Γ. Προβόπουλος: Δυο φορές απαγόρευσα στο Τ.Τ. να δίνει επιχειρηματικά δάνεια	Bad
8.	17/01/2014	13:12:00	Φάκελοι με σφαίρες στους Γ. Προβόπουλο και Γ. Πρετεντέρη	No Mover
9.	28/01/2014	13:22:25	ΤτΕ: Αύξηση 12,8% στις ταξιδιωτικές εισπράξεις	Good
10.	27/02/2014	12:16:00	Γ. Προβόπουλος: Πρώτη χρονιά ανάκαμψης το 2014	Good
11.	27/02/2014	12:38:42	Την επόμενη εβδομάδα τα αποτελέσματα των stress test	No Mover
12.	27/02/2014	14:36:00	ΤτΕ: Στο -4% η πιστωτική επέκταση τον Ιανουάριο	Bad
13.	27/02/2014	16:37:09	ΤτΕ: Μειώθηκαν οι τραπεζικές καταθέσεις τον Ιανουάριο	Bad
14.	04/03/2014	13:52:00	Νέα συνάντηση τρόικας - Γ. Προβόπουλου για τις τράπεζες	Bad
15.	05/03/2014	13:54:00	ΤτΕ: Μειώθηκαν τα επιτόκια καταθέσεων	Bad
16.	07/03/2014	17:01:00	Η ΤτΕ αναλαμβάνει τη σίτιση 1.300 άπορων μαθητών	No Mover
17.	10/04/2014	15:46:00	Σύλλογος Εργαζομένων ΤτΕ: Άναδρη ενέργεια η τρομοκρατική	No Mover

`

			επίθεση	
18.	11/04/2014	14:46:00	**Γ. Προβόπουλος: Διαφαίνεται η προοπτική εξόδου από την κρίση**	Good
19.	14/04/2014	12:08:00	**Βγαίνει στις αγορές και η Εθνική Τράπεζα**	Bad
20.	14/04/2014	17:25:32	**ΤΡΑΠΕΖΑ ΤΗΣ ΕΛΛΑΔΟΣ : ΜΗΝΙΑΙΑ ΣΥΝΟΠΤΙΚΗ ΛΟΓΙΣΤΙΚΗ ΚΑΤΑΣΤΑΣΗ ΜΑΡΤΙΟΥ 2014**	Bad
21.	17/04/2014	11:14:15	**ΤτΕ: Στα 709 εκατ. το έλλειμμα τρεχουσών συναλλαγών**	Bad
22.	17/04/2014	12:19:00	**ΤτΕ: Στα 932 εκατ. το ταμειακό πρωτογενές πλεόνασμα**	Good
23.	17/04/2014	13:19:14	**ΤτΕ: Αύξηση 11,7% στις ταξιδιωτικές εισπράξεις**	Good
24.	05/05/2014	13:33:18	**Space Hellas: Ανάληψη έργου για την Τράπεζα της Ελλάδος**	Bad
25.	06/05/2014	13:55:34	**ΤτΕ: Μικρές διακυμάνσεις στα επιτόκια καταθέσεων και δανείων**	Bad
26.	14/05/2014	17:31:54	**ΤΡΑΠΕΖΑ ΤΗΣ ΕΛΛΑΔΟΣ : ΜΗΝΙΑΙΑ ΣΥΝΟΠΤΙΚΗ ΛΟΓΙΣΤΙΚΗ ΚΑΤΑΣΤΑΣΗ ΑΠΡΙΛΙΟΥ 2014**	No Mover
27.	20/05/2014	13:21:19	**ΤτΕ: Πτώση 7,5% στις τιμές των διαμερισμάτων**	Good
28.	23/05/2014	15:07:00	**Πλεόνασμα 73 εκατ. ευρώ στο ταξιδιωτικό ισοζύγιο**	Good
29.	30/05/2014	13:03:06	**Ολοκληρώθηκε η συνάντηση Γουάτσα - Προβόπουλου**	Good
30.	12/06/2014	12:53:00	**ΤτΕ: Κανένα περιθώριο εφησυχασμού**	Good
31.	12/06/2014	14:41:00	**Financial TiMes: Τα επισφαλή δάνεια πρόκληση για τον Γ.Στουρνάρα**	Good
32.	12/06/2014	14:53:31	**ΤτΕ: Ενδείξεις σταθεροποίησης στην αγορά ακινήτων**	Good
33.	12/06/2014	17:20:41	**ΤΡΑΠΕΖΑ ΤΗΣ ΕΛΛΑΔΟΣ : ΜΗΝΙΑΙΑ ΣΥΝΟΠΤΙΚΗ ΛΟΓΙΣΤΙΚΗ ΚΑΤΑΣΤΑΣΗ ΜΑΙΟΥ 2014**	No Mover
34.	13/06/2014	12:42:24	**Spiegel: Φόβος μπροστά στον πανικό των καταθετών**	Bad
35.	13/06/2014	14:19:44	**ΤτΕ: Ταμειακό πρωτογενές πλεόνασμα 1,057 δισ. στο 5μηνο**	Bad
36.	16/06/2014	15:19:07	**ΕΛΙΑΜΕΠ: Οι πολιτικές που χρηματοδοτούνται από την Ε.Ε. και η ελληνική οικονομία**	No Mover
37.	17/06/2014	15:26:45	**Επιστολή Γ. Στουρνάρα στους Ευρωπαίους υπουργούς Οικονομικών**	Bad
38.	17/06/2014	15:55:00	**ΤτΕ: Νέο πλαίσιο για τα «κόκκινα» δάνεια**	Good

`

39.	20/06/2014	11:24:41	**ΤτΕ: Μειώθηκε το έλλειμμα τρεχουσών συναλλαγών στο α' τετράμηνο**	Good
40.	20/06/2014	12:49:18	**Ευ. Βενιζέλος: Τεράστια η συμβολή του Γ. Προβόπουλου στη διάσωση της οικονομίας**	Bad
41.	27/06/2014	12:49:00	**ΤτΕ: Ορκίστηκε νέος διοικητής ο Γ. Στουρνάρας**	Good
42.	27/06/2014	14:00:04	**Γ. Στουρνάρας: Σε καλύτερη πορεία η οικονομία**	Good
43.	27/06/2014	16:11:28	**Το γενικό συμβούλιο της ΤτΕ**	No Mover
44.	15/07/2014	17:23:11	**ΤΡΑΠΕΖΑ ΤΗΣ ΕΛΛΑΔΟΣ : ΜΗΝΙΑΙΑ ΣΥΝΟΠΤΙΚΗ ΛΟΓΙΣΤΙΚΗ ΚΑΤΑΣΤΑΣΗ ΙΟΥΝΙΟΥ 2014**	No Mover
45.	18/07/2014	14:32:00	**ΤτΕ: Καμία απόφαση για τον νέο υποδιοικητή**	No Mover
46.	18/07/2014	14:46:00	**ΤτΕ: Το 20% του συνόλου των επιχειρηματικών δανείων είναι προβληματικά**	Bad
47.	25/07/2014	12:21:09	**ΤτΕ: Στο -3,5% ο ρυθμός χρηματοδότησης του ιδιωτικού τομέα τον Ιούνιο**	Bad
48.	06/08/2014	11:56:59	**Μείωση 7,3% στις τιμές των διαμερισμάτων**	Good
49.	11/08/2014	17:11:32	**ΤΡΑΠΕΖΑ ΤΗΣ ΕΛΛΑΔΟΣ : ΜΗΝΙΑΙΑ ΣΥΝΟΠΤΙΚΗ ΛΟΓΙΣΤΙΚΗ ΚΑΤΑΣΤΑΣΗ ΙΟΥΛΙΟΥ 2014**	No Mover
50.	25/08/2014	17:13:00	**ΤτΕ: Εγκρίθηκε ο Κώδικας Λεοντολογίας για τα μη εξυπηρετούμενα δάνεια**	No Mover
51.	28/08/2014	13:28:07	**ΤτΕ: Στο -3,7% η πιστωτική επέκταση τον Ιούλιο**	Bad
52.	28/08/2014	16:00:05	**Τι συζήτησαν Γ. Στουρνάρας - Μπ. Κερέ**	Good
53.	28/08/2014	16:55:57	**ΤτΕ: Μικρή αύξηση καταθέσεων τον Ιούλιο**	Good
54.	23/09/2014	14:19:23	**ΤτΕ: Αύξηση 15,1% στο ταξιδιωτικό πλεόνασμα τον Ιούλιο**	Bad
55.	03/12/2014	13:27:11	**ΤτΕ: Μειώθηκαν τα επιτόκια νέων καταθέσεων**	Bad
56.	10/12/2014	17:23:49	**ΤΡΑΠΕΖΑ ΤΗΣ ΕΛΛΑΔΟΣ : ΜΗΝΙΑΙΑ ΣΥΝΟΠΤΙΚΗ ΛΟΓΙΣΤΙΚΗ ΚΑΤΑΣΤΑΣΗ ΝΟΕΜΒΡΙΟΥ 2014**	No Mover
57.	15/12/2014	13:24:08	**ΤτΕ: Ταμειακό πρωτογενές πλεόνασμα 2,6 δισ. στο 11μηνο**	Good
58.	22/12/2014	11:16:45	**ΤτΕ: Στα 3,6 δισ. ευρώ το πλεόνασμα τρεχουσών συναλλαγών 10μηνου**	Good
59.	30/12/2014	13:21:25	**ΤτΕ: Στο -3,0% η πιστωτική επέκταση τον Νοέμβριο**	Good

`

❖ Η μετοχή της Τράπεζας Κύπρου:

1.	Ημερομηνία	Ώρα	Δελτίο Τύπου/ Άρθρο	Κατηγορία
2.	16/12/2014	11:57:49	Επέστρεψε σε ΧΑΚ – Χ.Α. η Τράπεζα Κύπρου	Bad
3.	17/12/2014	13:31:44	Κύπρος: Στόχος μια νέα έξοδος στις αγορές το 2015	Good
4.	17/12/2014	16:44:00	Έρευνα για πιθανή χειραγώγηση μετοχών της Τράπεζας Κύπρου	Bad
5.	31/12/2014	13:04:05	ΤΡΑΠΕΖΑ ΚΥΠΡΟΥ ΔΗΜΟΣΙΑ ΕΤΑΙΡΙΑ ΛΙΜΙΤΕΔ : Παραίτηση Διευθυντή Διεύθυνσης Εσωτερικού Ελέγχου	Bad

❖ Η μετοχή της Τράπεζας Πειραιώς:

	Ημερομηνία	Ώρα	Δελτίο Τύπου/ Άρθρο	Κατηγορία
1.	02/01/2014	17:13:41	ΤΡΑΠΕΖΑ ΠΕΙΡΑΙΩΣ Α.Ε. : ΑΝΑΚΟΙΝΩΣΗ ΟΛΟΚΛΗΡΩΣΗΣ ΤΗΣ ΔΙΑΔΙΚΑΣΙΑΣ ΕΚΠΟΙΗΣΗΣ ΚΛΑΣΜΑΤΙΚΩΝ ΥΠΟΛΟΙΠΩΝ, ΠΟΥ ΠΡΟΕΚΥΨΑΝ ΑΠΟ ΤΟ REVERSE SPLIT ΤΩΝ ΜΕΤΟΧ	No Mover
2.	09/01/2014	13:58:37	JPMorgan: Ξεκινά εκ νέου την κάλυψη τριών ελληνικών τραπεζών	Good
3.	16/01/2014	10:56:20	Geniki: Διερευνητικές επαφές για διεθνείς συνεργασίες	Good
4.	31/01/2014	15:52:49	Χρ. Αντωνιάδης: Έρχονται επιχειρηματικές συγχωνεύσεις	Good
5.	21/02/2014	11:04:00	Reuters: Νέα κεφάλαια 5 δισ. χρειάζονται οι συστημικές τράπεζες	Good
6.	24/02/2014	17:42:41	ΤΡΑΠΕΖΑ ΠΕΙΡΑΙΩΣ Α.Ε. : ΑΝΑΚΟΙΝΩΣΗ ΡΥΘΜΙΖΟΜΕΝΗΣ ΠΛΗΡΟΦΟΡΙΑΣ Ν. 3556/2007	No Mover
7.	26/02/2014	13:08:39	ΤΡΑΠΕΖΑ ΠΕΙΡΑΙΩΣ Α.Ε. : ΑΝΑΚΟΙΝΩΣΗ	No Mover
8.	05/03/2014	10:41:27	Τράπεζα Πειραιώς: Πιστοποίηση Tier 4 για το data center	Good
9.	06/03/2014	17:03:58	Υπεγράφη η πρώτη σύμβαση χρηματοδότησης έργου στην Κεντρική Μακεδονία	Good
10.	06/03/2014	17:19:50	ΤΡΑΠΕΖΑ ΠΕΙΡΑΙΩΣ Α.Ε. : ΔΕΛΤΙΟ ΤΥΠΟΥ	No Mover
11.	11/03/2014	11:51:23	Ξεκινά κάλυψη για τις ελληνικές τράπεζες η Wood	Good
12.	18/03/2014	10:46:27	S&P: Επιβεβαιώνει το «CCC/C» για Πειραιώς, Alpha	Bad
13.	18/03/2014	11:52:00	Τράπεζα Πειραιώς: Άνοιξε βιβλίο προσφορών για τριετές ομόλογο	Good
14.	18/03/2014	14:37:00	Τρ. Πειραιώς: Στο 5,125% αναμένεται η απόδοση του ομολόγου	Bad
15.	20/03/2014	15:17:30	Reuters: Αλλάζουν σελίδα οι ελληνικές τράπεζες	Good
16.	26/03/2014	11:14:00	Με άνοδο και εντυπωσιακό τζίρο ξεκίνησε το Χ.Α.	Good

`

17.	26/03/2014	13:04:37	**Πήγασος: Σύναψη ενυπόθηκου ομολογιακού 80 εκατ. ευρώ**	Bad
18.	26/03/2014	14:15:52	**Reuters: Προσφορές άνω των 3 δισ. στην ΑΜΚ της Πειραιώς**	Good
19.	26/03/2014	15:05:23	**ΤΡΑΠΕΖΑ ΠΕΙΡΑΙΩΣ Α.Ε. : ΔΕΛΤΙΟ ΤΥΠΟΥ**	Good
20.	26/03/2014	15:19:00	**Τράπεζα Πειραιώς: Στα 1,70 ευρώ η τιμή διάθεσης των νέων μετοχών**	Good
21.	27/03/2014	12:54:37	**BlooMberg: Ανακάμπτουν οι ελληνικές τράπεζες**	Good
22.	28/03/2014	13:07:00	**Εγκρίθηκε η ΑΜΚ της Τράπεζας Πειραιώς**	Good
23.	28/03/2014	13:14:33	**ΤΧΣ: «Πράσινο» για τις ΑΜΚ σε Alpha και Πειραιώς**	Good
24.	28/03/2014	13:33:35	**ΤΡΑΠΕΖΑ ΠΕΙΡΑΙΩΣ Α.Ε. : ΔΕΛΤΙΟ ΤΥΠΟΥ**	Bad
25.	31/03/2014	17:19:00	**Γ. Προβόπουλος: Οι αγορές ξανανοίγουν για την Ελλάδα**	No Mover
26.	01/04/2014	12:43:42	**ΤΡΑΠΕΖΑ ΠΕΙΡΑΙΩΣ Α.Ε. : ΑΠΟΤΕΛΕΣΜΑΤΑ ΨΗΦΟΦΟΡΙΑΣ ΤΗΣ ΕΚΤΑΚΤΗΣ ΓΕΝΙΚΗΣ ΣΥΝΕΛΕΥΣΗΣ ΤΩΝ ΜΕΤΟΧΩΝ ΚΑΤΟΧΩΝ ΚΟΙΝΩΝ ΜΕΤΟΧΩΝ ΤΗΣ ΤΡΑΠΕΖΑΣ ΠΕΙΡΑΙΩΣ ΠΟΥ**	No Mover
27.	01/04/2014	16:06:13	**ΑΒ Βασιλόπουλος: Νέες κάρτες ΑΒ Plus MasterCard**	Good
28.	02/04/2014	10:40:01	**Fitch: Αναβάθμιση προοπτικών για ελληνικά καλυμμένα ομόλογα**	Good
29.	02/04/2014	15:56:30	**ΤΡΑΠΕΖΑ ΠΕΙΡΑΙΩΣ Α.Ε. : ΑΝΑΚΟΙΝΩΣΗ ΔΙΑΘΕΣΗΣ ΕΝΗΜΕΡΩΤΙΚΟΥ ΔΕΛΤΙΟΥ**	Good
30.	07/04/2014	16:47:00	**Τρ. Πειραιώς: Ανέτοιμη ή απρόθυμη η ΕΚΤ για μη-συμβατικά μέτρα;**	Bad
31.	16/04/2014	12:22:41	**ΤΡΑΠΕΖΑ ΠΕΙΡΑΙΩΣ Α.Ε. : ΑΝΑΚΟΙΝΩΣΗ ΡΥΘΜΙΖΟΜΕΝΗΣ ΠΛΗΡΟΦΟΡΙΑΣ Ν. 3556/2007**	Bad
32.	17/04/2014	14:57:06	**ΤΡΑΠΕΖΑ ΠΕΙΡΑΙΩΣ Α.Ε. : ΓΝΩΣΤΟΠΟΙΗΣΗ ΣΗΜΑΝΤΙΚΩΝ ΑΛΛΑΓΩΝ ΣΤΑ ΔΙΚΑΙΩΜΑΤΑ ΨΗΦΟΥ ΣΥΜΦΩΝΑ ΜΕ ΤΟ Ν. 3556/2007**	Bad
33.	24/04/2014	16:59:12	**ΤΡΑΠΕΖΑ ΠΕΙΡΑΙΩΣ Α.Ε. : ΓΝΩΣΤΟΠΟΙΗΣΗ ΣΗΜΑΝΤΙΚΩΝ ΑΛΛΑΓΩΝ ΣΤΑ ΔΙΚΑΙΩΜΑΤΑ ΨΗΦΟΥ ΣΥΜΦΩΝΑ ΜΕ ΤΟ Ν. 3556/2007**	Bad
34.	24/04/2014	16:59:43	**ΤΡΑΠΕΖΑ ΠΕΙΡΑΙΩΣ Α.Ε. : ΓΝΩΣΤΟΠΟΙΗΣΗ ΣΗΜΑΝΤΙΚΩΝ ΑΛΛΑΓΩΝ ΣΤΑ ΔΙΚΑΙΩΜΑΤΑ ΨΗΦΟΥ ΣΥΜΦΩΝΑ ΜΕ ΤΟ Ν. 3556/2007**	Bad
35.	06/05/2014	12:06:53	**Περιβαλλοντική διάκριση για την Τράπεζα Πειραιώς**	Good
36.	16/05/2014	14:51:00	**Το νέο δ.σ. της Τράπεζας Πειραιώς**	Bad
37.	16/05/2014	16:56:42	**ΤΡΑΠΕΖΑ ΠΕΙΡΑΙΩΣ Α.Ε. : ΔΕΛΤΙΟ ΤΥΠΟΥ**	Bad
38.	19/05/2014	12:02:10	**Τράπεζα Πειραιώς:**	Bad

`

			Διευκρινίσεις για τη συμφωνία με MIG	
39.	19/05/2014	13:51:18	ΤΡΑΠΕΖΑ ΠΕΙΡΑΙΩΣ Α.Ε. : ΔΕΛΤΙΟ ΤΥΠΟΥ	Bad
40.	22/05/2014	11:08:29	Τρ. Κύπρου: Πώληση δανείων στη Σερβία στην Τρ. Πειραιώς	Good
41.	22/05/2014	11:27:32	Τρ. Πειραιώς: Αποπληρωμή προνομιούχων μετοχών του Δημοσίου	Good
42.	22/05/2014	11:34:16	ΤΡΑΠΕΖΑ ΠΕΙΡΑΙΩΣ Α.Ε. : ΑΠΟΠΛΗΡΩΜΗ ΠΡΟΝΟΜΙΟΥΧΩΝ ΜΕΤΟΧΩΝ ΕΛΛΗΝΙΚΟΥ ΔΗΜΟΣΙΟΥ	Good
43.	28/05/2014	16:32:00	Προς απορρόφηση της Γενικής η Τράπεζα Πειραιώς	Good
44.	02/06/2014	12:24:00	Intrasoft: Ολοκλήρωση έργου για την Τράπεζα Πειραιώς	Good
45.	04/06/2014	10:54:49	Νέα προγράμματα συμβολαιακής γεωργίας από την Τράπεζα Πειραιώς	Bad
46.	13/06/2014	13:41:40	ΤΡΑΠΕΖΑ ΠΕΙΡΑΙΩΣ Α.Ε. : ΑΝΑΚΟΙΝΩΣΗ	No Mover
47.	13/06/2014	15:58:00	Τρ. Πειραιώς: Δεν γνωρίζουμε απόφαση του ΤΧΣ για διάθεση μετοχών	No Mover
48.	03/07/2014	15:32:00	Τράπεζα Πειραιώς: Νέα συμφωνία για Συμβολαιακή Κτηνοτροφία	Good
49.	11/07/2014	15:18:34	ΤΡΑΠΕΖΑ ΠΕΙΡΑΙΩΣ Α.Ε. : ΑΝΑΚΟΙΝΩΣΗ ΗΜΕΡΟΜΗΝΙΑΣ ΑΠΟΤΕΛΕΣΜΑΤΩΝ Α' ΕΞΑΜΗΝΟΥ 2014	Good
50.	11/07/2014	16:20:18	Τράπεζα Πειραιώς: Στις 28/8 τα αποτελέσματα α' εξαμήνου	No Mover
51.	16/07/2014	13:58:06	ΤΡΑΠΕΖΑ ΠΕΙΡΑΙΩΣ Α.Ε. : ΑΝΑΚΟΙΝΩΣΗ ΗΜΕΡΟΜΗΝΙΑΣ ΑΠΟΤΕΛΕΣΜΑΤΩΝ Α' ΕΞΑΜΗΝΟΥ 2014 (ΟΡΘΗ ΕΠΑΝΑΛΗΨΗ)	No Mover
52.	16/07/2014	14:05:00	Τράπεζα Πειραιώς: Στις 29/8 τα αποτελέσματα α' εξαμήνου	No Mover
53.	22/07/2014	14:31:52	ΤΡΑΠΕΖΑ ΠΕΙΡΑΙΩΣ Α.Ε. : ΑΝΑΚΟΙΝΩΣΗ	Good
54.	31/07/2014	15:20:00	Τράπεζα Πειραιώς: Νέα συμφωνία συμβολαιακής κτηνοτροφίας	Good
55.	26/08/2014	13:36:44	Θετικές συστάσεις Deutsche Bank για τις ελληνικές τράπεζες	Good
56.	27/08/2014	11:49:41	Συστάσεις από Citi για τις τραπεζικές μετοχές	Good
57.	02/09/2014	15:00:33	Ανθ. Θωμόπουλος: Ανάγκη αναδιάρθρωσης του ιδιωτικού χρέους	Bad
58.	23/09/2014	13:59:53	Μείωση επιτοκίων από Τράπεζα Πειραιώς	Bad
59.	24/09/2014	12:04:00	Geniki Bank: Αναπροσαρμογή επιτοκίων καταθέσεων	Bad
60.	24/09/2014	13:15:00	Νέες διακρίσεις για την ηλεκτρονική τραπεζική της Τρ. Πειραιώς	Good
61.	26/09/2014	15:30:13	ΤΡΑΠΕΖΑ ΠΕΙΡΑΙΩΣ Α.Ε. : ΓΝΩΣΤΟΠΟΙΗΣΗ ΔΙΚΑΙΩΜΑΤΩΝ ΨΗΦΟΥ	Good

`

			ΔΥΝΑΜΕΙ ΤΟΥ Ν. 3864/2010	
62.	02/10/2014	12:48:00	Τρ. Πειραιώς: Νέα συμφωνία για συμβολαιακή γεωργία	Bad
63.	07/10/2014	15:13:34	Διεθνής διάκριση για την Τράπεζα Πειραιώς	Good
64.	08/10/2014	12:51:15	Τρ. Πειραιώς: Ανανέωση πιστοποίησης για τις πληρωμές ενισχύσεων αγροτών	Bad
65.	09/10/2014	10:11:45	Τράπεζα Πειραιώς: Υποχρεωτική δημόσια πρόταση για Trastor	Good
66.	09/10/2014	13:21:00	Αυξάνει τις τιμές - στόχους για Εθνική και Alpha Bank η BofA	Bad
67.	10/10/2014	11:12:00	Προσωρινή αναστολή διαπραγμάτευσης για Trastor	Bad
68.	10/10/2014	11:45:00	Trastor: Αναστέλλεται η δημόσια πρόταση της Πειραιώς	Bad
69.	10/10/2014	14:04:53	Άρση αναστολής διαπραγμάτευσης για Trastor	Good
70.	13/10/2014	16:30:07	Υπηρεσίες easypay Point και e-Παράβολο από την Τράπεζα Πειραιώς	Good
71.	17/10/2014	11:09:37	Τράπεζα Πειραιώς: Νέος τομέας InvestMent Banking	Good
72.	29/10/2014	11:04:00	Ορόσημο για τις τράπεζες η απεμπλοκή από τα κρατικά κεφάλαια	Good
73.	03/11/2014	17:24:07	ΤΡΑΠΕΖΑ ΠΕΙΡΑΙΩΣ Α.Ε. : ΑΠΟΤΕΛΕΣΜΑΤΑ ΨΗΦΟΦΟΡΙΑΣ ΤΗΣ ΕΚΤΑΚΤΗΣ ΓΕΝΙΚΗΣ ΣΥΝΕΛΕΥΣΗΣ ΤΩΝ ΜΕΤΟΧΩΝ ΤΗΣ ΤΡΑΠΕΖΑΣ ΠΕΙΡΑΙΩΣ ΠΟΥ ΠΡΑΓΜΑΤΟΠΟΙΗΘΗΚΕ ΣΤΙΣ	No Mover
74.	07/11/2014	16:07:41	Τρ. Πειραιώς: Στις 25/11 τα αποτελέσματα 9μηνου	Bad
75.	19/11/2014	13:26:42	ΥΠΟΙΚ: Εκδόθηκαν 1,2 εκατ. e-παράβολα μέσα σε 10 μήνες	Good
76.	20/11/2014	14:01:58	«Πράσινο φως» για τη συγχώνευση Πειραιώς - Geniki	Good
77.	21/11/2014	10:23:01	Το Σαββατοκύριακο η ενοποίηση των πληροφοριακών συστημάτων Πειραιώς - Geniki	Good
78.	25/11/2014	17:22:49	ΤΡΑΠΕΖΑ ΠΕΙΡΑΙΩΣ Α.Ε. : ΑΠΟΤΕΛΕΣΜΑΤΑ ΕΝΝΕΑΜΗΝΟΥ 2014	No Mover
79.	28/11/2014	16:10:14	Φορολογικό σύστημα με ρήτρες 15ετίας προτείνει ο Μ. Σάλλας	Bad
80.	18/12/2014	11:54:37	Στην Τράπεζα Πειραιώς το 24,5% της ΑΝΕΚ	Good
81.	22/12/2014	17:22:09	ΤΡΑΠΕΖΑ ΠΕΙΡΑΙΩΣ Α.Ε. : ΑΠΟΤΕΛΕΣΜΑΤΑ ΨΗΦΟΦΟΡΙΑΣ ΤΗΣ ΕΚΤΑΚΤΗΣ ΓΕΝΙΚΗΣ ΣΥΝΕΛΕΥΣΗΣ ΤΩΝ ΜΕΤΟΧΩΝ ΤΗΣ ΤΡΑΠΕΖΑΣ ΠΕΙΡΑΙΩΣ ΠΟΥ ΠΡΑΓΜΑΤΟΠΟΙΗΘΗΚΕ ΣΤΙΣ	No Mover
82.	23/12/2014	11:25:11	ΕΣΕΕ: Τόνωση ρευστότητας στις ΜμΕ από το Επενδυτικό Ταμείο	Good

`

❖ Η μετοχή της Eurobank:

	Ημερομηνία	Ώρα	Δελτίο Τύπου/ Άρθρο	Κατηγορία
1.	15/01/2014	15:46:00	**Eurobank: Καθυστερούν οι αποφάσεις για δημοσιονομικό κενό και μεταρρυθμίσεις**	Bad
2.	20/01/2014	11:27:00	**Ο χρόνος φεύγει - πάλι**	Bad
3.	29/01/2014	16:52:00	**Eurobank: Χρειάζεται λύση για το χρέος στο α' εξάμηνο**	No Mover
4.	30/01/2014	14:58:18	**Eurobank: Στις 28/2 τα ετήσια αποτελέσματα**	Good
5.	30/01/2014	14:58:33	**ΤΡΑΠΕΖΑ EUROBANK ERGASIAS A.E. : Ανακοίνωση - Ημερομηνία Ανακοίνωσης Αποτελεσμάτων Έτους 2013**	Good
6.	14/02/2014	10:51:00	**Eurobank: Σε καθοδική τροχιά ο λόγος των NPLs από το 2015**	Bad
7.	17/02/2014	10:51:00	**Eurobank: Β΄κύκλος του προγράμματος egg-enter•grow•go**	Good
8.	19/02/2014	16:26:20	**Eurobank: Διατήρηση του πρωτογενούς πλεονάσματος μόνο μέσω διαρθρωτικών μεταρρυθμίσεων**	Bad
9.	21/02/2014	11:04:00	**Reuters: Νέα κεφάλαια 5 δισ. χρειάζονται οι συστημικές τράπεζες**	Bad
10.	26/02/2014	16:52:48	**Eurobank: Σημαντικά περιθώρια αισιοδοξίας για το 2014**	Bad
11.	05/03/2014	10:32:07	**Eurobank: Στις 13 Μαρτίου τα αποτελέσματα του 2013**	No Mover
12.	05/03/2014	17:01:35	**Eurobank: Απαραίτητη η απεμπλοκή των τραπεζών από τις διαπραγματεύσεις**	Bad
13.	07/03/2014	10:35:36	**Eurobank: Σύγκληση γενικής συνέλευσης για την ΑΜΚ**	No Mover
14.	10/03/2014	16:27:00	**Περαιτέρω πτώση στις τιμές κατοικιών βλέπει η Eurobank**	Bad
15.	11/03/2014	11:51:23	**Ξεκινά κάλυψη για τις ελληνικές τράπεζες η Wood**	Good
16.	12/03/2014	17:06:00	**Eurobank: Στενεύουν τα περιθώρια για συμφωνία με την τρόικα**	Bad
17.	13/03/2014	12:43:00	**Eurobank: Χρειάζονται πολιτικές ενίσχυσης της παραγωγικότητας**	Good
18.	13/03/2014	13:32:32	**ΤΡΑΠΕΖΑ EUROBANK ERGASIAS A.E. : ΑΝΑΚΟΙΝΩΣΗ**	Bad
19.	13/03/2014	13:48:00	**Eurobank: Αναβολή λήψης απόφασης για σύγκληση Γ.Σ.**	Bad
20.	14/03/2014	16:38:00	**Eurobank: Ενδιαφέρον διεθνών επενδυτών για την ΑΜΚ**	Good
21.	20/03/2014	16:58:04	**Eurobank: Πρώτο θετικό βήμα για έξοδο στις αγορές**	Good
22.	26/03/2014	16:17:08	**Eurobank: «Ανέπαφες συναλλαγές» - νέα γενιά καρτών**	Good
23.	27/03/2014	16:58:27	**Eurobank: Επείγει η ψήφιση του νομοσχεδίου**	Good
24.	31/03/2014	12:36:00	**Eurobank: Να συνεχιστούν οι μεταρρυθμίσειςEurobank: Να συνεχιστούν οι μεταρρυθμίσεις**	No Mover
25.	02/04/2014	10:40:01	**Fitch: Αναβάθμιση προοπτικών για ελληνικά καλυμμένα ομόλογα**	Good

`

26.	03/04/2014	15:58:00	**Eurobank: Εφικτή η έξοδος στις αγορές εάν συνεχιστεί η εφαρμογή του προγράμματος**	Bad
27.	04/04/2014	13:32:04	**ΤΡΑΠΕΖΑ EUROBANK ERGASIAS A.E. : Πρόσκληση Έκτακτης Γενικής Συνέλευσης των Μετόχων 12 Απριλίου 2014**	No Mover
28.	04/04/2014	14:01:00	**Eurobank: Στις 12/4 η Γ.Σ. για την αύξηση μετοχικού κεφαλαίου**	No Mover
29.	04/04/2014	17:01:00	**Eurobank: Στηρίζει την αύξηση μετοχικού κεφαλαίου το ΤΧΣ**	Good
30.	15/04/2014	13:56:00	**Προβάδισμα της Fairfax για Eurobank**	Good
31.	16/04/2014	11:54:00	**Eurobank: Υπαρκτός ο κίνδυνος αποπληθωρισμού στην Ευρωζώνη**	Bad
32.	17/04/2014	11:55:00	**Χρ. Μεγάλου: Στόχος να ολοκληρωθεί το συντομότερο η ΑΜΚ**	Good
33.	17/04/2014	12:43:34	**Δεν βλέπει αποπληθωριστικό σπιράλ η Eurobank**	Good
34.	17/04/2014	14:01:00	**ΣΥΡΙΖΑ: Έγκλημα σε τιμή ευκαιρίας στην ΑΜΚ της Eurobank**	Bad
35.	28/04/2014	10:45:00	**Eurobank: Στις 30/4 η τελική τιμή διάθεσης των μετοχών**	Good
36.	29/04/2014	14:00:00	**Eurobank: Ενέκρινε το σχέδιο αναδιάρθρωσης η Κομισιόν**	Good
37.	02/05/2014	15:46:09	**Eurobank: Αδιαμφισβήτητη επιτυχία το πρωτογενές πλεόνασμα**	Bad
38.	06/05/2014	15:37:19	**FT: Συμμετοχή της PiMco στην ΑΜΚ της Eurobank**	Good
39.	08/05/2014	16:41:00	**Eurobank: «Κλειδί» οι διαρθρωτικές μεταρρυθμίσεις**	No Mover
40.	09/05/2014	16:51:10	**Χρ. Μεγάλου: Νέα εποχή για τη Eurobank**	No Mover
41.	12/05/2014	13:21:00	**Eurobank: Πιο βιώσιμο το χρέος χάρη στα δημοσιονομικά μέτρα**	Good
42.	21/05/2014	16:36:00	**Eurobank: Προσήλωση στις μεταρρυθμίσεις και πολιτική σταθερότητα**	Good
43.	28/05/2014	17:21:13	**ΤΡΑΠΕΖΑ EUROBANK ERGASIAS A.E. : ΑΝΑΚΟΙΝΩΣΗ ΑΠΟΤΕΛΕΣΜΑΤΩΝ Α' ΤΡΙΜΗΝΟΥ 2014**	No Mover
44.	30/05/2014	10:50:00	**Συνάντηση Γ. Στουρνάρα με τον επικεφαλής της Fairfax**	Good
45.	30/05/2014	12:29:58	**ΤΡΑΠΕΖΑ EUROBANK ERGASIAS A.E. : ΑΝΑΚΟΙΝΩΣΗ - ΔΕΛΤΙΟ ΤΥΠΟΥ**	Good
46.	30/05/2014	12:48:00	**Γουάτσα: Η Eurobank μπορεί να υπερβεί τις προσδοκίες**	Good
47.	30/05/2014	13:03:06	**Ολοκληρώθηκε η συνάντηση Γουάτσα - ΠGoodροβόπουλου**	Good
48.	03/06/2014	17:03:25	**Eurobank: Στρατηγική συνεργασία με Σκλαβενίτη**	Bad
49.	03/06/2014	17:44:07	**ΤΡΑΠΕΖΑ EUROBANK ERGASIAS A.E. : ΑΝΑΚΟΙΝΩΣΗ**	No Mover
50.	04/06/2014	15:44:04	**ΤΡΑΠΕΖΑ EUROBANK ERGASIAS A.E. : Ανακοίνωση - Γνωστοποίηση σημαντικών μεταβολών σε δικαιώματα ψήφου σύμφωνα με τον ν. 3864/2010**	Bad
51.	04/06/2014	16:16:00	**Eurobank: Με 13,58% η Fairfax**	Good

`

			- Στο 20,224% η συμμετοχή της Capital	
52.	06/06/2014	11:29:42	ΤΡΑΠΕΖΑ EUROBANK ERGASIAS A.E. : ΠΡΟΣΚΛΗΣΗ ΤΑΚΤΙΚΗΣ ΓΕΝΙΚΗΣ ΣΥΝΕΛΕΥΣΗΣ ΤΩΝ ΜΕΤΟΧΩΝ 28.06.2014	No Mover
53.	06/06/2014	11:47:49	ΤΡΑΠΕΖΑ EUROBANK ERGASIAS A.E. : ΑΝΑΚΟΙΝΩΣΗ	Bad
54.	06/06/2014	11:48:00	Eurobank: Στις 28/6 η τακτική γενική συνέλευση	Bad
55.	06/06/2014	12:01:45	ΤΡΑΠΕΖΑ EUROBANK ERGASIAS A.E. : Πρόσκληση Τακτικής Γενικής Συνέλευσης των Μετόχων 28.06.2014	No Mover
56.	10/06/2014	15:48:03	ΤΡΑΠΕΖΑ EUROBANK ERGASIAS A.E. : ΑΝΑΚΟΙΝΩΣΗ	Good
57.	10/06/2014	15:57:00	Αποχώρηση Γκ. Χαρδούβελη από Eurobank	Good
58.	13/06/2014	10:39:50	ΤΡΑΠΕΖΑ EUROBANK ERGASIAS A.E. : ΑΝΑΚΟΙΝΩΣΗ	Bad
59.	13/06/2014	12:24:07	Χρ. Μεγάλου: Η Ελλάδα πέρασε τον κάβο	Good
60.	13/06/2014	15:54:00	Eurobank: Μέσω μεταρρυθμίσεων θα αποφευχθούν νέα μέτρα	Bad
61.	19/06/2014	10:52:00	Eurobank: «Κλειδώνει» το ομόλογο των 500 εκατ. ευρώ	Good
62.	08/07/2014	17:04:00	Χρ. Μεγάλου: Προτεραιότητα η διαχείριση των NPLs	No Mover
63.	11/07/2014	15:19:35	ΤΡΑΠΕΖΑ EUROBANK ERGASIAS A.E. : ΑΝΑΚΟΙΝΩΣΗ ΗΜΕΡΟΜΗΝΙΑΣ ΑΠΟΤΕΛΕΣΜΑΤΩΝ Α' ΕΞΑΜΗΝΟΥ 2014	No Mover
64.	11/07/2014	16:07:00	Eurobank: Επικεφαλής του Retail Banking ο Θ. Καλαντώνης	No Mover
65.	11/07/2014	16:38:11	Επιτροπή Κεφαλαιαγοράς: Ανανεώθηκε η θητεία του Κ. Μποτόπουλου	No Mover
66.	11/07/2014	17:01:32	Eurobank: Παραμένουν επιφυλακτικές οι αγορές	Bad
67.	15/07/2014	16:46:37	Eurobank: Στις 29/8 τα αποτελέσματα α' εξαμήνου	No Mover
68.	17/07/2014	16:40:47	Eurobank: Οι επενδύσεις θα κρίνουν την επιστροφή στη βιώσιμη ανάπτυξη	Good
69.	31/07/2014	16:39:42	Eurobank: Οι αγορές αξιολογούν θετικά την πρόοδο της οικονομίας	Good
70.	18/08/2014	10:38:43	ΤΡΑΠΕΖΑ EUROBANK ERGASIAS A.E. : ΑΝΑΚΟΙΝΩΣΗ	Good
71.	18/08/2014	12:12:26	Eurobank: Συμφωνία πώλησης της θυγατρικής της στην Ουκρανία	Good
72.	26/08/2014	13:36:44	Θετικές συστάσεις Deutsche Bank για τις ελληνικές τράπεζες	Good
73.	29/08/2014	15:26:56	Κέρδη 26,3 εκατ. ευρώ για τη Eurobank Κύπρου	Good
74.	24/09/2014	15:44:27	ΤΡΑΠΕΖΑ EUROBANK ERGASIAS A.E. : ΕΠΙΚΑΙΡΟΠΟΙΗΜΕΝΗ ΠΑΡΟΥΣΙΑΣΗ Α' ΕΞΑΜΗΝΟΥ 2014	Bad

`

75.	29/09/2014	15:15:55	**Eurobank: Επαφές με τη διοίκηση της θυγατρικής στην Κύπρο**	Good
76.	29/09/2014	16:30:49	**Eurobank: Προϋπόθεση για μεγαλύτερες επενδύσεις η αύξηση της αποταμίευσης**	Bad
77.	01/10/2014	13:13:35	**Eurobank: Πιο βιώσιμο σήμερα το ελληνικό χρέος**	Good
78.	09/10/2014	13:21:00	**Αυξάνει τις τιμές - στόχους για Εθνική και Alpha Bank η BofA**	Bad
79.	10/10/2014	12:02:11	**Eurobank: Επενδύσεις και εξαγωγές οι βασικοί μοχλοί ανάπτυξης**	Good
80.	16/10/2014	16:57:55	**Eurobank: Αγκάθι το υψηλό ποσοστό των μακροχρόνιων ανέργων**	Bad
81.	17/10/2014	10:54:29	**ΤΡΑΠΕΖΑ EUROBANK ERGASIAS Α.Ε. : ΠΡΟΣΚΛΗΣΗ ΕΚΤΑΚΤΗΣ ΓΕΝΙΚΗΣ ΣΥΝΕΛΕΥΣΗΣ ΤΩΝ ΜΕΤΟΧΩΝ 07.11.2014**	Bad
82.	17/10/2014	11:45:00	**Συνελεύσεις στις τράπεζες για τον αναβαλλόμενο φόρο**	Bad
83.	17/10/2014	16:04:02	**ΤΡΑΠΕΖΑ EUROBANK ERGASIAS Α.Ε. : ΠΡΟΣΚΛΗΣΗ ΕΚΤΑΚΤΗΣ ΓΕΝΙΚΗΣ ΣΥΝΕΛΕΥΣΗΣ ΤΩΝ ΜΕΤΟΧΩΝ 7 ΝΟΕΜΒΡΙΟΥ 2014**	Good
84.	22/10/2014	15:54:11	**Ανάπτυξη 1% το γ' τρίμηνο βλέπει η Eurobank**	Good
85.	23/10/2014	16:36:07	**Eurobank: Τροχοπέδη για την ανάκαμψη η πολιτική αβεβαιότητα**	Bad
86.	27/10/2014	14:03:00	**Χωρίς κεφαλαιακές ανάγκες η Eurobank**	Good
87.	29/10/2014	11:04:00	**Ορόσημο για τις τράπεζες η απεμπλοκή από τα κρατικά κεφάλαια**	Good
88.	30/10/2014	13:32:00	**Eurobank: Με άτοκες δόσεις η αγορά πετρελαίου θέρμανσης από την ΕΚΟ**	Good
89.	07/11/2014	11:10:00	**«Πράσινο» από Alpha Bank, Eurobank για τον αναβαλλόμενο φόρο**	No Mover
90.	19/11/2014	14:13:00	**Eurobank: Διπλασιασμός χορηγήσεων στα 2 δισ. ευρώ**	Good
91.	28/11/2014	16:52:18	**Eurobank: Βελτιώνεται το γενικό οικονομικό κλίμα**	No Mover
92.	02/12/2014	13:01:40	**Eurobank: Συνεργασία της Business Exchanges με Mellon Technologies**	Good
93.	02/12/2014	15:38:00	**Χρ. Μεγάλου: Εστιάζουμε στη χρηματοδότηση της οικονομίας**	Good
94.	03/12/2014	16:36:00	**Eurobank: Πρόγραμμα επιβράβευσης αριστούχων μαθητών**	No Mover
95.	11/12/2014	15:30:54	**Eurobank: «Όχι» σε πάγωμα των μεταρρυθμίσεων**	Bad
96.	15/12/2014	13:09:59	**Λάμψα: Τροποποίηση συμβολαίων εγγραφής υποθηκών**	Bad
97.	22/12/2014	16:32:34	**Eurobank: Τρία βραβεία για την εξυπηρέτηση πελατών**	Good
98.	23/12/2014	11:25:11	**ΕΣΕΕ: Τόνωση ρευστότητας στις ΜμΕ από το Επενδυτικό Ταμείο**	Good

`

Π.2.1. ΣΥΓΚΕΝΤΡΩΤΙΚΟΣ ΠΙΝΑΚΑΣ ΚΕΙΜΕΝΙΚΩΝ ΔΕΔΟΜΕΝΩΝ

Συγκεντρωτικά, ο παρακάτω πίνακας δίνει τον ακριβή αριθμό κατηγοριοποίησης των δελτίων τύπου για κάθε μετοχή:

Όνομα Τράπεζας	*Good News*	*Bad News*	*No Movers*	*Σύνολο*
Alphabank	35	22	34	91
Attica Bank	7	3	8	18
Εθνική	33	23	29	85
Ελλάδος	22	22	15	59
Κύπρου	1	3	-	4
Πειραιώς	44	24	14	82
Eurobank	47	31	20	98

Π.2 ΠΑΡΟΥΣΙΑΣΗ ΑΡΙΘΜΗΤΙΚΩΝ ΔΕΔΟΜΕΝΩΝ

Προκειμένου για υβριδικά δεδομένα, πέρα από τα κειμενικά που παρατέθηκαν παραπάνω, είναι αναγκαία και η παράθεση των αριθμητικών δεδομένων. Συνεπώς, θα παρουσιαστούν και οι χρονοσειρές των τιμών κλεισίματος για κάθε μετοχή για τις ημερομηνίες κατά τις οποίες εκδόθηκαν δελτία τύπου, ανεξάρτητα, βέβαια, από το αντίκτυπό τους στην τιμή της μετοχή καθώς και ο όγκος:

- Για την Alphabank έχουμε:

Trade Date	***Close***	***Volume***
9/1	0,709	22434800
24/1	0,646	10570749
31/1	0,670	2344044
3/2	0,679	4477553
7/2	0,703	9931147
14/2	0,699	13725551
21/2	0,702	13915049
6/3	0,712	19481169
11/3	0,684	28863666
13/3	0,727	20246815
18/3	0,728	19805364
21/3	0,738	19935035
26/3	0,730	83765556
27/3	0,730	25918259
28/3	0,725	17806380
31/3	0,715	16312966
1/4	0,700	20447076
2/4	0,695	77702228
3/4	0,729	17028651
4/4	0,720	17317358
8/4	0,723	12842501
10/4	0,733	25892155
17/4	0,695	19519851
24/4	0,680	9459751
25/4	0,687	16979277
30/4	0,697	14258903
6/5	0,690	10974345
8/5	0,696	14711152
15/5	0,615	14077926
16/5	0,618	8158391
22/5	0,655	14299690
29/5	0,698	28056159
2/6	0,711	79891777
6/6	0,760	22506571
13/6	0,715	30208382
18/6	0,716	20871114
30/6	0,680	47401272
3/7	0,705	28997594
9/7	0,620	22187437
11/7	0,617	6947988
21/8	0,632	43724360
26/8	0,653	38816448
27/8	0,645	10302523
28/8	0,656	21740492
29/8	0,666	11156113
12/9	0,670	25205221
15/9	0,675	41588546
19/9	0,650	15481715

`

26/9	0,640	46264299
3/10	0,630	20856035
6/10	0,620	22546808
9/10	0,621	24066182
14/10	0,558	16243557
17/10	0,570	41763881
29/10	0,560	7986960

4/11	0,534	8450696
5/11	0,536	11639764
7/11	0,520	21865127
11/11	0,520	14551836
19/11	0,530	45215031
28/11	0,522	66709295
3/12	0,560	27935517

4/12	0,539	18413798
5/12	0,565	8623658
12/12	0,470	27755808
15/12	0,484	15174163
31/12	0,468	45646163

- Για την Attica Bank έχουμε:

Trade Date	***Close***	***Volume***
6/2	0,221	1672535
7/2	0,222	857384
18/3	0,150	3211979
24/3	0,172	1461788

19/5	0,125	982235
27/5	0,128	2283688
18/6	0,147	2466192
29/7	0,127	1202000
4/8	0,130	479558
9/9	0,112	2313687

30/9	0,082	1977839
13/10	0,072	518106
14/10	0,066	1354219
10/11	0,057	1021464
12/11	0,052	1772390
10/12	0,058	9645761

- Για την Εθνική Τράπεζα Ελλάδος η χρονοσειρά είναι η εξής:

Trade Date	***Close***	***Volume***
2/1	4,12	1115576
9/1	4,26	1341310
29/1	3,16	4110951
30/1	3,33	2407287
10/2	3,78	1769958
17/2	3,66	845800
21/2	3,50	1110359
10/3	3,66	3109994
11/3	3,98	4835599
26/3	3,93	923370
14/4	3,26	5311321
15/4	3,04	3705602
16/4	2,97	2061996
22/4	3,18	807042
24/4	3,10	1677779
6/5	2,81	2417865
8/5	2,58	5644799
13/5	2,50	9970582
14/5	2,42	16645638
16/5	2,12	39734809

20/5	2,14	58297227
21/5	2,19	38010504
4/6	2,61	16302978
5/6	2,75	39588405
6/6	2,90	58708151
16/6	2,75	9791261
20/6	2,84	8094787
24/6	2,75	9720365
26/6	2,68	11841495
27/6	2,64	13627780
30/6	2,67	7005033
1/7	2,74	16900115
4/7	2,75	6051429
7/7	2,73	6189106
11/7	2,47	6155142
18/7	2,48	9050067
23/7	2,41	5092280
24/7	2,47	8826261
28/7	2,57	3923073
20/8	2,48	6315314
26/8	2,73	9967834
27/8	2,68	9235742

29/8	2,61	6153364
17/9	2,45	8796882
19/9	2,45	17085169
29/9	2,28	6010264
30/9	2,32	8983034
9/10	2,20	9808168
17/10	2,04	12522149
21/10	2,18	11159126
22/10	2,29	13629915
24/10	2,31	10822988
27/10	2,13	16867925
29/10	2,00	17395375
7/11	1,82	15051966
12/11	1,61	17291409
19/11	1,86	9064801
2711	1,86	7628531
16/12	1,45	14762110
19/12	1,52	9290577
23/12	1,55	6233688
30/12	1,43	4676768

`

- Ομοίως για τη Τράπεζα της Ελλάδος:

Trade Date	*Close*	*Volume*
3/1	15,70	6170
8/1	16,40	8424
13/1	15,92	6019
14/1	15,62	9316
16/1	16,15	5567
17/1	16,01	1479
28/1	14,90	16864
27/2	15,44	12664
4/3	15,79	6347
5/3	15,50	5118
7/3	15,53	3829
10/4	15,56	6353
11/4	15,30	2493
14/4	14,92	5363
17/4	15,50	3373
5/5	14,76	1654
6/5	14,78	2455
14/5	14,38	3465
20/5	13,99	6670
23/5	13,95	3142
30/5	14,15	2004
12/6	14,97	4329
13/6	14,74	2321
16/6	14,74	2904
17/6	14,61	3024
20/6	14,20	5871
27/6	14,22	3767
15/7	13,62	2594
18/7	13,60	3585
25/7	13,88	4056
6/8	13,41	3254
11/8	13,24	775
25/8	13,44	1604
28/8	13,45	512
23/9	12,74	2170
3/12	10,39	2618
10/12	9,98	5454
15/12	9,59	3673
22/12	9,62	2047
30/12	9,00	6406

- Για τη Τράπεζα Κύπρου:

Trade Date	*Close*	*Volume*
16/12	0,177	10348328
17/12	0,183	10298807
31/12	0,215	533180

- Για τη Πειραιώς έχουμε:

Trade Date	*Close*	*Volume*
2/1	1,55	3142341
9/1	1,68	9522934
16/1	1,84	14313809
31/1	1,73	4926185
21/2	1,94	6702707
24/2	1,90	6352414
26/2	1,98	9206125
5/3	2,09	9677758
6/3	1,99	11418891
11/3	1,84	9738567
18/3	20	16997672
20/3	2,05	9510518
26/3	1,90	25621685
27/3	1,84	23618241
28/3	1,90	8497860
31/3	2,00	16856355
1/4	1,98	9310397
2/4	1,99	10630341
7/4	1,87	5422006
16/4	1,75	165470815
17/4	1,71	29161999
24/4	1,72	10735001
6/5	1,75	8442917
16/5	1,45	42282749
19/5	1,39	27170141
22/5	1,63	7599191
28/5	1,73	11116978
2/6	1,82	14948884
4/6	1,81	4887645
13/6	1,70	7479484
3/7	1,74	6638542
11/7	1,52	13039827
16/7	1,62	5285799
22/7	1,46	13943074
31/7	1,58	7779260
26/8	1,53	8191005
27/8	1,54	6955758
2/9	1,38	10775639
23/9	1,33	11482195
24/9	1,30	10934387
26/9	1,30	9010642

`

2/10	1,37	18437783
7/10	1,24	15349577
8/10	1,18	10160994
9/10	1,23	13959605
10/10	1,20	14209826
13/10	1,23	13452610

17/10	1,06	29447764
29/10	1,19	26307521
3/11	1,16	7111600
7/11	1,07	11845254
19/11	1,19	13721226
20/11	1,20	9484849

21/11	1,22	9024403
25/11	1,15	9558522
28/11	1,22	7637422
18/12	0,935	9672280
22/12	1,00	7732306
23/12	0,97	12515976

- Τέλος, για την Eurobank, η χρονοσειρά είναι η ακόλουθη:

Trade Date	***Close***	***Volume***
15/1	0,589	1648111
20/1	0,562	704434
29/1	0,447	1704621
30/1	0,495	1553446
14/2	0,505	617657
17/2	0,534	1977885
19/2	0,512	996210
21/2	0,50	507171
26/2	0,515	1383579
5/3	0,455	1917623
7/3	0,405	4748447
10/3	0,438	3629635
11/3	0,440	1928950
12/3	0,430	1391144
13/3	0,430	2329453
14/3	0,438	1508015
20/3	0,490	2121788
26/3	0,468	1970593
27/3	0,470	1474423
31/3	0,484	3645443
2/4	0,470	1340650
3/4	0,458	1440626
4/4	0,453	1720390
15/4	0,406	2186726

16/4	0,381	3600410
17/4	0,394	2076493
28/4	0,369	4747498
29/4	0,390	4863938
2/5	0,483	20915981
6/5	0,360	1,97E+08
8/5	0,366	54473843
9/5	0,351	4,91E+08
12/5	0,350	1,03E+08
21/5	0,360	45942107
28/5	0,370	40483654
30/5	0,410	6,56E+08
3/6	0,405	49647861
4/6	0,411	52949208
6/6	0,420	80024378
10/6	0,434	1E+08
13/6	0,410	21040921
19/6	0,390	34023899
8/7	0,359	31366243
11/7	0,353	18611776
15/7	0,350	40902904
17/7	0,354	6461026
31/7	0,336	14471320
18/8	0,305	19004512
26/8	0,333	23930817

29/8	0,323	14241678
24/9	0,329	34670551
29/9	0,310	22538604
1/10	0,307	48397285
9/10	0,298	28042087
10/10	0,290	56380895
16/10	0,247	50454776
17/10	0,265	66061214
22/10	0,283	14469494
23/10	0,280	26218172
27/10	0,290	49772339
29/10	0,280	31531372
30/10	0,266	26680537
7/11	0,241	21931359
19/11	0,273	19913577
28/11	0,249	21686186
2/12	0,260	31138800
3/12	0,268	33190095
11/12	0,215	81749704
15/12	0,203	22622535
22/12	0,207	29510221
23/12	0,207	13235842

`

Π.3 ΛΙΣΤΑ STOPWORDS

1.	αδιάκοπα
2.	αι
3.	ακόμα
4.	ακόμη
5.	ακριβώς
6.	αλήθεια
7.	αληθινά
8.	αλλά
9.	αλλάχου
10.	άλλες
11.	άλλη
12.	άλλην
13.	άλλης
14.	αλλιώς
15.	αλλιώτικα
16.	άλλο
17.	άλλοι
18.	αλλιώς
19.	αλλιώτικα
20.	άλλον
21.	άλλος
22.	άλλοτε
23.	αλλού
24.	άλλους
25.	άλλων
26.	άμα
27.	άμεσα
28.	αμέσως
29.	αν
30.	ανά
31.	ανάμεσα
32.	αναμεταξύ
33.	άνευ
34.	αντί
35.	αντίπερα
36.	αντί
37.	άνω
38.	ανωτέρω
39.	άξαφνε
40.	απ
41.	απέναντι
42.	από
43.	απόψε
44.	άρα
45.	άραγε
46.	αργά
47.	αργότερο
48.	αριστερά
49.	αρκετά
50.	αρχικά
51.	ας
52.	αύριο
53.	αυτά
54.	αυτές
55.	αυτή
56.	αυτήν
57.	αυτής
58.	αυτό
59.	αυτοί
60.	αυτόν
61.	αυτός
62.	αυτού
63.	αυτούς
64.	αυτών
65.	αφότου
66.	αφού
67.	βεβαία
68.	βεβαιότατα
69.	γι
70.	για
71.	γρήγορα
72.	γύρω
73.	δα
74.	δε
75.	δεινά
76.	δεν
77.	δεξιά
78.	δήθεν
79.	δηλαδή
80.	δι
81.	δια
82.	διαρκώς
83.	δικά
84.	δικό
85.	δικοί
86.	δικός
87.	δικού
88.	δικούς
89.	διόλου
90.	δίπλα
91.	δίχως
92.	εάν
93.	εαυτό
94.	εαυτόν
95.	εαυτού
96.	εαυτούς
97.	εαυτών
98.	έγκαιρα
99.	εγκαίρως
100.	εγώ
101.	άδω
102.	ειδεμή
103.	είθε
104.	είμαι
105.	είμαστε
106.	είναι
107.	εις
108.	είσαι
109.	είσαστε
110.	είστε
111.	είτε

`

112.	είχα
113.	είχαμε
114.	είχαν
115.	είχατε
116.	είχε
117.	είχες
118.	έκαστα
119.	έκαστες
120.	έκαστη
121.	εκάστην
122.	εκάστης
123.	έκαστο
124.	έκαστοι
125.	έκαστον
126.	έκαστος
127.	έκαστου
128.	έκαστους
129.	εκάστων
130.	εκεί
131.	εκείνα
132.	εκείνες
133.	εκείνη
134.	εκείνην
135.	εκείνης
136.	εκείνο
137.	εκείνοι
138.	εκείνον
139.	εκείνος
140.	εκείνου
141.	εκείνους
142.	εκείνων
143.	εκτός
144.	εμάς
145.	εμείς
146.	έμενα
147.	εμπρός
148.	εν
149.	ένα
150.	έναν
151.	ένας

152.	ενός
153.	εντελώς
154.	όντος
155.	εντωμεταξύ
156.	ενώ
157.	εξ
158.	έξανα
159.	έξης
160.	εξίσου
161.	έξω
162.	επάνω
163.	επειδή
164.	έπειτα
165.	επί
166.	επίσης
167.	επομένως
168.	εσάς
169.	εσείς
170.	εσένα
171.	έστω
172.	εσύ
173.	ετέρα
174.	έτεραι
175.	ετέρας
176.	ετέρες
177.	έτεροι
178.	έτερες
179.	έτερο
180.	έτεροι
181.	έτερον
182.	έτερος
183.	έτερου
184.	ετέρους
185.	ετέρων
186.	ετούτε
187.	ετούτε
188.	ετούτε
189.	τούτην
190.	τούτης
191.	ετούτε

192.	τούτοι
193.	τούτον
194.	τούτος
195.	τούτου
196.	τούτους
197.	τούτων
198.	έτσι
199.	εύγε
200.	ευθύς
201.	ευτυχώς
202.	έφεξες
203.	έχει
204.	έχεις
205.	έχετε
206.	εχθές
207.	έχομε
208.	έχουμε
209.	έχουν
210.	εχτές
211.	έχω
212.	έως
213.	η
214.	ήδη
215.	ήμασταν
216.	ήμαστε
217.	ήμουν
218.	ήσασταν
219.	ήσαστε
220.	ήσουν
221.	ήσαν
222.	ήταν
223.	ήτοι
224.	ήττον
225.	θα
226.	ι
227.	ίδια
228.	ίδιαν
229.	ίδιας
230.	ίδιες
231.	ίδιο

`

232.	ίδιοι
233.	ίδιον
234.	ίδιος
235.	ίδιου
236.	ίδιους
237.	ίδιων
238.	ιδίως
239.	ιέ
240.	ιοί
241.	ίσαμε
242.	ίσια
243.	ίσως
244.	κάθε
245.	καθεμία
246.	καθεμίας
247.	καθένα
248.	καθένας
249.	καθενός
250.	καθετί
251.	καθόλου
252.	καθώς
253.	και
254.	κακά
255.	κακώς
256.	καλά
257.	καλώς
258.	καμιά
259.	καμιάν
260.	καμιάς
261.	καμπόσα
262.	καμπόσες
263.	κάμποση
264.	καμπόσην
265.	καμπόσης
266.	κάμποσο
267.	καμπόσοι
268.	κάμποσον
269.	κάμποσος
270.	κάμποσου
271.	καμπόσους

272.	καμπόσων
273.	κάνεις
274.	κανών
275.	κανένα
276.	κανέναν
277.	κανένας
278.	κανενός
279.	κάποια
280.	κάποιαν
281.	κάποιας
282.	κάποιες
283.	κάποιο
284.	κάποιοι
285.	κάποιον
286.	κάποιος
287.	κάποιου
288.	κάποιους
289.	κάποιων
290.	κάποτε
291.	κάπου
292.	κάπως
293.	κατ
294.	κατά
295.	κάτι
296.	κατιτί
297.	κατόπιν
298.	κάτω
299.	κιόλας
300.	κλπ
301.	κοντά
302.	κτλ
303.	κυρίως
304.	λιγάκι
305.	λίγο
306.	λιγότερο
307.	λογά
308.	λοιπά
309.	λοιπόν
310.	μα
311.	μαζί

312.	μακάρι
313.	μακριά
314.	μάλιστα
315.	μάλλον
316.	μας
317.	με
318.	μεθαύριο
319.	μείον
320.	μέλει
321.	μέλλεται
322.	μεμιάς
323.	μεν
324.	μερικά
325.	μερικές
326.	μερικοί
327.	μερικούς
328.	μερικών
329.	μέσα
330.	μετ
331.	μετά
332.	μεταξύ
333.	μέχρι
334.	μη
335.	μήδε
336.	μην
337.	μήπως
338.	μήτε
339.	μια
340.	μιαν
341.	μιας
342.	μόλις
343.	μολονότι
344.	μοναχά
345.	μονές
346.	μονή
347.	μόνην
348.	μόνης
349.	μονό
350.	μονοί
351.	μονομιάς

`

352.	μόνος
353.	μονού
354.	μόνους
355.	μονών
356.	μου
357.	μπορεί
358.	μπορούν
359.	μπράβο
360.	μπρος
361.	να
362.	ναι
363.	νωρίς
364.	ξανά
365.	ξαφνικά
366.	ο
367.	οι
368.	όλα
369.	όλες
370.	όλη
371.	όλων
372.	όλης
373.	όλο
374.	ολόγυρα
375.	όλοι
376.	όλον
377.	ολονέν
378.	όλος
379.	ολότελα
380.	όλου
381.	όλους
382.	όλων
383.	όλως
384.	ολωσδιόλου
385.	όμως
386.	οποία
387.	οποιαδήποτε
388.	όποιαν
389.	οποιανδήποτε
390.	οποίας
391.	οποιασδήποτε
392.	οποιοδήποτε
393.	οποίες
394.	οποιεσδήποτε
395.	όποιο
396.	οποιοιδήποτε
397.	όποιοι
398.	όποιον
399.	οποιονδήποτε
400.	οποίος
401.	οποιοσδήποτε
402.	όποιου
403.	οποιουδήποτε
404.	οποίους
405.	οποιουσδήποτε
406.	οποίων
407.	οποιωνδήποτε
408.	όποτε
409.	οποτεδήποτε
410.	όπου
411.	οπουδήποτε
412.	όπως
413.	ορισμένα
414.	ορισμένες
415.	ορισμένων
416.	ορισμένως
417.	όσα
418.	οσαδήποτε
419.	όσες
420.	οσεσδήποτε
421.	όση
422.	οσηδήποτε
423.	όσην
424.	οσηνδήποτε
425.	όσης
426.	οσησδήποτε
427.	όσο
428.	οσοδήποτε
429.	όσοι
430.	οσοιδήποτε
431.	όσον
432.	οσονδήποτε
433.	όσος
434.	οσοσδήποτε
435.	όσου
436.	οσουδήποτε
437.	όσους
438.	οσουσδήποτε
439.	όσων
440.	οσωνδήποτε
441.	όταν
442.	ότι
443.	οτιδήποτε
444.	ότου
445.	ου
446.	ουδέ
447.	ούτε
448.	όχι
449.	πάλι
450.	πάντοτε
451.	παντού
452.	πάντως
453.	παρά
454.	περά
455.	περί
456.	περίπου
457.	περισσότερο
458.	πέρσι
459.	πέρυσι
460.	πια
461.	πιθανόν
462.	πιο
463.	πίσω
464.	πλάι
465.	πλέον
466.	πλην
467.	ποια
468.	ποιαν

`

469.	ποιας
470.	ποιες
471.	ποιο
472.	ποιοι
473.	ποιον
474.	ποιος
475.	ποιου
476.	ποιους
477.	ποιων
478.	πολύ
479.	πόσες
480.	πόση
481.	πόσην
482.	πόσης
483.	πόσοι
484.	πόσος
485.	πόσους
486.	ποτέ
487.	που
488.	πούθε
489.	πουθενά
490.	πρέπει
491.	πριν
492.	προ
493.	προκειμένου
494.	πρόκειται
495.	πρόπερσι
496.	προς
497.	προτού
498.	προχθές
499.	προχθές
500.	πρωτύτερα
501.	πως
502.	σαν
503.	σας
504.	σε
505.	σεις
506.	σήμερα
507.	σιγά
508.	σου

509.	στα
510.	στη
511.	στην
512.	στης
513.	στις
514.	στο
515.	στον
516.	στου
517.	στους
518.	στων
519.	συγχρόνως
520.	συν
521.	συνάμα
522.	συνεπώς
523.	συνήθως
524.	συχνά
525.	συχνός
526.	συχνές
527.	συχνή
528.	συχνήν
529.	συχνής
530.	συχνό
531.	συχνοί
532.	συχνών
533.	συχνός
534.	συχνού
535.	συχνού
536.	συχνούς
537.	συχνών
538.	συχνός
539.	σχεδόν
540.	σωστά
541.	τα
542.	τάδε
543.	ταύτα
544.	ταύτες
545.	ταύτη
546.	ταυτόν
547.	ταύτης
548.	ταύτο,

	ταύτον
549.	ταύτος
550.	ταύτου
551.	ταύτων
552.	τάχα
553.	τύχατε
554.	τελικά
555.	τελικός
556.	τες
557.	τέτοια
558.	τέτοιαν
559.	τέτοιας
560.	τέτοιες
561.	τέτοιο
562.	τέτοιοι
563.	τέτοιων
564.	τη
565.	την
566.	της
567.	τι
568.	τίποτα
569.	τίποτε
570.	τις
571.	το
572.	τού
573.	τον
574.	τας
575.	τόσα
576.	τόσες
577.	τόση
578.	τόσην
579.	τόσης
580.	τόσο
581.	τόσοι
582.	τόσον
583.	τόσος
584.	τόσου
585.	τόσους
586.	τόσων
587.	τότε

`

588.	του
589.	τουλάχιστο
590.	τουλάχιστον
591.	τους
592.	τούτα
593.	τούτες
594.	τούτη
595.	τούτην
596.	τούτης
597.	τούτο
598.	τούτοι
599.	τούτοις
600.	τούτον
601.	τούτος

602.	τούτου
603.	τούτους
604.	τούτων
605.	τυχόν
606.	των
607.	τώρα
608.	υπ
609.	υπέρ
610.	υπό
611.	υπόψη
612.	υπόψιν
613.	ύστερα
614.	φέτος
615.	χαμηλά

616.	χθες
617.	χτες
618.	χωρίς
619.	χωριστά
620.	ψηλά
621.	ω
622.	ωραία
623.	ως
624.	ωσάν
625.	ωσότου
626.	ώσπου
627.	ώστε
628.	ωστόσο
629.	ωχ

`

ΒΙΒΛΙΟΓΡΑΦΙΑ

- *Basic Text Mining in R.* (n.d.). Retrieved from rstudio-pubs-static.s3.amazonaws.com: *https://rstudio-pubs static.s3.amazonaws.com/31867_8236987cf0a8444e962ccd2aec46d9c3.html#k-means-clustering*
- Breiman, L., Friedman, J., Olsen, R., & Stone, C. (1984). *Classification and Regression Trees.* FL: CRC Press.
- Brock, G., Pihur, V., Datta, S., & Datta, S. (2008). *clValid: An R Package for Cluster Validation.* Retrieved from Journal of Statistical Software: *http://www.jstatsoft.org/*
- Choudhury, P. R. (2015). *Analytics Vidhya| Learn Everything About Analytics*. Retrieved from *http://www.analyticsvidhya.com/blog/2015/08/learning-concept-knn-algorithms-programming/*
- Cover, T. M., & P.E.Hart. (1967). *Nearest Neghbor Pattern Classification.* IEEE Trans. Inform. Theory.
- *DnI Institute|Build Data And Decision Science Experience*. (2014). Retrieved from *http://dni-institute.in/blogs/decision-tree-using-rpart-in-r/*
- Fayyad et al, U. (1996). *Advances in Knowledge Discovery and Data Mining.* AAAI Press.
- Garonfolo, H. J. (2015). *R: Text Classification using a K Nearest Neighbour Model*. Retrieved from garonfolo.dk: *http://garonfolo.dk/herbert/2015/05/r-text-classification-using-a-k-nearest-neighbour-model/*
- Hahsler, M. (2016). *R Code for Chapter 8 of Introduction to Data Mining: Clustering*. Retrieved from michael.hahsler.net: *http://michael.hahsler.net/SMU/EMIS7332/R/chap8.html*
- Han, J., & M.Kamber. (2001). *Data Mining: Concepts and Techniques.* Morgan Kaufman Publishers.
- Hilbe, J. (2009). *Logistic Regression Models.* Chapman & Hall/ CRC Press.
- Janson, G., & Gaur, M. (2015). *Quora*. Retrieved from In Text Mining, Why Should We Remove The Sparse Term From The Document Term Matrix?: *https://www.quora.com/In-text-mining-why-should-we-remove-the-sparse-term-from-the-document-term-matrix*
- Khalifa, M. b. (2015). *benKhalifa.com*. Retrieved from How I used R to create a word cloud: *http://www.benkhalifa.com/tm-wordcloud-R-english-spanish*
- Krzanowski, W. (1988). *Principles of Multivariate Analysis: A User' s Perspective.* New York: Oxford University Press.

`

- Liu, W. (2014). *beyondvalence.blogspot.gr*. Retrieved from Text Mining: 5. Hierarchical Clustering for Frequent Terms in R: *http://beyondvalence.blogspot.gr/2014/01/text-mining-5-hierarchical-clustering.html*
- M.Bramer. (2007). *Principles of Data Mining. Undergraduate Topics in Computer Science*. Springer.
- Ma, M. (2014). *RStudio Pubs Static*. Retrieved from Basic Text Mining in R*: https://rstudio-pubs-static.s3.amazonaws.com/31867_8236987cf0a8444e962ccd2aec46d9c3.html#word-clouds*
- Ma, M. (2014). *RStudio Pumbs Static*. Retrieved from *http://rstudio-pubs-static.s3.amazonaws.com/27179_e64f0de316fc4f169d6ca300f18ee2aa.html*
- Mitchell, T. (1997). *Machine Learning*. McGraw Hill.
- *Positron Investments*. (n.d.). Retrieved from *http://positron-investments.com*
- *R Statistics.net|TheNo.1 Educational Reference For Statistical Computing With R~ Seek.Learn.Grow*. (2015). Retrieved from *http://rstatistics.net/discriminant-analysis/*
- R. I. Kabacoff, P. (2014). *Quick-R|accesing the power of R*. Retrieved from *http://www.statmethods.net/advgraphs/trellis.html*
- R.Kelly. (2014). *RPumbs*. Retrieved from *http://rpubs.com/ryankelly/LDA-QDA*
- S.V.Stehman. (1997). *Selecting and Intepreting Measures of Thematic Classification Accuracy*. Elesvier.
- Seber, G. A. (1984). *Multivariate Observations*. NJq John Wiley & Sons, Inc.
- Shakhnarovich, G., Darell, T., & Indyk, P. (2005). *Nearest-Neighbor Methods in Learning and Vision: Theory and Practise*. MIT Press.
- *STHDA| Statistical Tools For High-Throuput Data Analysis*. (2015). Retrieved from Text Mining And Word Cloud Fundamentals in R: 5 Steps You Should Know: *http://www.sthda.com/english/wiki/text-mining-and-word-cloud-fundamentals-in-r-5-simple-steps-you-should-know*
- *TRANSLATUM| The Greek Translation Vortal*. (2006). Retrieved from Topic: Greek stop-words (stopwords, stop words) : *http://www.translatum.gr/forum/index.php?topic=3550.0*
- *Validate cluster analysis in R*. (2013). Retrieved from CrossValidated: *http://stats.stackexchange.com/questions/64790/validate-cluster-analysis-in-r*
- *Wordpress.com*. (2015). Retrieved from Eight ti Late|A gentle Introduction To Text Using R: *https://eight2late.wordpress.com/2015/05/27/a-gentle-introduction-to-text-mining-using-r/*
- Δόβα, Σ. (2013, Νοέμβριος). Μελέτη Αλγορίθμων Κατηγοριοποίησης με Χρήση του Matlab. Θεσσαλονίκη.

`

- Κούτρας, Μ. (2007). *Εφαρμομένη Πολυμεταβλητή Ανάλυση:Ανάλυση κατά Συστάδες.* Πανεπιστήμιο Πειραιώς: Π.Μ.Σ. στην Εφαρμοσμένη Στατιστική.
- Νασίκας, Ι. (2006, Ιούνιος). Μια προτεινόμενη μέθοδος με τη χρήση κανόνων συσχέτισης. Πάτρα.
- *ΝΑΥΤΕΜΠΟΡΙΚΗ*. (n.d.). Retrieved from *http://www.naftemporiki.gr/finance/sector/IB1130/trapezes*
- Σταυλιώτης, Γ. Ν. (2008, Ιούνιος). Retrieved from Εξόρυξη Δεδομένων (Data Mining) σε κατηγορικά δεδομένα.