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

ΤΜΗΜΑ Σ Τ Α Τ Ι Σ Τ Ι Κ Η Σ
ΚΑΙ ΑΣΦΑΛΙΣΤΙΚΗΣ Ε Π Ι Σ Τ Η Μ Η Σ

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΣΤΗΝ ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ

ΧΡΗΣΗ ΤΕΧΝΙΚΩΝ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ ΣΤΟΝ ΑΝΑΛΟΓΙΣΜΟ

Μιχάλης Μακρής

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 Αναλογιστική Επιστήμη και Διοικητική Κινδύνου

Πειραιάς
Φεβρουάριος 2015

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:

- Καθηγητής Μ. Κούτρας (Επιβλέπων)
- Καθηγητής Κ. Τσίμπος
- Λέκτορας Ν. Πελέκης

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

U N I V E R S I T Y O F P I R A E U S

DEPARTMENT OF STATISTICS AND INSURANCE SCIENCE

POSTGRADUATE PROGRAM IN APPLIED STATISTICS

DATA MINING TECHNIQUES IN ACTUARY AND INSURANCE

By

Michael Makris

MSc Dissertation

submitted to the Department of Statistics and Insurance Science of the University of Piraeus in partial fulfilment of the requirements for the degree of Master of Science in Actuarial Science and Risk Management

Piraeus, Greece
February 2014

Περίληψη

Περίπου το 80% των δεδομένων που διακινείται καθημερινά στον κόσμο είναι αδόμητα. Τα σύγχρονα συστήματα υποστήριξης λήψης επιχειρηματικών αποφάσεων βασίζονται σε πληροφόρηση που προέρχεται κυρίως από δομημένα δεδομένα, αγνοώντας τα αδόμητα τα οποία όμως μπορούν να προσφέρουν σημαντική πληροφορία. Υπάρχει επομένως η ανάγκη νέων δυναμικών εργαλείων μετατροπής των αδόμητων δεδομένων σε δομημένα, τα οποία σε συνδυασμό με τα εξαρχής δομημένα δεδομένα να βοηθήσει τους καταναλωτές πληροφοριών να λαμβάνουν καλύτερες αποφάσεις. Αυτή την ανάγκη καλείται να καλύψει ένας νέος κλάδος της επιστήμης η εξόρυξη από κείμενα (text mining) που είναι συνδυασμός ετερόκλητων επιστημονικών πεδίων όπως της στατιστικής, της μηχανικής εκμάθησης, της θεωρίας της πληροφορίας και των υπολογιστικών διαδικασιών.

Στην παρούσα εργασία εξετάζονται δύο εφαρμογές του Text Mining κυρίως στην περιοχή του ασφαλιστικού χώρου. Στην πρώτη εφαρμογή παρουσιάζεται αναλυτικά η ομαδοποίηση που οργανώνει τις αδόμητες πληροφορίες σε ομάδες βοηθώντας καθοριστικά στην ανάλυση τους και στην εξαγωγή χρήσιμων συμπερασμάτων ενώ στην δεύτερη χρησιμοποιούνται τεχνικές πρόβλεψης σε κειμένα όπως παραδείγματός χάριν αν ένα σχόλιο είναι αρνητικό ή θετικό ή αν μία ασφαλιστική δήλωση είναι απάτη ή όχι. Συγχρόνως παρουσιάζονται οι σημαντικότερες έννοιες και μέθοδοι που χρησιμοποιούνται κατά τη διάρκεια της ομαδοποίησης και πρόβλεψης.

Σκοπός της Διπλωματικής εργασίας είναι, εξηγώντας τις μεθόδους του Text Mining, να δείξουμε πόσο χρήσιμο εργαλείο είναι για κάθε ασφαλιστική επιχείρηση και με πόσο απλό και εύκολο τρόπο μπορεί να προβεί στην εξαγωγή σημαντικών συμπερασμάτων από αδόμητα και δύσκολα στη χρήση τους αρχεία.

Abstract

Almost 80% of the data utilized on a daily basis all over the world are unstructured. Modern decision support systems are based on information extracted from structured data, thereof neglecting unstructured data that can also provide significant information. Therefore there is a need for new dynamic tools that will help to transform unstructured data into structured information which, combined with the information extracted from structured data will help information consumers make better decisions. A new field of science, named Text Mining promises to fill this gap. Text Mining is a combination of different sciences such as statistics, machine learning, information theory and computational procedures.

In this thesis we examine two Text Mining applications which are of major inportance for the insurance sector. In the first text we illustrate how can be classified into groups based on text subjects or according to other attributes (common words, frequency of words etc.). This application is very useful for document analysis and useful information extraction.

In the second we present how forecasting techniques are applied in text mining to predict, for example if a document contains positive or negative comments or if an insurance claim is fraud or not. At the same time we present the most important concepts and methods used by classification and prediction techniques in text mining.

The aim of this thesis is to present how useful text mining tools can be to an insurance company and elucidate an easy and simple way to extract important information from unstructured or semi-structured documents.

Περιεχόμενα

ΚΕΦΑΛΑΙΟ 1

Το Text Mining και η ιστορία του

1.1. Εισαγωγή

Την εποχή που διανύουμε η πληροφορία έχει σημαντική θέση στον επιχειματικό κόσμο και όχι μόνο. Όλο και περισσότερα δεδομένα αποθηκεύονται καθημερινά σε βάσεις δεδομένων με αποτέλεσμα αυτές οι βάσεις να γιγαντώνονται από στοιχεία τα οποία στην πλειοψηφία τους δεν χρησιμοποιούνται πουθενά. Σύμφωνα με εκτιμήσεις, το 80% των διαθέσιμων πληροφοριών εμφανίζονται σε μορφή ελεύθερου κειμένου γεγονός το οποίο κάνει την ανάλυσή τους αδύνατη, Feldman (2003). Έτσι λοιπόν παρουσιάσθηκε η ανάγκη αυτά τα κείμενα να αναλυθούν, να διερευνηθεί αν μπορούν να προσφέρουν κάποια χρήσιμη πληροφορία και σε ποιους τομείς. Γι' αυτούς τους λόγους τα τελευταία είκοσι χρόνια έχει αναπτυχθεί μία καινούργια επιστήμη το Text Mining.

Το Text Mining μπορεί να ορισθεί ως η ανάλυση των ημι-δομημένων ή αδόμητων δεδομένων κειμένων. Ο στόχος είναι να μετατρέψει τις πληροφορίες κειμένου σε αριθμούς, ώστε να μπορούν να εφαρμοστούν αλγόριθμοι εξόρυξης δεδομένων και να προκύψουν χρήσιμα και ενδιαφέροντα συμπεράσματα. Τα βασικά βήματα του Text Mining σύμφωνα με τον Lachenbruch (2006) είναι τα εξής:

1. Ορισμός του σκοπού της μελέτης

2. Καθορισμός της διαθεσιμότητας των δεδομένων: αυτό συνεπάγεται την ανεύρεση και την καταγραφή του συνόλου των δεδομένων
3. Προετοιμασία των δεδομένων: αυτό σημαίνει κωδικοποίηση των δεδομένων ώστε να γίνεται πιο εύκολα η ανάλυση, εξαγωγή χαρακτηριστικών όπως σημεία στίξης και λέξεων που δεν έχουν καμία χρησιμότητα και μείωση των στοιχείων του κειμένου τα οποία δεν έχουν ιδιαίτερη στατιστική χρησιμότητα.
4. Ανάπτυξη και αξιολόγηση των μοντέλων: δηλαδή να αξιολογήσουμε ποιο μοντέλο ταιριάζει καλύτερα στο σκοπό της μελέτης μας προκειμένου να το χρησιμοποιήσουμε
5. Συμπεράσματα της όλης διαδικασίας.

Η βασική ιδέα του Text Mining είναι να πάρει επιθυμητές πληροφορίες από ‘’βουνά’’ γραπτών δεδομένων χωρίς να χρειάζεται ο χρήστης να τα διαβάσει όλα. Σήμερα για αυτή την καινούργια διαδικασία έχει αυξηθεί το ενδιαφέρον κυρίως λόγω του Internet στο οποίο υπάρχουν πολλά άρθρα και αναφορές. Μπορεί να θεωρηθεί ως μία μη παραδοσιακή μέθοδος όπου επιχειρούμε να χειριστούμε ολιστικά μεγάλες συλλογές κειμένων αποφεύγοντας την μεροληψία από ανθρώπινες παρεμβάσεις και αφήνοντας τα κείμενα να «μιλούν» από μόνα τους. Αποτελεί ουσιαστικά μία υπόσχεση για την βελτίωση της τεχνολογίας των πληροφοριών. Το Text mining εφαρμόζεται σε κείμενα τα οποία μπορεί να είναι αδόμητα, άμορφα και πολλές φορές δυσνόητα, ωστόσο οι πληροφορίες που περιλαμβάνουν κρίνονται σε πολλές περιπτώσεις εξαιρετικά σημαντικές. Ως εκ τούτου το κίνητρο για την προσπάθεια εξόρυξης σημαντικών αποτελεσμάτων από πλήθος γραπτών είναι τεράστιο. Ο στόχος είναι κυρίως να μειωθεί η προσπάθεια που χρειάζεται ένας χρήστης να αποκτήσει χρήσιμες πληροφορίες από μεγάλο όγκο μηχανογραφημένων κειμένων, Humphreys, Demetriou, & Gaizauskas, (2000).

1.2 Ιστορία του TEXT MINING

Η πρώτη αναφορά σε μία διαδικασία εξόρυξης πληροφοριών από κείμενα έγινε σε ένα άρθρο εφημερίδας από την H.P Luhn το 1958. Το συγκεκριμένο άρθρο ανέφερε πως, πριν από την έναρξη της εν λόγω διαδικασίας, είναι σημαντικό να προσδιορισθεί ο λόγος για τον οποίο πραγματοποιείται. Στην συνέχεια σημείωνε ότι για να αναλυθεί ένα κείμενο μέσω ηλεκτρονικού υπολογιστή πρέπει πρώτα να

προσδιορισθούν οι λέξεις που χαρακτηρίζουν καλύτερα τον σκοπό της ανάλυσης, Tefko Saracevic (2001) .Συνέχεια σε αυτό έδωσε ο Lauren B. Doyle το 1961 ο οποίος ασχολήθηκε με το πνεύμα του Text Mining και άλλων σχετικών μεθόδων και ισχυριζόταν ότι οι παράγοντες που επηρεάζουν περισσότερο το Text Mining είναι η συχνότητα με την οποία εμφανίζονται οι λέξεις και ο τρόπος που αυτές είναι κατανεμημένες σε ένα κείμενο. Το Text Mining μπορεί να είναι καινούργιο αλλά το όνειρο του να βρεθεί η μέθοδος με την οποία οι υπολογιστές να είναι σε θέση να αποσπούν πληροφορίες από ‘’βουνά’’ γραπτών είναι πάρα πολύ παλιό ανέφερε χαρακτηριστικά.

Στα τέλη της δεκαετίας του 1980-1990 ο καθηγητής Marti A. Hearst σε ένα άρθρο με τίτλο Untangling Text Data Mining ανέφερε ότι για σχεδόν μία δεκαετία η κοινότητα της υπολογιστικής γλωσσολογίας είδε μεγάλες συλλογές κειμένων ως πηγή που πρέπει να αξιοποιηθεί και γι’ αυτό θα έπρεπε να παραχθούν αλγόριθμοι για την ανάλυσή τους οι οποίοι θα ανοίξουν πόρτες σε νέα συναρπαστικά αποτελέσματα.

Το 1988 ο Don R Swanson σε ένα άρθρο του αναφέρθηκε στην ύπαρξη επιστημονικής βιβλιοθήκης η οποία θα πρέπει να θεωρηθεί ως ένα φυσικό φαινόμενο που αξίζει της ‘’εξερεύνησης, της συσχέτισης και της σύνθεσης’’. Ο Don Swanson υποστηρίζει ότι είναι εύλογο να αναμένουμε νέες πληροφορίες από συλλογές κειμένων. Οι εμπειρογνώμονες μπορούν να διαβάσουν μόνο ένα μικρό υποσύνολο του τι δημοσιεύεται στο αντικείμενο τους και συχνά είναι αποκομμένοι από τις εξελίξεις σε συναφείς τομείς. Έτσι θα πρέπει να είναι δυνατό να βρουν χρήσιμες διασυνδέσεις μεταξύ των πληροφοριών που σχετίζονται με συναφή αντικείμενα. Απέδειξε ότι στο πλαίσιο της ιατρικής βιβλιογραφίας μπορούμε να οδηγηθούμε σε αίτια σπάνιων ασθενειών ορισμένες από τις οποίες έχουν αποδειχθεί και πειραματικά, Swanson & Smalheiser (1999).

Πιο συγκεκριμένα ανέπτυξε ένα λογισμικό που σήμερα ονομάζεται ARROWSMITH και είναι ελεύθερα διαθέσιμο στο internet (http://kiwi.uchicago.edu) το οποίο βοηθά με την εξερεύνηση κοινών λέξεων-κλειδιών και κοινών φράσεων και αναφορών, να βρεθουν συσχετίσεις μεταξύ των κειμένων και των περιεχομένων τους. Για παράδειγμα αν έχουμε ένα κείμενο το οποίο υποστηρίζει τη φυσική σχέση ενός αντικειμένου Α με ένα Β, και έχουμε και ένα άλλο το οποίο υποστηρίζει τη σχέση του Β με ένα Γ, τότε μπορούμε να οδηγηθούμε στο συμπέρασμα ότι υπάρχει σχέση και μεταξύ του Α με το Γ. Ο Swanson έχει ανακαλύψει με αυτήν την μέθοδο δύο τουλάχιστον σημαντικές βιοϊατρικές σχέσεις: του ιχθυελαίου με το σύνδρομο του

Raynaud και του μαγνησίου με τις ημικρανίες, Lindsay & Gordon (1999).

Οι χρήσεις βέβαια του Text Mining δεν περιορίζονται μόνο στην βιοϊατρική αλλά και σε πολλούς και διαφορετικούς τομείς, για παράδειγμα στο φιλτράρισμα ανεπιθύμητων μηνυμάτων όπως συμβαίνει σε διάφορα email, στη δημιουργία προτάσεων ή συστάσεων για πανομοιότυπα προϊόντα όπως στο Amazon και στο Ebay, στην παρακολούθηση της κοινής γνώμης με εφαρμογή σε blogs ή σε ιστοσελίδες αναθεώρησης, στη μέτρηση προτιμήσεων των πελατών με την ανάλυση συνεντεύξεων όπως συμβαίνει σε τμήματα marketing σε μεγάλες επιχειρήσεις και στην καταπολέμηση της παρενόχλησης ή του εγκλήματος στον κυβερνοχώρο με την ανάλυση συνομιλιών.

Στην συγκεκριμένη διπλωματική εργασία θα ασχοληθούμε αποκλειστικά με τα οφέλη που μπορεί να προσφέρει το Text mining στον ασφαλιστικό χώρο.

ΚΕΦΑΛΑΙΟ 2

Απάτη στον ασφαλιστικό χώρο

2.1. Εισαγωγή

Οι δηλώσεις των ατυχημάτων και το κόστος που μπορούν να προκαλέσουν είναι ένας σημαντικός τομέας για τις ασφαλιστικές εταιρείες. Ο λόγος είναι ότι η προληπτική διαχείριση μπορεί να μειώσει σημαντικά τις ζημιές που μπορούν να προκληθούν. Το ζήτημα είναι ιδιαίτερα σημαντικό όταν θεωρείται ότι ένα μεγάλο μέρος των συνολικών απαιτήσεων προέρχεται από ένα μικρό αριθμό υποθέσεων.

Από έρευνες που έχουν γίνει στις Ηνωμένες Πολιτείες Αμερικής, πάνω από 30 δισεκατομμύρια δολάρια χάνονται από τις ασφαλιστικές λόγω ασφαλιστικής απάτης, Holm (2011). Η National Insurance Crime Bureau (NICB) θεωρεί την ασφαλιστική απάτη ως το δεύτερο πιο δαπανηρό υπαλληλικό έγκλημα στην Αμερική μετά την φοροδιαφυγή. Τα αποτελέσματα αυτής δεν γίνονται αισθητά μόνο από ιδιώτες ή εταιρείες που βλέπουν τα ασφάλιστρα τους να αυξάνονται αλλά και από τις ίδιες τις ασφαλιστικές οι οποίες συνεχώς αντιμετωπίζουν μεγαλύτερη πίεση προκειμένου να μειώσουν τα συνολικά τους έξοδα.

Η ολοένα και πιο υψηλή επικράτηση της απάτης είναι ένας από τους κύριους παράγοντες της αύξησης του combined ratio σήμερα. Ο combined ratio είναι ένας δείκτης ή με άλλα λόγια ένας μετρητής ο οποίος δείχνει πόσο καλά αποδίδει η ασφαλιστική εταιρεία στην καθημερινότητα της. Είναι ένας συνδυασμένος δείκτης μεταξύ του claims ratio (αξιώσεις που οφείλονται ως ποσοστό των ασφαλίστρων) και του expense ratio (λειτουργικό κόστος ως ποσοστό των εσόδων που προκύπτουν από τα ασφάλιστρα). Αν το combined ratio είναι κάτω του 100% τότε η εταιρεία κινείται κερδοσκοπικά, από την άλλη αν ξεπερνά το 100% σημαίνει ότι πληρώνει περισσότερα χρήματα στις αξιώσεις από όσα δέχεται από τα ασφάλιστρα.

Τα προηγούμενα χρόνια τα ποσοστά απάτης ήταν μονοψήφια και χαρακτηριζόντουσαν ως κόστος της επαγγελματικής δραστηριότητας. Τα τελευταία

όμως χρόνια τα ποσοστά έχουν φτάσει σε διψήφιους αριθμούς με αποτέλεσμα η ανάγκη για νέες μεθόδους προσέγγισης και μεωσή τους να κρίνεται επιτακτική.

Η απάτη είναι δυναμική δηλαδή, όταν το βιομηχανικό κύκλωμα ανακαλύπτει ένα συγκεκριμένο τύπο απάτης τότε βάζει εμπόδια στο να επαναληφθεί. Το κακό είναι όμως ότι ένα άλλο είδος απάτης εμφανίζεται στη θέση του και η διαδικασία ξεκινάει από την αρχή. Παράλληλα οι οργανωτικοί, νομικοί και εμπορικοί περιορισμοί υπό τους οποίους η ασφαλιστική βιομηχανία λειτουργεί, συχνά έχει αρνητικό αντίκτυπο στην επιτυχία των ήδη υπαρχόντων μεθόδων πρόληψης, ανίχνευσης και διερεύνησης της απάτης γι' αυτό το λόγο οι ασφαλιστικές αναζητούν συνεχώς νέους και πιο αποτελεσματικούς τρόπους αντιμετώπισης αυτού του φαινομένου.

2.2. Είδη απάτης και απατεώνων

Οι άνθρωποι διαπράττουν ασφαλιστικές απάτες με πολλούς τρόπους. Οι Gill, Woolley and Gill (1994) ορίζουν ως απάτη την γνώση του τρόπου να δημιουργήσεις μία εικονική δήλωση, το «φούσκωμα» ή η πρόσθεση επιπλέον στοιχείων σε μια δήλωση με ανέντιμο τρόπο προκειμένου να κερδίσουν περισσότερα χρήματα από όσα δικαιούνται από την ασφαλιστική.

Το Crime and Fraud Prevention Bureau στην ετήσια έκθεση του για το 2000 αναφέρει τέσσερις βασικούς τύπους απάτης στην ασφάλεια οχημάτων και σε συναφείς ασφαλίσεις. Η πρώτη είναι οι εντελώς ψευδείς ισχυρισμοί (12%), η δεύτερη είναι η σκόπιμη παραποίηση των συνθηκών της δήλωσης (39%), η τρίτη είναι η φουσκωμένη αξία απώλειας (30%), η τέταρτη είναι η υποστήριξή του από πολλαπλές ασφαλιστικές (3%) και το υπόλοιπο 14% να οφείλεται σε άλλου είδους απάτες.

Έκτος από τα είδη απάτης υπάρχουν και διάφορα είδη απατεώνων τα οποία έχουν ταξινομηθεί από τον Clarke (1989). Σύμφωνα λοιπόν με τον τελευταίο τα είδη είναι ο καιροσκόπος, ο ερασιτέχνης και ο επαγγελματίας. Ο καιροσκόπος εκμεταλλεύεται ένα πραγματικό γεγονός για την διάπραξη απάτης, για παράδειγμα υποστηρίζει ότι σε μία ληστεία που πραγματοποιήθηκε εις βάρος του απολέσθηκαν πολύ περισσότερα αντικείμενα από αυτά που στην πραγματικότητα κλάπηκαν. Ο ερασιτέχνης μπορεί να ξεκινήσει από την ευκαιριακή απάτη και στην συνέχεια να προχωρήσει ένα βήμα παραπέρα για παράδειγμα υποβάλλοντας αίτημα για

αντικείμενα που έχουν κλαπεί σε μία ληστεία που δεν συνέβη ποτέ. Τέλος ο επαγγελματίας επιδίδεται σε συστηματικές απάτες τόσο μεμονωμένα όσο και σε οργανωμένα δίκτυα.

Η έρευνα έχει επίσης επικεντρωθεί στην προσπάθεια να γίνει κατανοητό γιατί οι άνθρωποι διαπράττουν απάτη. Προσωπικές περιστάσεις και η δυσαρέσκεια προς τις ασφαλιστικές εταιρίες φαίνεται να επηρεάζουν σημαντικά την προθυμία για απάτη, Gill (1994). Η προσπάθεια να διακρίνουμε μία δήλωση-απάτη από μία γνήσια με βάση μόνο προσωπικά και κοινωνικά χαρακτηριστικά είναι αρκετά δύσκολη. Η έρευνα έχει δείξει ότι τα χαρακτηριστικά των απατεώνων σε σχέση με τους υπόλοιπους ασφαλισμένους δεν έχουν ιδιαίτερες διαφορές, Dodd (1998). Το γεγονός ότι οι επαγγελματίες απατεώνες δεν χρησιμοποιούν γνήσιες ταυτότητες ευνοεί ακόμα περισσότερο τα αποτελέσματα της παραπάνω έρευνας.

Ένας μεγάλος αριθμός νομικών, εμπορικών και οργανωτικών παραγόντων προσφέρουν ποικίλα κίνητρα απάτης. Μία νομικά δεσμευτική συμφωνία μεταξύ ασφαλιστή και ασφαλισμένου βασίζεται στην έννοια της καλής πίστης που σημαίνει ότι κάθε συμβαλλόμενο μέρος έχει την υποχρέωση να αποκαλύψει κάθε πληροφορία που θα μπορούσε να επηρεάσει την σύμβαση μεταξύ τους ακόμα και αν δεν έχει ζητηθεί ρητά. Αυτή η συμφωνία βασίζεται στην εμπιστοσύνη και στην ειλικρίνεια εκ μέρους και των δύο πλευρών.

Ο ανταγωνιστικός χαρακτήρας του ασφαλιστικού κλάδου δίνει την δυνατότητα σε πελάτες να αποκτήσουν πολιτική απάτης καθώς υπάρχει μικρή ως ελάχιστη πιθανότητα επικύρωσης των πληροφοριών που υποβάλλονται στις εκάστοτε ασφαλιστικές εταιρίες προκειμένου να ελεχθεί η καλή πίστη των δύο πλευρών.

Παράλληλα η αύξηση ευαισθητοποίησης των καταναλωτών μέσω της δημοσιοποίησης των καταναλωτικών προγραμμάτων σημαίνει ότι ένας κακός χειρισμός κάποιας κατάστασης μπορεί να προκαλέσει εκατομμύρια ευρώ ζημιάς με αποτέλεσμα οι ασφαλιστικές να είναι πολύ προσεκτικές στο τρόπο με τον οποίο αντιμετωπίζουν πιθανά περιστατικά απάτης, Clarke (1989). Συγχρόνως η απάτη είναι δύσκολο να αποδειχθεί νόμιμα και ενδέχεται να προκαλέσει περαιτέρω απώλειες τόσο σε χρήμα όσο και σε χρόνο με τα συνεχόμενα δικαστήρια και τα δικηγορικά έξοδα με αποτέλεσμα οι ασφαλιστικές εταιρείες συχνά να διστάζουν την δίωξη των παραβατών και να προτιμούν να πληρώνουν τις αξιώσεις που απαιτούν. Τέλος υπάρχει και ο φόβος αποτυχημένης διεκδίκησης που έχει αντίκτυπο οικονομικό αλλά και στην φήμη της εταιρείας.

2.3. Προσεγγίσεις για τον εντοπισμό ενδεχόμενης απάτης.

Παρά τα προβλήματα που συνδέονται με την αντιμετώπιση της απάτης και των δόλιων απαιτήσεων το σίγουρο είναι ότι υπάρχουν τρόποι να ανιχνευτούν. Ανέκδοτες αναφορές από συζητήσεις με τους διαχειριστές της απάτης υποστηρίζουν ότι η ανίχνευση συμβαίνει με διάφορους τρόπους κυρίως εμπειρικούς: από την ανακάλυψη ανωμαλιών ή ασυμφωνιών στις πληροφορίες γύρω από το αίτημα (π.χ. όταν οι συνθήκες της απαίτησης δεν ταιριάζουν με την περιγραφή που δίνεται από τον αιτούντα), από επαναλαμβανόμενες συμπεριφορές, παραδείγματος χάριν, παρόμοιες απώλειες ή παρόμοιες δηλώσεις σε διαφορετικές καταστάσεις και γεγονότα, ή από την αναγνώριση διστακτικότητας στην γενικότερη συμπεριφορά αντισυμβαλλόμενου όπως είναι η αβεβαιότητα και η ανεπαρκής παροχή πληροφοριών γύρω από ένα συμβάν.

Οι ερευνητές που έχουν αναλάβει την προστασία της ασφαλιστικής από τον κίνδυνο της απάτης ειδοποιούνται για τις ύποπτες αξιώσεις και πραγματοποιούν περαιτέρω έρευνες για να βρουν τρόπο να τις αντικρούσουν. Η διαδικασία της έρευνας γενικά περιλαμβάνει τον εντοπισμό περισσότερων πληροφοριών είτε από τον αιτούντα είτε από τρίτες πηγές και τη δημιουργία ενός σαφούς απολογισμού των ανωμαλιών, των ανακολουθιών στην αξίωση σε συνδυασμό με πιθανά, οικονομικά κυρίως, κίνητρα του αιτούντος.

Η ευθύνη για την ανίχνευση δόλιων απαιτήσεων στις ασφαλιστικές επιχειρήσεις στηρίζεται σε μεγάλο βαθμό στο προσωπικό της «πρώτης γραμμής» του χειρισμού των διαδικασιών των δηλώσεων. Οι εργαζόμενοι αυτοί όμως είναι συχνά άπειροι και στερούνται κατάλληλης ή επαρκούς εκπαίδευση στην ανίχνευση της απάτης, Doig, Jones and Wait (1999). Σε πολλές περιπτώσεις δεν είναι καλά πληροφορημένοι για το πόσο σημαντικός είναι ο ρόλος και τι οφέλη μπορούν να προσφέρουν στην εταιρεία. Κατά συνέπεια το ποσοστό των δόλιων απαιτήσεων που ανιχνεύονται είναι γύρω στο 10%, γεγονός που υποδηλώνει ότι μεγάλος αριθμός περιπτώσεων απάτης παραμένουν απαρατήρητες.

Για να αυξηθούν οι πιθανότητες ανίχνευσης δόλιων απαιτήσεων, οι ασφαλιστικές έχουν προσλάβει χειριστές οι οποίοι χρησιμοποιούν κάποιους δείκτες

απάτης και με βάση τα στοιχεία του πελάτη και της δήλωσης ελέγχονται συνεχώς οι εισερχόμενες αξιώσεις, Doig (1999). Αναφορές δύο εταιρειών που χρησιμοποιούν τέτοιες μεθόδους προσέγγισης υποστηρίζουν ότι ένας από τους δείκτες απάτης που χρησιμοποιούν βασίζεται στην εμπειρία της επιχείρησης από ενδείξεις απάτης. Κάθε εταιρεία λοιπόν χρησιμοποιεί δικούς της δείκτες από εμπειρίες που έχει από διαφορετικά συμβάντα απάτης που έχει αντιμετωπίσει. Το πιο σωστό βέβαια θα ήταν να υπάρχουν κατάλογοι με δείκτες από όλες τις ασφαλιστικές. Το εμπορικό όμως απόρρητο εμποδίζει την δημοσίευση ενός τέτοιου καταλόγου. Ακόμα πιο σημαντικό είναι ότι οι εταιρείες δεν θέλουν το κοινό, συμπεριλαμβανομένου και των απατεώνων, να έχει πρόσβαση σε τέτοιου είδους πληροφορίες διότι και οι ήδη υπάρχοντες δείκτες θα χάσουν την όποια προβλεπτική ικανότητα έχουν.

Ένας διμοφιλής δείκτης απάτης είναι ο δείκτης επαλήθευσης στοιχείων και γεγονότων. Ο έλεγχος της ημερομηνίας κυκλοφορίας του οχήματος αποτελεί έναν τέτοιον δείκτη αφού όσο πιο σύγχρονο είναι ένα αυτοκίνητο τόσο πιο δύσκολη είναι η διάρρηξη του αφού τα σύγχρονα οχήματα διαθέτουν υψηλής ποιότητας κλειδαριές. Άρα οι συνθήκες κλοπής θα πρέπει να δικαιολογούν την χρονολογία του.

Υπάρχουν και δείκτες οι οποίοι εξετάζουν την συμπεριφορά του ασφαλισμένου. Αν είναι νευρική και οι πληροφορίες οι οποίες αναφέρει στον ασφαλιστή γύρω από ένα συμβάν είναι συγκεχυμένες τότε η πιθανότητα απάτης είναι ιδιαίτερα αυξημένη.

2.4. Ελαττώματα των μέχρι σήμερα προσεγγίσεων

Οι ασφαλιστικές εταιρείες για να χρησιμοποιήσουν ένα δείκτη, θα πρέπει να αποφασίσουν ποιο θα είναι το όριο με βάση το οποίο θα προβαίνουν σε περαιτέρω εξερεύνηση της υπόθεσης. Πολλοί από τους δείκτες που χρησιμοποιούνται δεν έχουν πολύ καλή προγνωστική ικανότητα για τον εντοπισμό της απάτης. Παραδείγματος χάριν, η συμπεριφορά του ασφαλισμένου που έγινε θύμα κλοπής δεν είναι απαραίτητα νευρική λόγω δόλιας συμπεριφοράς αλλά μπορεί να οφείλεται στο σοκ που δέχτηκε από αυτό το γεγονός. Ένας πρόσθετος παράγοντας είναι ότι δεν έχει μπει ξανά σε μια τέτοια διαδικασία και δεν ξέρει τους τρόπους με τους οποίους θα λάβει την αποζημίωση τους και πόση θα είναι αυτή.

Συνεπώς οι κατάλογοι των δεικτών ίσως προσφέρουν μία ένδειξη αλλά σε γενικές γραμμές αποτυγχάνουν να αντανακλούν την δυναμική του χαρακτήρα της

απάτης. Συγχρόνως αυτοί οι στατικοί δείκτες καθιστούν δύσκολη τη διαπίστωση νέων παραλλαγών απάτης καθώς συνήθως βασίζονται στην εμπειρία των στελεχών της εταιρείας και αν οι τελευταίοι δεν τις έχουν συναντήσει στο παρελθόν δεν μπορούν και να τις προβλέψουν.

Εκτός των όσων προείπαμε, οι ασφαλιστικές εταιρείες έχουν λάβει περαιτέρω προληπτικά μέτρα για την βελτίωση της ανίχνευσης της απάτης κατά την διαδικασία εξέτασης των δηλώσεων. Έχουν δημιουργήσει ξεχωριστές βάσεις δεδομένων για κάθε κατηγορία περιστατικών προκειμένου να βοηθηθεί ο εντοπισμός των ανώμαλων πληροφοριών. Αυτές οι βάσεις μπορεί να είναι δηλώσεις για συνάλλαγμα, ή δηλώσεις για ασφαλίσεις αυτοκινήτων ή για κλοπή ή για ασφάλιση πιστώσεων. Ο στόχος αυτών των βάσεων είναι τριπλός. Πρώτον παρέχει έναν τρόπο για την επαλήθευση των πληροφοριών που παρέχονται από τους ενάγοντες. Δεύτερον επιτρέπουν στις επιχειρήσεις να αξιολογούν κατά πόσον οι ενάγοντες έχουν παρόμοιες ιστορίες ύποπτων συμπεριφορών ή απαιτήσεων στο παρελθόν και τέλος αποτελούν αποθετήρια για το ιστορικό των δηλώσεων με αποτέλεσμα την ανταλλαγή χρήσιμων πληροφοριών μεταξύ των τμημάτων της εταιρείας ανά πάσα στιγμή.

Όπως στην περίπτωση δεικτών έτσι και στα συστήματα των βάσεων δεδομένων παρουσιάζονται αρκετά προβλήματα ως προς την ανίχνευση της απάτης. Συγκεκριμένα τα δεδομένα για έναν πελάτη μπορεί να έχουν καταχωρηθεί πριν από αρκετό διάστημα με αποτέλεσμα πολλά από τα στοιχεία πιθανόν θα έχουν μεταβληθεί. Παράλληλα η εισαγωγή δεδομένων συχνά γίνεται μέσω τηλεφώνων ή βάσει των πληροφοριών που παρέχονται από τον πελάτη. Σε πολλές περιπτώσεις λοιπόν μπορεί να υπάρχουν ορθογραφικά λάθη, να λείπουν κάποια στοιχεία χρήσιμα για το νόημα μιας φράσης, κάποια δεδομένα να έχουν περαστεί περισσότερο από μία φορά και να υπάρχουν πολλές παλιές πληροφορίες που κάνουν τον όγκο τους τεράστιο και δύσκολο σε έναν αναγνώστη να το επεξεργαστεί. Κατά συνέπεια η αναζήτηση συστημάτων βάσεων δεδομένων με τις παραδοσιακές μεθόδους μπορεί να οδηγήσει σε προβλήματα τόσο με την αποτυχία κάποιας συγκεκριμένης αναζήτησης όσο και με την παραγωγή λανθασμένων εντυπώσεων λόγω κάποιων φράσεων που δεν έχουν νόημα.

2.5. Σύγχρονες μέθοδοι αντιμετώπισης

Για τον εντοπισμό της απάτης έχουν αναπτυχθεί διάφορες σύγχρονες τεχνολογίες. Κάποιες από αυτές έχουν βασιστεί στις προσεγγίσεις του προφίλ του υπόπτου και είναι εμπνευσμένες από εγκληματολογικές ιατροδικαστικές έρευνες όπως είναι ανάλυση τάσεων φωνής, Hovarth (1982).

Πρόσφατες αναφορές ασφαλιστικών εταιρειών δείχνουν ότι πολλές εταιρείες χρησιμοποιούν τέτοιου είδους τεχνολογίες στην προσπάθεια αντιμετώπισης δόλιων συμπεριφορών, Steed (2003). Οι προσεγγίσεις αυτές προσδιορίζουν πελάτες που εμφανίζουν παράξενα χαρακτηριστικά στη συμπεριφορά τους. Υπάρχουν και άλλες διαδικασίες που συνδυάζουν τους δείκτες απάτης με τις προσεγγίσεις προφίλ.

Αυτές οι τεχνολογικές προσεγγίσεις δείχνουν φιλόδοξες αλλά σε μεγάλο βαθμό η αποτελεσματικότητά τους δεν έχει τεκμηριωθεί επαρκώς. Μία ανησυχία είναι ότι οι συγκεκριμένες προσπάθειες ανίχνευσης φαίνεται να έχουν αναπτυχθεί χωρίς την πλήρη κατανόηση των χρηστών τους δηλαδή τον χειρισμό των δηλώσεων και των ερευνών του προσωπικού στον ασφαλιστικό κλάδο. Είναι μέθοδοι που δεν έχουν ως στόχο να επεκτείνουν τις ήδη υπάρχουσες ασφαλιστικές γνώσεις γύρω από την απάτη αλλά χρησιμοποιούν στοιχεία και δεδομένα από εγκληματικές συμπεριφορές από άλλους κλάδους.

Είναι πλέον ευρέως αναγνωρισμένο και σε άλλους τομείς της ανθρώπινης αλληλεπίδρασης με την τεχνολογία των υπολογιστών ότι είναι απαραίτητη η λεπτομερής κατανόηση των αναγκών, των ικανοτήτων, των αδυναμιών και της εμπειρίας που έχουν οι χρήστες και θα πρέπει να αποτελεί ουσιαστικό μέρος του αρχικού σχεδιασμού, Schneiderman (1998).

ΚΕΦΑΛΑΙΟ 3

Η διαδικασία του Text Mining

3.1. Εισαγωγή

Όπως ήδη αναφέρθηκε το Text Mining είναι μία ιδιαίτερα τεχνολογία η οποία τα τελευταία χρόνια αναπτύσσεται με γρήγορους ρυθμούς.

Πολλές ασφαλιστικές κυρίως του εξωτερικού έχουν επενδύσει πάνω σε αυτήν την καινούργια τεχνολογία. Οι εταιρείες χρησιμοποιούν το Text Mining όχι μόνο για την ανίχνευση της απάτης που αποτελεί βέβαια τον κύριο στόχο, αλλά και για βασικές διαδικασίες λήψης αποφάσεων, ανάλυσης αγοράς, ομαδοποίησης δεδομένων, τάσης των αξιώσεων, τιμολόγηση των προϊόντων και ανάλυση των τιμών.

Κατά την ανάλυση μεγάλων όγκων δεδομένων, η ποιότητα και η κάλυψη όλων των δεδομένων είναι δύο αντίθετοι πόλοι. Με περισσότερη κάλυψη η ποιότητα των δεδομένων δεν υφίσταται. Δεν είναι πρακτικό ή τουλάχιστον δεν γινόταν μέχρι σήμερα, οι πληροφορίες να είναι αποθηκευμένες σε δομημένα δεδομένα. Από δημοσιευμένα αποτελέσματα ο δείκτης όγκου δεδομένων είναι 80:20 υπέρ των αδόμητων. Αυτό είχε ως συνέπεια όλες οι παραπάνω διαδικασίες που αναφέραμε για τον εντοπισμό της απάτης να βασίζονται ουσιαστικά στο 20% των συνολικών δεδομένων γεγονός που μας δείχνει πόσο ελλιπείς ήταν οι έρευνες.

Για αρκετό καιρό υπήρξε αναγνώριση ότι τα δεδομένα σχετικά με συμβάντα περιέχουν πληροφορίες που επιτρέπουν μια προληπτική προσέγγιση διαχείρισης κινδύνου (risk management), Feyer and Williamson (1998). Το μεγάλο μέγεθος των ασφαλιστικών δεδομένων κάνει το Text mining ιδιαίτερα ελκυστικό εργαλείο για την ανάλυση τους σε σχέση με άλλες παραδοσιακές μεθόδους ανάλυσης, Kolyshkina, Steinberg (2003). Η αξιοποίηση της αξίας των πληροφοριών που υπάρχουν σε αυτά τα δεδομένα κάνει το ενδιαφέρον των ασφαλιστικών για τις νέες εφαρμογές εξόρυξης πληροφοριών από κείμενα να αυξάνεται συνεχώς Feyer, Stout (2001). Πριν φτάσουμε όμως στην ανάλυση, την αξιοποίηση και στο τέλος στην πρόβλεψη κάποιας απάτης υπάρχουν σημαντικά βήματα σε αυτή την καινούργια τεχνολογική μέθοδο.

3.2. Κατανόηση της μεθόδου του Text Mining

Πριν την έναρξη του Text Mining είναι πρώτα απαραίτητο να κατανοήσουμε τα εννοιολογικά θεμέλιά του και μετά να κατανοήσουμε πώς να ξεκινήσουμε την διαδικασία αξιοποιώντας τη δύναμη των δεδομένων μας. Σε αυτό το κεφάλαιο λοιπόν θα παρουσιάσουμε τα θεωρητικά θεμέλια του Text mining και θα περιγράψουμε τα βασικά βήματα προεπεξεργασίας και προετοιμασίας των κειμένων για ανάλυση.

Η πλειοψηφία των δεδομένων που αντιμετωπίζουμε είναι αδόμητα κείμενα. 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 στην μία και στην άλλη περίπτωση η μέθοδος που ακολουθούμε είναι ίδια. Συγχρόνως, λόγω του μεγάλου όγκου των κειμένων, είναι απαραίτητο να στραφούμε σε αυτοματοποιημένα μέσα που θα μας βοηθήσουν να κατανοήσουμε και να αξιοποιήσουμε τα διαθέσιμα έγγραφα.

Ο σκοπός των αλγορίθμων εξόρυξης κειμένου είναι να οδηγήσει στην κατανόηση και στην επεξεργασία χωρίς να είναι απαραίτητο κάποιος να το διαβάσει. Ωστόσο ένας υπολογιστής μπορεί να εξετάσει τους χαρακτήρες των λέξεων και την θέση τοποθετούντα οι λέξεις σε μία πρόταση. Δεν μπορεί να γνωρίζει τις πληροφορίες που ένα κείμενο μεταδίδει αλλά μπορεί να κατανοήσει τη δομή και τη σύνταξή του. Με τον όρο σύνταξη εννοούμε τη δομή της γλώσσας και πως μεμονωμένες λέξεις συνδυάζονται για να κάνουν προτάσεις και παραγράφους. Αυτή αποτελεί μία οργανωμένη διαδικασία. Ειδικοί κανόνες γραμματικής και γλώσσας διέπουν τον τρόπο με τον οποίο χρησιμοποιείται το λεξιλόγιο. Αυτή η δομή είναι σχετικά εύκολη για έναν υπολογιστή να την επεξεργαστεί. Από μόνη της όμως η σύνταξη δεν επαρκεί για την πλήρη κατανόηση της σημασίας ενός κειμένου. Η σημασιολογία των μεμονωμένων λέξεων που χρησιμοποιούνται είναι εξίσου σημαντική. Κοινά ιδιώματα είναι χρήσιμα για να τονιστεί η διαφορά μεταξύ της σύνταξης και της σημασιολογίας. Ιδιώματα είναι λέξεις ή φράσεις με μεταφορική σημασία που είναι δηλαδή διαφορετική από την κυριολεκτική έννοια των λέξεων. Για

παράδειγμα «η Μαρία έχει πεταλούδες στο στομάχι πριν από κάθε παράσταση» είναι συντακτικά σωστή και έχει δύο σημασιολογικές ερμηνείες: η γραμματική που ερμηνεύει ότι η Μαρία πριν από την παράσταση τρώει πεταλούδες και η ιδιωματική ερμηνεία είναι ότι 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. Πολλές εργασίες Text Mining όπως ταξινόμηση εγγράφων και ανάκτηση πληροφοριών ασχολούνται 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. Για παράδειγμα η έννοια εκχύλιση είναι περίπου η αυτόματη αναγνώριση λέξεων και φράσεων που έχουν την ίδια έννοια. Και πάλι οι προσεγγίσεις της εξόρυξης δεδομένων πρέπει να βασίζεται στη σύνταξη για να καταλήξει στην σημασιολογική σχέση. Στους περισσότερους αλγορίθμους το κείμενο αντιπροσωπεύεται από στοιχεία που δείχνουν την εμφάνιση των λέξεων μέσα στο κείμενο. Αυτό οδηγεί σε ένα μεγάλο αριθμό διαστάσεων. Σε όλα αυτά υπάρχει και μία σιωπηρή παραδοχή η οποία είναι ότι η σειρά στο έγγραφο δεν έχει καμία σημασία αλλά μεγάλη σημασία έχει πόσο συχνά βρίσκονται κοντά δύο ή και παραπάνω λέξεις μέσα στο κείμενο. Αυτό μπορεί να φαίνεται σαν μία παράξενη υπόθεση καθώς το κείμενο για να γίνει κατανοητό πρέπει να διαβαστεί με μία συγκεκριμένη σειρά. Για τις περισσότερες εφαρμογές όμως αυτό δεν είναι πρόβλημα. 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 να βρίσκουν όλες τις λέξεις κλειδιά ενός κειμένου. Συχνά οι λέξεις κλειδιά από μόνες τους δεν διαφοροποιούν ένα έγγραφο ενώ ο συνδυασμός με τις δευτερεύοντες λέξεις το διαφοροποιεί.

3.3. Προεπεξεργασία Κειμένου

Αφού κατανοήσουμε τη λογική του Text Mining το επόμενο βήμα είναι η συλλογή και η προεπεξεργασία όλων των δεδομένων που θέλουμε να αναλύσουμε. Δηλαδή να μετατρέψουμε τα αδόμητα και ημιδομημένα κείμενα σε δομημένα έγγραφα ικανά για ανάλυση. Αυτή η διαδικασία είναι απαραίτητο να γίνεται πριν από οποιοδήποτε εφαρμογή εξόρυξης κειμένου.

Τα βασικά βήματα αυτής είναι τα εξής:

α. Επιλογή των εγγράφων

Υπάρχουν περιπτώσεις στις οποίες είναι εύκολο να συλλέξουμε και να διαλέξουμε ποια έγγραφα επιθυμούμε να αναλύσουμε. Σε πολλές περιπτώσεις, όπως για παράδειγμα στα δεδομένα που προέρχονται από ασφαλιστικές εταιρείες και είναι έγγραφα τα οποία συλλέγονται για πολύ μεγάλο χρονικό διάστημα, πρέπει να αποφασίσουμε αν θα σπάσουν σε ενότητες παραγράφους ή προτάσεις. Η επιλογή αυτή εξαρτάται κυρίως από τον λόγο που πραγματοποιούμε την εξόρυξη κειμένου. Αν ο σκοπός μας είναι η ομαδοποίηση ή η ταξινόμηση των εγγράφων συνήθως χρησιμοποιούμε όλα τα δεδομένα που έχουμε στην διάθεση μας. Από την άλλη αν ο λόγος του Text Mining είναι η πρόβλεψη, παραδείγματος χάριν η πιθανότητα απάτης κάποιας δήλωσης, τότε είναι απαραίτητο τα δεδομένα μας να χωριστούν και να επιλεχθούν τα πιο χρήσιμα για την εν λόγω διαδικασία.

β. 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. Συγχρόνως μπορεί να δημιουργήσουν 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.

γ. 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. Αυτό λοιπόν που

κάνει το Tokenize είναι να αναγνωρίζει τις λέξεις, σε συνδυασμό με κάποιο λεξικό, και να τις διαχωρίζει με κενά προκειμένου να αναλυθούν ξεχωριστά.

δ. Διαγραφή των stopwords

Στις εφαρμογές Text Mining είναι χρήσιμο να αφαιρέσουμε κάποιες λέξεις οι οποίες παρουσιάζονται πολύ συχνά με στόχο να εξοικονομήσουμε χώρο αποθήκευσης και να επιταχύνουμε τη διαδικασία επεξεργασίας του κειμένου. Οι λέξεις αυτές ονομάζονται "stopwords" και η διαδικασία αφαίρεσης ονομάζεται "stopping". Κάθε αλγόριθμος Text Mining περιλαμβάνει αυτή την διαδικασία. Η διαγραφή των stopwords γίνεται χωρίς την απώλεια σημαντικών πληροφοριών γιατί στα περισσότερα κείμενα αυτές οι λέξεις δεν έχουν καμία επίδραση στα τελικά αποτελέσματα.

ε. Stemming

Το επόμενο βήμα στην προεπεξεργασία κειμένου ονομάζεται stemming και σκοπό έχει λέξεις οι οποίες έχουν την ίδια βάση δηλαδή προέρχονται από την ίδια λέξη να γίνονται μία. Αυτή η διαδικασία περιλαμβάνει τον εντοπισμό και την αφαίρεση των προθεμάτων, των επιθεμάτων και του πληθυντικού πάλι σε συνδυασμό με λεξικό και μας βοηθάει στη συρρίκνωση του κειμένου, στην βελτίωση του αλγορίθμου και στη βελτιώση της ακρίβειας των αποτελεσμάτων μας.

στ. Διόρθωση της ορθογραφίας

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 και συνήθως γίνεται με την βοήθεια λεξικών τα οποία συνδυάζονται με τον αλγόριθμο. Δηλαδή 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.

Έχοντας ολοκληρώσει την περιγραφή της διαδικασία της προεπεξεργασίας είναι σημαντικό να παραθέσουμε σημαντικές τεχνικές που χρησιμοποιούνται στο Text Mining και μας βοηθούν στην πλήρη κατανόηση της εν λόγω διαδικασίας.

ΚΕΦΑΛΑΙΟ 4

Μέτρα απόστασης και ομοιότητας

4.1. Εισαγωγή

Κατ' αρχήν θα επικεντρωθούμε στην ομαδοποίηση (clustering) η οποία αποτελεί βασικό εργαλείο για την τμηματοποίηση (segmentation) των δεδομένων. Κάθε ομάδα οφείλει να είναι ομοιογενής με βάση κάποιο μέτρο και συγχρόνως κάθε ομάδα οφείλει να είναι διαφορετική από την άλλη με βάση τα χαρακτηριστικά της, Sharma (1995). Σε αυτήν την διαδικασία σκοπός μας είναι να διατηρήσουμε μαζί στοιχεία τα οποία έχουν όμοια χαρακτηριστικά και να χωρίσουμε αυτά που έχουν ανόμοια χαρακτηριστικά.

Το πρώτο πράγμα που πρέπει να κάνουμε πριν ξεκινήσουμε οποιαδήποτε ομαδοποίηση είναι να βρούμε με ποιο μέτρο θα γίνει αυτός ο διαχωρισμός ή αλλιώς με ποιον τρόπο θα γίνει η ένωση όμοιων στοιχείων. Αυτό το μέτρο μπορεί, ανάλογα με την μέθοδο που χρησιμοποιούμε να είναι διαφορετικό. Παρακάτω παρουσιάζονται τα βασικότερα μέτρα που χρησιμοποιούνται στο Text Mining.

4.2. Απόσταση - Distance

Το πρώτο μέτρο με το οποίο θα ασχοληθούμε ονομάζεται απόσταση και συμβολίζεται με d (distance). Σκοπό έχει να δείξει πόσο κοντά βρίσκονται δύο ή και περισσότερες λέξεις μεταξύ τους.

Η απόσταση πρέπει να ικανοποιεί τις επόμενες βασικές ιδιότητες:

1. Η απόσταση μεταξύ δύο οποιοδήποτε σημείων–λέξεων θα πρέπει να είναι πάντα μεγαλύτερη ή ίση του μηδενός δηλαδή $d(x,y) \geq 0$ για κάθε x, y.

2. Η απόσταση μπορεί να πάρει την τιμή 0 μόνο αν τα δύο σημεία-λέξεις είναι ίδια δηλαδή $d(x,y) = 0$ αν $x = y$.

3. Η απόσταση δύο σημείων-λέξεων θα πρέπει να είναι συμμετρική. Αυτό σημαίνει ότι η απόσταση ενός σημείου-λέξης x με ένα άλλο y είναι ίση με την απόσταση του y από το x δηλαδή $d(x,y) = d(y,x)$.

4. Η απόσταση θα πρέπει να ικανοποιεί την τριγωνική ανισότητα δηλαδή $d(x,z) \leq d(x,y) + d(y,z)$.

Δίνουμε στην συνέχεια τις πιο συνηθισμένες απόστάσεις που χρησιμοποιούνται στην πράξη.

i. Ευκλείδεια Απόσταση

Η πιο συχνή μορφή απόστασης ονομάζεται ευκλείδεια απόσταση και είναι γνωστή περισσότερο για την επίλυση γεωμετρικών προβλημάτων. Ικανοποιεί τους τέσσερις βασικούς κανόνες που προαναφέραμε και χρησιμοποιείται ευρέως στην ομαδοποίηση του Text Mining. Χρησιμοποιείται ιδιαίτερα και στην K-means μέθοδο με την οποία θα ασχοληθούμε παρακάτω.

Έστω ότι συμβολίζουμε με $D = \{ d_1, d_2, d_3, \dots, d_n\}$ τα διαφορετικά κείμενα που έχουμε να αναλύσουμε και με $T = \{ t_1, t_2, t_3, \dots, t_m\}$ τις διαφορετικές λέξεις που χρησιμοποιούνται στα κείμενα. Ένα κείμενο ουσιαστικά περιγράφεται από ένα διάνυσμα $\boldsymbol{t_r}$ διάστασης m.

Ιδιαίτερη σημασία στο Text Mining έχει η συχνότητα που εμφανίζεται η κάθε λέξη και γι' αυτό θα πρέπει να την συμπεριλάβουμε στο τρόπο που συμβολίζουμε ένα κείμενο. Γι' αυτό με $f(t,d)$ θα συμβολίζουμε τη συχνότητα της λέξης t στο κειμένου d.

Πιο αντιπροσωπευτικός δείκτης από την συχνότητα είναι η λεγόμενη συχνότητα όρου (term frequency) και η οποία συμβολίζεται με τον εξής τύπο:

$$tf(t,d) = \frac{f(t,d)}{maxf(t,k)}$$

όπου το $maxf(t,k)$ συμβολίζει τον μεγαλύτερο αριθμό που εμφανίζεται η λέξη σε οποιοδήποτε κείμενο (k).

Το διάνυσμα του κειμένου θα παίρνει την μορφή :

$$\boldsymbol{t_r} = [tf(t_1, d_r), tf(t_2, d_r), tf(t_3, d_r) \dots . tf(t_m, d_r)\ ]$$

Το αρνητικό όμως της συχνότητας όρου είναι ότι σημαντικές λέξεις είναι αυτές που εμφανίζονται πιο συχνά. Παραδείγματος χάριν το «και» μπορεί να εμφανίζεται πολλές φορές και αυτό μας δείχνει ότι θα είναι σημαντική λέξη για ένα κείμενο. Ωστόσο σύμφωνα με αυτά που είπαμε παραπάνω αυτό δεν ισχύει πάντοτε. Σε πολλές περιπτώσεις οι λεξεις με μικρή συχνότητα είναι πιο σημαντικές από αυτές με μεγάλη συχνότητα. Γι' αυτόν τον λόγο χρησιμοποιούμε τη λεγόμενη αντίστροφη συχνότητα κειμένου (Inverse Document Frequence) το οποίο ορίζεται ως:

$$IDf(t) = log\frac{N}{n(t)}$$

όπου με N συμβολίζουμε τον συνολικό αριθμό των κειμένων και με $n(t)$ συμβολίζουμε τον αριθμό των κειμένων στα οποία αναφέρεται η λέξη t.

Αυτό που μας παρέχει η αντίστροφη συχνότητα κειμένου είναι ότι δίνει μικρό «σκορ» στις λέξεις με μεγάλη συχνότητα και μεγάλο «σκορ» στις λέξεις με μικρή συχνότητα. Πολλαπλασιάζοντας λοιπόν τις δύο αυτές συχνότητες έχουμε μία πλήρη εικόνα της χρησιμότητας μιας λέξης σε ένα κείμενο. Κατά συνέπεια το διάνυσμα που περιγράφει καλύτερα ένα κείμενο είναι της μορφής:

$$\boldsymbol{t_r} = [tf(t_1, d_r)IDf(t_1),\ tf(t_2, d_r)IDf(t_2),\ tf(t_3, d_r)IDf(t_3)\ldots.tf(t_m, d_r)IDf(t_m)]$$

Σύμφωνα λοιπόν με τους συμβολισμούς που αναλύσαμε παραπάνω η ευκλείδεια απόσταση δύο κειμένων υπολογίζεται από τον τύπο:

$$D_E(\boldsymbol{t_a}, \boldsymbol{t_b}) = \sqrt{\sum_{i=1}^{m}[tf(t_i, d_a)IDf(t_i) - tf(t_i, d_b)IDf(t_i)]^2}$$

ii. Manhattan ή City Block απόσταση

Υπάρχουν και άλλου είδους αποστάσεις εκτός της Ευκλείδειας όπως η **απόσταση Manhattan** ή City Block η οποία μοιάζει πολύ στην ευκλείδεια με την διαφορά ότι αντί για τετραγωνικές αποκλίσεις χρησιμοποιούμε απόλυτες αποκλίσεις. Και αυτή ικανοποιεί τους τέσσερις κανόνες και η απόσταση δύο κειμένων υπολογίζεται από τον τύπο:

$$D_{Manhattan}(\boldsymbol{t_a}, \boldsymbol{t_b}) = \sum_{i=1}^{m}|tf(t_i, d_a)IDf(t_i) - tf(t_i, d_b)IDf(t_i)|$$

iii. Απόσταση Minkowski

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:

$$D_{Minkowski}(\boldsymbol{t_a}, \boldsymbol{t_b}) = \left(\sum_{i=1}^{m} |tf(t_i, d_a) IDf(t_i) - tf(t_i, d_b) IDf(t_i)\ |^{\frac{1}{\lambda}} \right)^{\lambda}$$

όπου $\lambda \geq 1$.

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

Υπάρχουν και άλλες μορφές αποστάσεων αλλά δεν χρησιμοποιούνται ιδιαίτερα στις εφαρμογές του Text Mining.

4.3. Άλλου είδους μέτρα

Γενικά όλα τα είδη αποστάσεων είναι γνωστά ως μέτρα απόστασης. Υπάρχουν και άλλου είδους μέτρα τα οποία είναι πιο γνωστά ως μέτρα ομοιότητας. Ως ομοιότητα ορίζουμε την ύπαρξη κοινών χαρακτηριστικών ή την μέχρι κάποιου βαθμού συμμετρία μεταξύ δύο ή περισσότερων οντοτήτων. Στην συνέχεια γίνεται σύντομη παρουσίαση κάποιων από τα σημαντικότερα μέτρα ομοιότητας για Text Mining.

i. Ομοιότητα Συνημιτόνου

Το πρώτο που θα αναφέρουμε είναι η **Ομοιότητα Συνημιτόνου** (Cosine Similarity) η οποία στην συγκεκριμένη περίπτωση έχει πολλά πλεονεκτήματα. Αυτό το κριτήριο δείχνει την συσχέτιση που έχουν δύο διανύσματα μεταξύ τους. Ποσοτικοποιείται ως το συνιμήτονο της γωνίας μεταξύ διανυσμάτων και επειδή στην περίπτωση του Text Mining τα κείμενα παρουσιάζονται ως διανύσματα αποτελεί ένα από τα πιο δημοφιλή κριτήρια ομαδοποίησης.

Ο τρόπος υπολογισμού της ομοιότητας συνιμητόνου δύο κειμένων είναι ο εξής:

$$SIM_C(t_a, t_b) = \frac{t_a \cdot t_b}{|t_a| \cdot |t_b|}$$

Η ομοιότητα συνιμητόνου είναι μη αρνητική τιμή και να λαμβάνει τιμές μεταξύ (0,1). Το θετικό αυτού του κριτηρίου είναι ότι είναι ανεξάρτητο του μεγέθους ενός κειμένου. Με άλλα λόγια 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. Τέλος θα πρέπει να αναφέρουμε ότι αν δύο κείμενα μοιάζουν μεταξύ τους τότε η ομοιότητα συνημιτόνου είναι κοντά στην μονάδα (1) ενώ αν είναι διαφορετικά τείνει στο 0.

ii. Συντελεστής Jaccard

Άλλο κριτήριο ομοιότητας είναι ο **συντελεστής Jaccard** ή αλλιώς **συντελεστής Tanimoto** ο οποίος μετράει την ομοιότητα ως τομή των αντικειμένων διαρούμενη με την ένωση τους.

Στην περίπτωση του Text Mining ο συντελεστής Jaccard συγκρίνει το άθροισμα των βαρών των κοινών λέξεων δύο κειμένων προς το άθροισμα των βαρών όλων λέξεων που παρουσιάζονται στα δύο κείμενα αλλά δεν είναι κοινά. Ο επίσημος ορισμός είναι ο εξής:

$$SIM_j(t_a, t_b) = \frac{t_a \cdot t_b}{|t_a|^2 + |t_b|^2 - 2t_a \cdot t_b}$$

Οι τιμές που λαμβάνει ο συντελεστής Jaccard κυμαίνονται μεταξύ του (0,1). Όπως και στην ομοιότητα συνημιτόνου αν δύο κείμενα μοιάζουν μεταξύ τους τότε ο συντελεστής Jaccard είναι κοντά στην μονάδα ενώ αν δεν έχουν κοινά στοιχεία τότε είναι κοντά στο μηδέν.

iii. Συντελεστής συσχέτισης Pearson

Άλλο ένα κριτήριο που δείχνει το βαθμό στον οποίο δύο διανύσματα έχουν ομοιότητες ή όχι είναι ο συντελεστής συσχέτισης Pearson. Υπάρχουν διάφορες μορφές που ορίζεται αυτός ο συντελεστής αλλά ο πιο συνηθισμένος στο Text Mining είναι ο εξής:

$$SIM_P(t_a, t_b) = \frac{\sum_{i=1}^{m} w_{t_i,a} \times w_{t_i,b} - TF_a \times TF_b}{\sqrt{[m\sum_{i=1}^{m} w_{t_i,a}^2 - TF_a^2] - [m\sum_{i=1}^{m} w_{t_i,b}^2 - TF_b^2]}}$$

Όπου $T = \{t_1, t_2, t_3, \ldots, t_m\}$ το σύνολο λέξεων, $w_{t_i,a} = tf(t_i, d_a)IDf(t_i)$, $w_{t_i,b} = tf(t_i, d_b)IDf(t_i)$
και

$$TF_a = \sum_{i=1}^{m} w_{t_i,a},$$
$$TF_b = \sum_{i=1}^{m} w_{t_i,b}.$$

Σε αντίθεση με τα δύο παραπάνω κριτήρια ομοιότητας ο συντελεστής συσχέτισης Pearson λαμβάνει τιμές μεταξύ (-1,+1) όπου 1 όταν τα διανύσματα είναι ίσα ενώ όταν η τιμή είναι κοντά στο μείον ένα (-1) υποδηλώνει ότι τα διανύσματα έχουν ελάχιστες ομοιότητες.

iv. Κατά μέσο όρο Kullback-Leibler Απόκλιση

Το τελευταίο και πιο πολύπλοκο ίσως κριτήριο ομοιότητας στο Text Mining είναι η κατά μέσο όρο απόκλιση Kullback-Leibler. Ένα κείμενο θεωρείται ως μία κατανομή πιθανοτήτων των λέξεων. Σε αυτήν την περίπτωση το κριτήριο ομοιότητας δεν είναι άλλο από την ομοιότητα που έχουν οι δύο κατανομές πιθανοτήτων. Έτσι έστω ότι έχουμε την κατανομή $P(x)$ και την κατανομή $Q(x)$ τότε η Kullback-Leibler απόκλιση ορίζεται ως:

$$D_{KL}(P(x)||Q(x)) = \sum_{x \in X} P(x) log(\frac{P(x)}{Q(x)})$$

Στην περίπτωση όμως του Text Mining αυτό το μέτρο ορίζεται ως:

$$D_{KL}(t_a \,||\, t_b) = \sum_{i=1}^{m} w_{t_i,a} \times \log(\frac{w_{t_i,a}}{w_{t_i,b}})$$

Επειδή στο συγκεκριμένο μέτρο δεν ισχύει το κριτήριο της συμμετρίας αφού $D_{KL}(P(x)||Q(x)) \neq D_{KL}(Q(x)||P(x))$, συνήθως χρησιμοποιούμε την κατά μέσο όρο Kullback-Leibler απόκλιση, η οποία ορίζεται ως εξής:

$$D_{AvgKL}(P(x)||Q(x)) = \pi_1 D_{KL}(P(x)||M(x)) + \pi_2 D_{KL}(Q(x)||M(x))$$

Όπου $\pi_1 = \frac{P(x)}{P(x)+Q(x)}$, $\pi_2 = \frac{Q(x)}{P(x)+Q(x)}$, $M(x) = \pi_1 P(x) + \pi_2 Q(x)$.

Στο Text Mining η κατά μέσο όρο Kullback-Leibler απόκλιση υπολογίζεται με τον εξής τρόπο.

$$D_{AvgKL}(\boldsymbol{t_a} \,||\, \boldsymbol{t_b}) = \sum_{i=1}^{m} (\pi_1 \times D(w_{t_i,a} || w_{t_i}) + \pi_2 \times D(w_{t_i b} || w_{t_i}))$$

Όπου $\pi_1 = \frac{w_{\boldsymbol{t},a}}{w_{\boldsymbol{t},a}+w_{\boldsymbol{t},b}}$, $\pi_2 = \frac{w_{\boldsymbol{t},b}}{w_{\boldsymbol{t},a}+w_{\boldsymbol{t},b}}$, $w_{\boldsymbol{t}} = \pi_1 \times w_{\boldsymbol{t},a} + \pi_2 \times w_{\boldsymbol{t},b}$.

Με αυτήν λοιπόν την στάθμιση μεταξύ των δύο διανυσμάτων επιτύχαμε η κατά μέσο όρο Kullback-Leibler απόκλιση να ικανοποιεί όλα τα αρχικά κριτήρια και να είναι ένα από τα πιο συνηθισμένα μέτρα ομοιότητας σε εφαρμογές Text Mining.

ΚΕΦΑΛΑΙΟ 5

Τεχνικές Ομαδοποίησης

5.1. Εισαγωγή

Αφού επιλέξουμε ποιο κριτήριο ομοιότητας ή απόστασης θα χρησιμοποιήσουμε το επόμενο βήμα είναι να επιλέξουμε ποια τεχνική τεχνική ομαδοποίησης θα χρησιμοποιήσουμε.

Οι ομαδοποιήσεις χωρίζονται κατ' αρχάς σε ιεραρχικές και μη ιεραρχικές ανάλογα με τον τρόπο που διαμορφώνουν τις ομάδες.

Οι ιεραρχικές μέθοδοι ομαδοποίησης χωρίζονται σε δύο βασικές κατηγορίες στις συσσωρευτικές (agglomerative methods) και στις διαιρετικές (divisive methods). Η ιεραρχική ομαδοποίηση χρειάζεται πολύ χώρο και χρόνο για την εφαρμογή της καθώς σε κάθε βήμα χρησιμοποιεί έναν πίνακα με τις αποστάσεις όλων των λέξεων από όλες τις υπόλοιπες.

Σε αντίθεση με την ιεραρχική ομαδοποίηση στην μη ιεραρχική ο αριθμός των ομάδων μπορεί να ορισθεί είτε εκ των προτέρων είτε να καθοριστεί σαν μέρος της διαδικασίας ομαδοποίησης. Στις μη ιεραρχικές δεν είναι απαραίτητο να καθοριστεί πίνακας αποστάσεων όλων των λέξεων από όλες τις υπόλοιπες, γι' αυτό δεν χρειάζεται πολύ μεγάλος χώρος αποθήκευσης και έχουν την ικανότητα να εφαρμόζονται σε πολύ μεγάλους όγκους δεδομένων.

5.2. Ιεραρχικές Μέθοδοι Ομαδοποίησης

Η γενική λογική των συσσωρευτικών μεθόδων είναι ότι ξεκινούν με n ομάδες όπου κάθε ομάδα αντιπροσωπεύει μία λέξη και με διαδοχικές συγχωνεύσεις καταλήγουν σε μία ομάδα. Στην αρχή υπολογίζονται όλες οι αποστάσεις των λέξεων $d(x_i, y_i)$και δημιουργείται ένας πίνακας αποστάσεων (distance ή dissimilarity matrix) $D = [d(x_i, y_i)]$ με n γραμμές και n στήλες ο οποίος περιλαμβάνει όλες τις

αποστάσεις. Στην συνέχεια εντοπίζονται οι δύο πρώτες πλησιέστερες λέξεις και συγχωνεύονται σε μία ομάδα άρα οι ομάδες μας γίνονται $n-1$. Συγχρόνως μειώνεται ο πίνακας D αφού διαγράφεται η απόσταση αυτών των δύο λέξεων. Η ίδια διαδικασία πραγματοποιείται για κάθε λέξη και για κάθε ομάδα που δημιουργείται με αποτέλεσμα στο τέλος να έχουμε μία ομάδα n λέξεων. Σε κάθε συγχώνευση αλλάζει και ο πίνακας αποστάσεων δηλαδή διαγράφονται οι γραμμές και οι στήλες των ομάδων ή λέξεων που συγχωνεύονται και προστίθενται οι αποστάσεις της νέας ομάδας με όλες τις άλλες ομάδες ή λέξεις.

Η συγχώνευση των λέξεων γίνεται με το κριτήριο της απόστασης που είδαμε προηγουμένως. Στην περίπτωση των ομάδων το κριτήριο παραμένει η απόσταση αλλά αλλάζει η μέθοδος.

Υπάρχουν διάφορες τεχνικές υπολογισμού αποστάσεων των ομάδων με σκοπό την συγχώνευση τους. Οι πιο γνωστές είναι, S. Sharma (1995):

i. Μέθοδος της απλής συνένωσης (Nearest neighbor ή Single linkage method)

Στην περίπτωση αυτή ψάχνουμε να βρούμε την μικρότερη απόσταση μεταξύ μίας λέξης μιας ομάδας με μία λέξη από μία άλλη ομάδα δηλαδή:

$$d_{\text{ομάδων}} = min(d(x_i, y_i))$$

όπου x_i είναι οι λέξεις της μιας ομάδας και y_i της άλλης.

Έστω ότι συγχωνεύθηκαν οι ομάδες *A,B* και πήραμε την ομάδα *Γ=(AB)*. Τότε οι νέες αποστάσεις του πίνακα αποστάσεων θα υπολογισθούν από την επόμενη σχέση όπου με Δ συμβολίζουμε μία ήδη υπάρχουσα ομάδα:

$$d_{\Gamma,\Delta} = min\{d(A,\Delta)d(B,\Delta)\} = \frac{1}{2}\ [d(A,\Delta) + d(B,\Delta)] - \frac{1}{2}\,|d(A,\Delta) + d(B,\Delta)|$$

Το μειονέκτημα αυτής της μεθόδου είναι ότι όταν δύο ομάδες με σημαντικές διαφορές έχουν δύο λέξεις σε κοντινή απόσταση μεταξύ τους υπάρχει περίπτωση να συγχωνευθούν με αποτέλεσμα οι ομάδες που δημιουργούνται να μην είναι συμπαγείς και να δημιουργούνται μερικές πολύ μεγάλες ομάδες και κάποιες πολύ μικρές.

ii) Μέθοδος της πλήρους συνένωσης (Complete Linkage Method ή furthest neighbor)

Σε αυτήν την μέθοδο χρησιμοποιούμε για κριτήριο την μεγαλύτερη απόσταση από μία λέξη μιας ομάδας με κάποια παρατήρηση στην άλλη ομάδα δηλαδή:

$$d_{ομάδων} = max(d(x_i, y_i))$$

όπου x_i είναι οι λέξεις της μιας ομάδας και y_i της άλλης.

Ο υπολογισμός του πίνακα αποστάσεων γίνεται ως εξής:

$$d_{\Gamma,\Delta} = max\{d(A,\Delta)d(B,\Delta)\} = \frac{1}{2}\ [d(A,\Delta) + d(B,\Delta)] - \frac{1}{2}\ |d(A,\Delta) + d(B,\Delta)|$$

Το μειονέκτημα αυτής της μέθοδου είναι ότι σε περίπτωση που δύο ομάδες έχουν αρκετά όμοια στοιχεία αλλά δύο λέξεις είναι πολύ μακριά μεταξύ τους, ενώ θα έπρεπε να συγχωνευθούν, στις περισσότερες περιπτώσεις παραμένουν δύο ξεχωριστές ομάδες.

iii) Μέθοδος σταθμισμένων μέσων (Weighted Average Linkage method)

Η τρίτη μέθοδος ονομάζεται μέθοδος σταθμισμένων μέσων και κριτήριο αποτελεί ο μέσος όρος των αποστάσεων όλων των λέξεων της μιας ομάδας με τα στοιχεία της άλλης ομάδας. Έτσι αν μία ομάδα περιλαμβάνει τις λέξεις 1,2 και η άλλη ομάδα τις λέξεις 3,4 τότε ο τύπος με τον οποίο βρίσκουμε την απόσταση μεταξύ των ομάδων είναι ο εξής:

$$d_{(12)(34)} = \frac{d_{13} + d_{14} + d_{23} + d_{24}}{4}$$

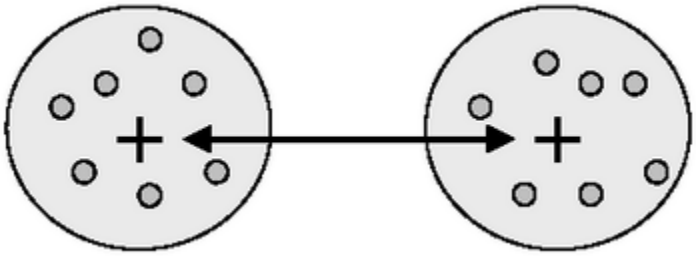
Υπάρχει και μία παραλλαγή αυτής της μεθόδου σύμφωνα με την οποία υπολογίζουμε την απόσταση όλων των λέξεων και των δύο ομάδων μεταξύ δηλαδή βρίσκουμε τον μέσο όρο των αποστάσεων όλων των λέξεων που ανήκουν σε ανά δύο

ομάδες. Η μέθοδος αυτή ονομάζεται Group Average Method και ο προηγούμενος τύπος αλλάζει και γίνεται:

$$d_{(12)(34)} = \frac{d_{12} + d_{13} + d_{14} + d_{23} + d_{24} + d_{34}}{6}$$

iv) Μέθοδος των κέντρων βάρους (Centroid method)

Σε αυτήν την μέθοδο υπολογίζουμε τα κέντρα της κάθε ομάδας και η απόσταση που θα κρίνει αν δύο ομάδες θα συγχωνευθούν ή όχι, είναι η απόσταση μεταξύ των κέντρων τους



Οι ομάδες με την μικρότερη απόσταση συγχωνεύονται.

v) Μέθοδος του Ward (Ward's method)

Η επόμενη μέθοδος είναι του Ward. Σε αυτή για κάθε λέξη μιας ομάδας υπολογίζουμε την απόστασή της από το κέντρο. Το κέντρο συμβολίζεται με $\bar{x}$ και με x_i η κάθε λέξη. Αμέσως μετά υπολογίζουμε για κάθε ομάδα ένα μέτρο συνεκτικότητας το οποίο ονομάζεται άθροισμα των τετραγωνικών αποκλίσεων (Error Sum of Squares) από τον τύπο:

$$ESS(\text{ομάδας}) = \sum [d(x_i, \bar{x}(\text{ομάδας}))]^2$$

Στην αρχή όλες οι ομάδες θα έχουν $ESS = 0$ αφού η κάθε λέξη αποτελεί μία ομάδα και άρα είναι και το κέντρο της. Στην συνέχεια πραγματοποιούνται όλες οι πιθανές συγχωνεύσεις και η ομάδα με την μικρότερη αύξηση του αθροίσματος των τετραγωνικών αποκλίσεων παραμένει ομάδα. Αυτή η διαδικασία θα γίνει ξανά και ξανά μέχρις ότου όλες οι λέξεις συγχωνευθούν σε μία ομάδα. Είναι μία μέθοδος που

σκοπό έχει την ελαχιστοποίηση της διακύμανσης της κάθε ομάδας και είναι αρκετά αποτελεσματική δημιουργώντας συμπαγείς ομάδες, Mardia (1979).

Όπως προείπαμε η άλλη μεγάλη κατηγορία ιεραρχικών μεθόδων είναι οι διαιρετικές. Έχουν την ακριβώς αντίθετη λογική από τις συσσωρευτικές καθώς η διαδικασία ξεκινάει με τη δημιουργία μίας ομάδος με όλα τα στοιχεία και την διαιρεί σε όλο και μικρότερες υποομάδες. Δηλαδή βρίσκουν υποομάδες που έχουν απομακρυσμένα στοιχεία σε σχέση με τις υπόλοιπες και τις διαχωρίζουν. Το κακό αυτής της μεθόδου είναι ότι χρειάζεται πολύ μεγάλο χώρο και χρόνο για την εκτέλεσή της και αυτός είναι ο λόγος που πολύ σπάνια θα συναντήσουμε εφαρμογές τέτοιου τύπου.

5.3. Μη ιεραρχική ομαδοποίηση

Εκτός των ιεραρχικών μεθόδων ομαδοποιήσεων υπάρχουν και οι μη ιεραρχικές. Η πρώτη μεγάλη διαφορά των δύο είναι ότι στις δεύτερες έχουμε προκαθορισμένο αριθμό υποομάδων που θέλουμε να δημιουργήσουμε.

Υπάρχουν δύο διαφορετικοί τρόποι που λειτουργούν οι μη ιεραρχικές ομαδοποιήσεις προκειμένου να δημιουργήσουν δεδομένο αριθμό ομάδων *k*. Ο πρώτος είναι να θεωρούν *k* συγκεκριμένες λέξεις οι οποίες ονομάζονται μητρικά σημεία (seed points) και με βάση αυτά γίνεται η ομαδοποίηση. Ο αρχικός προσδιορισμός των μητρικών σημείων μπορεί να πραγματοποιηθεί με ποικίλους τρόπους κατά την διάρκεια βέβαια της διαδικασίας τα μητρικά σημεία μπορεί να αλλάξουν πολλές φορές:

α. ο πρώτος και πιο εύκολος είναι να επιλέξουμε τις *k* πρώτες λέξεις οι οποίες θα αποτελέσουν αυτά τα σημεία.

β. αριθμούμε τις λέξεις από 1 μέχρι n (το πλήθος όλων των λέξεων) και διαλέγουμε τυχαία *k* λέξεις (μητρικά σημεία)

γ. δημιουργούμε τυχαίες *k* υποομάδες με όλες τις λέξεις και παίρνουμε ως μητρικά σημεία τα κέντρα βάρους τους

δ. μία άλλη πιο περίπλοκη μέθοδος την οποία έχει προτείνει ο Astrahan ξεκινάει υπολογίζοντας για κάθε λέξη πόσες λέξεις υπάρχουν γύρω τους ,με ακτίνα την οποία θα πρέπει από πριν να την καθορίσουμε, και αυτή θεωρείται ως πυκνότητα της κάθε λέξης. Στην συνέχεια βρίσκουμε την λέξη με την μεγαλύτερη πυκνότητα και την θεωρούμε ως το πρώτο μητρικό μας σημείο. Στην συνέχεια επιλέγουμε ως μητρικό σημείο την λέξη με την αμέσως

μικρότερη πυκνότητα με ένα επιπλέον κριτήριο το οποίο πάλι πρέπει να προσδιορίσουμε εμείς και είναι η απόσταση των μητρικών σημείων μεταξύ τους να μην πέφτει κάτω από ένα όριο. Αυτή η διαδικασία πραγματοποιείται μέχρις ότου φτάσουμε τα *k* μητρικά σημεία, Anderberg (1973).

Ο άλλος τρόπος είναι να ξεκινήσει η μέθοδος με *k* υποομάδες και στην συνέχεια να μετακινεί στοιχεία από την μία ομάδα στην άλλη μέχρις ότου οι τελικές ομάδες να κρίνονται συμπαγείς. Ομοίως ο προσδιορισμός των αρχικών υποομάδων ποικίλει, για παράδειγμα:

α. να επιλέγουμε τυχαία τις υποομάδες μας

β. να χρησιμοποιούμε μία ιεραρχική μέθοδο για τον προσδιορισμό τους

γ. να κάνουμε ένα αρχικό διαχωρισμό με εμπειρικά κριτήρια

Η πιο δημοφιλής μη ιεραρχική μέθοδος ομαδοποίησης είναι η μέθοδος *k* - means. Σε αυτήν την μέθοδο επιλέγουμε έναν από τους τρόπους προσδιορισμού των μητρικών σημείων και στην συνέχεια δημιουργούμε *k* ομάδες με κριτήριο η απόσταση της κάθε λέξης να είναι η μικρότερη δυνατή από τα μητρικά στοιχεία που δημιουργήσαμε. Στην συνέχεια υπολογίζουμε το κέντρο βάρους της κάθε ομάδας τα οποία θα αποτελέσουν τα νέα μητρικά σημεία και επαναλαμβάνουμε την ίδια διαδικασία για να δούμε αν κάποια λέξη θα μετακινηθεί σε άλλη ομάδα. Το τελευταίο βήμα θα επαναλαμβάνεται ξανά και ξανά μέχρις ότου καμία λέξη να μην χρειάζεται να μετακινηθεί.

Μία άλλη μέθοδος είναι η μέθοδος που προτάθηκε από τον Forgy στην οποία διαμερίζουμε στην αρχή όλες τις λέξεις σε *k* ομάδες και υπολογίζουμε τα κέντρα βάρους της κάθε ομάδας. Στην συνέχεια τοποθετούμε όλες τις λέξεις στην ομάδα που έχει την μικρότερη απόσταση λέξη-κέντρο βάρους και υπολογίζουμε ξανά τα κέντρα βάρους των καινούργιων ομάδων. Αν είναι τα ίδια με πριν τότε η διαδικασία σταματάει, αν είναι διαφορετικά επαναλαμβάνουμε την τοποθέτηση της κάθε λέξης στην κατάλληλη ομάδα και υπολογίζουμε ξανά τα κέντρα βάρους.

Η δυσκολία στην εφαρμογή όλων των μη ιεραρχικών ομαδοποιήσεων είναι κατ' αρχάς ο προσδιορισμός του κατάλληλου *k* προκειμένου οι ομάδες που θα δημιουργηθούν να είναι συμπαγής και στην συνέχεια ο προσδιορισμός των κατάλληλων αρχικών είτε υποομάδων είτε μητρικών στοιχείων γιατί σε περίπτωση λάθους δημιουργούνται λανθασμένες ομαδοποιήσεις.

ΚΕΦΑΛΑΙΟ 6

Τεχνικές Πρόβλεψης

6.1. Εισαγωγή

Στο προηγούμενο κεφάλαιο παρουσιάσαμε τις τεχνικές ομαδοποίησης που χρησιμοποιούμε προκειμένου να μπορέσουμε να κατανοήσουμε τι λέει ένα κείμενο να δημιουργήσουμε ομάδες με διαφορετικά νοήματα και περιεχόμενα προκειμένου να μην είναι αναγκασμένος ο αναλυτής να διαβάσει όλα τα δεδομένα αλλά να διαλέξει αυτά που τον ενδιαφέρουν. Θα προχωρήσουμε τώρα στην παρουσίαση τεχνικών που μας βοηθάνε στην πρόβλεψη.

Οι τεχνικές αυτές επιτρέπουν την ανίχνευση της απάτης από μία ασφαλιστική εταιρεία. Πιο συγκεκριμένα με βάση τις δηλώσεις που έχουμε από το ιστορικό αρχείο που διαθέτουμε θα χωρίσουμε τις δηλώσεις σε δύο ομάδες: εκείνες που έχουν αποδειχθεί απάτη και αυτές που δεν έχουν αποδειχθεί. Στην συνέχεια θα προσπαθήσουμε να δημιουργήσουμε ένα μοντέλο το οποίο θα προβλέπει αν μία δήλωση είναι πιο πιθανή από μία άλλη να είναι απάτη.

Σε κάθε μέθοδο πρόβλεψης το πρώτο βήμα είναι να διαλέξουμε ένα τυχαίο δείγμα από τα δεδομένα στα οποία θα εφαρμόσουμε τις τεχνικές πρόβλεψης και το υπόλοιπο δείγμα θα το χρησιμοποιήσουμε για να αξιολογήσουμε αν το μοντέλο πρόβλεψης που δημιουργήσαμε είναι καλό ή όχι και σε τι ποσοστό. Αυτή η διαδικασία γίνεται με τυχαίο τρόπο και σε τυχαίο ποσοστό προκειμένου να είναι αμερόληπτη η μελέτη.

Όπως και προηγουμένως στις τεχνικές ομαδοποίησης θα πρέπει να πραγματοποιήσουμε όλα τα βήματα της προεπεξεργασίας κειμένου:

α) Επιλογή των εγγράφων

β) Σημεία στίξης και κεφαλαία σε μικρά

γ)Tokenize

δ) Διαγραφή των λεγόμενων stopwords

ε) Stemming

στ) Διόρθωση της ορθογραφίας

προκειμένου να έχουμε μόνο την απαραίτητη για τη μελέτη μας πληροφορία.

Αφού ολοκληρώσουμε αυτό το βήμα θα πρέπει να προχωρήσουμε στην τεχνική πρόβλεψης θα χρησιμοποιήσουμε.

Στις επόμενες παραγράφους θα αναλύσουμε τις πιο γνωστές τεχνικές που με βάση μελέτες έχουν αποδειχθεί ότι είναι οι πιο αποτελεσματικές για τον προαναφερθέντα στόχο.

6.2. Λογιστική Παλινδρόμηση

Η πρώτη μέθοδος που θα ασχοληθούμε είναι η λογιστική παλινδρόμηση (logistic regression). Η λογιστική παλινδρόμηση είναι μια μέθοδος πολυπαραγοντικής στατιστικής ανάλυσης (multivariate statistical analysis) που χρησιμοποιεί ένα σύνολο ανεξαρτήτων μεταβλητών (independent variables) για τη διερεύνηση μιας κατηγορικής εξαρτημένης μεταβλητής (dependent variable). Η μέθοδος αυτή αποτελεί ειδική περίπτωση των γενικευμένων γραμμικών μοντέλων. Η λογιστική παλινδρόμηση (Logistic Regression) είναι χρήσιμη σε καταστάσεις στις οποίες επιθυμούμε την πρόβλεψη της ύπαρξης ή της απουσίας ενός χαρακτηριστικού ή ενός συμβάντος. Η πρόβλεψη αυτή βασίζεται στην κατασκευή ενός γραμμικού μοντέλου και συγκεκριμένα στον προσδιορισμό των τιμών που παίρνουν οι συντελεστές ενός συνόλου (set) ανεξάρτητων μεταβλητών που χρησιμοποιούνται ως μεταβλητές πρόβλεψης (predictor variables). Εκτός από την πρόβλεψη ένα μοντέλο λογιστικής παλινδρόμησης δίνει τη δυνατότητα να εκτιμήσουμε την επίδραση κάθε ανεξάρτητης μεταβλητής στη διαμόρφωση των τιμών της εξαρτημένης μεταβλητής. Στις περισσότερες εφαρμογές η εξαρτημένη μεταβλητή παίρνει δύο μόνο τιμές, για παράδειγμα αν μία δήλωση είναι γνήσια ή αν είναι απάτη. Οι τιμές της μεταβλητής αποτελούν μία αυθαίρετη κωδικοποίηση των δύο ενδεχομένων, συνήθως 0 και 1.

Το λογιστικό μοντέλο αποτελεί μια γενίκευση της απλής γραμμικής παλινδρόμησης για την περίπτωση που η εξαρτημένη μεταβλητή είναι δίτιμη. Είναι ένα μη γραμμικό μοντέλο, τα σφάλματα, του οποίου δεν υπακούν στην κανονική κατανομή και η μεταβλητή απόκρισης είναι διακριτή. Το μεγάλο πλεονέκτημα του λογιστικού μοντέλου έναντι του γραμμικού είναι ότι το γραμμικό μοντέλο είναι

αδύνατο να χρησιμοποιηθεί όταν η μεταβλητή είναι δυαδική λόγω των εξής προβλημάτων:

α) Τα σφάλματα δεν είναι κανονικά.

β) Τα σφάλματα έχουν άνισες διασπορές

γ) Περιορισμός στη συνάρτηση απόκριση (η μεταβλητή απόκρισης παίρνει τιμές 0 ή 1 και η μέση τιμή της βρίσκεται στο διάστημα [0,1])

Παρόλο που τα δύο πρώτα προβλήματα είναι δυνατό να τα αγνοήσουμε και να χρησιμοποιήσουμε την γραμμική παλινδρόμηση, εφαρμόζοντας κάποιες άλλες τεχνικές, το τρίτο πρόβλημα μας το απαγορεύει ρητά, γιατί δεν μπορεί να αντιμετωπιστεί με διαφορετικό τρόπο.

Ως γνωστόν η κατανομή Bernoulli είναι μία διακριτή κατανομής που περιγράφει ένα τυχαίο πείραμα με δύο πιθανά αποτελέσματα (επιτυχία – αποτυχία) και πιθανότητα επιτυχίας p. Η συνάρτηση πιθανότητας της κατανομής Bernoulli είναι:

$$f(x) = P(X = x) = \begin{cases} p, & x = 1 \\ q = p - 1, & x = 0 \end{cases}$$

Με μέση τιμή $E(X) = p$ και διακύμανση $V(X) = p(1 - p)$.

Η τυχαία μεταβλητή X που δίνει τον αριθμό επιτυχιών σε n δοκιμές Bernoulli με κοινή πιθανότητα επιτυχίας p λέγεται Διωνυμική (binomial) τυχαία μεταβλητή.

Η συνάρτηση πιθανότητας της Διωνυμικής κατανομής με παραμέτρους n και p δίνεται από τον τύπο:

$$f(x) = P(X = x) = \binom{n}{x} p^x q^{n-x} \ , x = 0,1,2,3, \dots \dots n$$

Με μέση τιμή $E(X) = np$ και διακύμανση $V(X) = npq$.

Εάν ορίσουμε την τιμή $Y = 1$ σαν επιτυχία και $Y = 0$ σαν αποτυχία τότε η Y είναι μία τυχαία μεταβλητή η οποία ακολουθεί την κατανομή Bernoulli, δηλαδή $Y \sim B(p)$. Η μέση τιμή λοιπόν της Y είναι $E(Y) = p$ και η διασπορά $V(Y) = p(1 - p)$.

Γενικεύοντας σε μία σειρά από n επαναλήψεις (δηλαδή πολλών πραγματοποιήσεων του πειράματος), ορίζουμε την τυχαία μεταβλητή:

Y =*αριθμός επιτυχιών σε* n *δοκιμές*

Αν η πιθανότητα επιτυχίας p είναι ίδια σε κάθε δοκιμή και οι δοκιμές είναι ανεξάρτητες μεταξύ τους τότε η νέα Y ακολουθεί την Διωνυμική κατανομή, δηλαδή $Y \sim b(n,p)$. Η μέση τιμή της Y είναι $E(Y) = np$ και διακύμανση $V(Y) = npq$.

Σε πολλές περιπτώσεις η τυχαία μεταβλητή Y εξαρτάται από κάποιες επεξηγηματικές μεταβλητές. Η εξάρτηση της Y από τις επεξηγηματικές μεταβλητές εισάγεται μέσω της εξάρτησης της πιθανότητας επιτυχίας p από τις $\boldsymbol{x}$. Για παράδειγμα, η πιθανότητα μία δήλωση να είναι απάτη εξαρτάται από την συχνότητα κάποιων λέξεων, την σειρά των λέξεων, το λεξιλόγιο που χρησιμοποιείται σε αυτή κλπ.

Το μοντέλο λογιστικής παλινδρόμησης είναι ένα γενικευμένο γραμμικό μοντέλο που περιγράφεται από τον τύπο:

$$n_x = g(E(Y_x)) = g(\mu_x) = \boldsymbol{x\beta}$$

και έχει την ακόλουθη δομή:

1. $Y_x \sim b(n_x, \mu_x)(n_x > 1, \delta\iota\omega\nu\upsilon\mu\iota\kappa\acute{\alpha}\ \delta\epsilon\delta o\mu\acute{\epsilon}\nu\alpha)$
2. $n_x = g(\mu_x) = ln\frac{n_x}{n_x - p_x} = ln\frac{p_x}{1-p_x} = logit(p_x) = \boldsymbol{x\beta}$ (συνάρτηση Logit)
3. $A\nu\epsilon\xi\acute{\alpha}\rho\tau\eta\sigma\acute{\iota}\alpha\ \mu\epsilon\tau\alpha\xi\acute{\upsilon}\ \tau\omega\nu\ \pi\alpha\rho\alpha\tau\eta\rho\acute{\eta}\sigma\epsilon\omega\nu\ Y_x$

Ο αριθμός n_x είναι το πλήθος των επαναλήψεων της τιμής του διανύσματος $\boldsymbol{x}$ των επεξηγηματικών μεταβλητών.

Έχουμε λοιπόν την σχέση:

$$ln\frac{p_x}{1-p_x} = n_x \Leftrightarrow p_x = \frac{e^{n_x}}{1+e^{n_x}}$$ όπου $0<p_x<1$.

Το $p_i = E(Y_i)$ συμβολίζει την πιθανότητα μία δήλωση να μην είναι απάτη ενώ το $1 - p_i = 1 - E(Y_i)$ συμβολίζει την πιθανότητα να είναι απάτη. Ο λόγος αυτών των δύο πιθανοτήτω ονομάζεται σχετική πιθανότητα (odd) του ενδεχομένου να μην είναι

απάτη η δήλωση και αν λογαριθμήσουμε αυτόν τον λόγο καταλήγουμε στην εξίσωση της Λογιστικής Παλινδρόμησης που είναι η εξής:

$$ln(\text{odds}) = ln\frac{E(Y_i)}{1-E(Y_i)} = ln\frac{P(\text{όχι απάτη})}{P(\text{απάτη})} = ln\frac{p_i}{1-p_i}$$

Για κάθε ξεχωριστή παρατήρηση i το μοντέλο γράφεται :

$$ln\frac{p_i}{1-p_i} = a + b_1x_{i1} + b_2x_{i2} + b_3x_{i3} + b_4x_{i4} \dots\dots + b_kx_{ik}, \quad i = 1, \dots n$$

Όπως μπορούμε να παρατηρήσουμε, το δεξί μέρος της τελευταίας ισότητας αποτελείται από ένα γραμμικό συνδυασμό ανεξάρτητων μεταβλητών ενώ στο αριστερό μέρος περιέχονται οι τιμές της εξαρτημένης μεταβλητής με τη μορφή του λογαρίθμου των odds. Αν η σχετική πιθανότητα είναι μεγαλύτερη του 1 τότε είναι πιο πιθανό η δήλωση μας να μην είναι απάτη.

Η πιθανότητα επιτυχίας (δηλαδή η πιθανότητα μια δήλωση να μην είναι απάτη) δίνεται από τον τύπο:

$$p_i = p_{xi} = \frac{\exp}{1+\exp} = \frac{1}{1+e^{-x_i\beta}}$$

Η ποσότητα:

$$\boldsymbol{x_i\beta} = a + b_1x_{i1} + b_2x_{i2} + b_3x_{i3} + b_4x_{i4} \dots\dots + b_nx_{ik}$$

λέγεται γραμμική συνάρτηση πρόβλεψης και μέσω αυτής υπολογίζεται η μέση τιμή με χρήση του τύπου:

$$E(Y_i) = n_ip_i = n_i\frac{e^{x_i\beta}}{1+e^{x_i\beta}}$$

Αφού αναλύσαμε την γραμμική συνάρτηση είναι σημαντικό να παρουσιάσουμε τον τρόπο με τον οποίο υπολογίζονται οι παράμετροι $a, b_1, b_2, b_3, \dots . b_n$.που συμμετάσχουν σε αυτό το μοντέλο καθώς και τον τρόπο με τον οποίο που πραγματοποιούνται έλεχγοι υποθέσεων για αυτές.

α. Εκτίμηση παραμέτρων με την Μέθοδο της Μέγιστης Πιθανοφάνειας

Η πιο αποτελεσματική μέθοδος εκτίμησης είναι η μέθοδος της μέγιστης πιθανοφάνειας. Ας υποθέσουμε ότι τα δεδομένα μας είναι χωρισμένα σε m κατηγορίες. Άρα $\sum_{i=1}^{m} n_i = n$ δηλαδή το πλήθος του δείγματος. Το μοντέλο μας βάσει της τελευταίας εξίσωσης έχει την μορφή:

$$E(Y_i) = n_i P(x_i)$$

όπου $P(x_i) = \frac{e^{x_i\beta}}{1+e^{x_i\beta}}$

και $y_1, y_2, y_3, \ldots . y_m$ είναι οι παρατηρούμενες τιμές των ανεξάρτητων τυχαίων διωνυμικών μεταβλητών.

Η συνάρτηση πιθανότητας μίας Διωνυμικής κατανομής όπως προείπαμε δίνεται από τον τύπο:

$$\binom{n}{x} p^y q^{n-y}$$

Άρα η πιθανοφάνεια για το λογιστικό μοντέλο παλινδρόμησης δίνεται από τον τύπο:

$$\ln[\mathcal{L}(P; \boldsymbol{y})] = \sum_{i=1}^{m} \{ y_i ln \left[\frac{P(x_i)}{1-P(x_i)}\right] + n_i \ln[1 - P(x_i)]\}$$

Ο όρος $\left\| ln \left[\frac{P(x_i)}{1-P(x_i)}\right]\right\|$ ονομάζεται *logit* και ισχύει όπως είδαμε και παραπάνω ότι:

$$ln \left[\frac{P(x_i)}{1-P(x_i)}\right] = \boldsymbol{x_i\beta} = a + b_1 x_{i1} + b_2 x_{i2} + b_3 x_{i3} + b_4 x_{i4} \ldots \ldots + b_n x_{ik} = \alpha + \sum_1^k x_{ij} b_j$$

Αντικαθιστώντας στην εξίσωση πιθανοφάνειας παίρνουμε:

$$\ln[\mathcal{L}(P; \boldsymbol{y})] = \sum_{i=1}^{m} \sum_{j=1}^{k} y_i \, y x_{ij} b_j - \sum_{i=1}^{m} n_i \ln \left(1 + exp \sum_{j=1}^{k} x_{ij} b_j\right)$$

ή αλλιώς

$$\ln[\mathcal{L}(P;\boldsymbol{y})] = \boldsymbol{\beta} X'\boldsymbol{y} - \sum_{i=1}^{m} n_i \ln\,(1+\exp\,(\boldsymbol{x_i\beta}))$$

Ο X' είναι ο πίνακας που συναντάμε και στην γραμμική παλινδρόμηση και $\boldsymbol{y}$ είναι το διάνυσμα απόκρισης.

Προκειμένου να εκτιμήσουμε τις παραμέτρους πρέπει να μεγιστοποιηθεί η παραπάνω παράσταση ως προς τις παραμέτρους. Επομένως χρειάζεται να παραγωγίσουμε την παράσταση αυτή ως προς β και να βρούμε σε ποια σημεία μηδενίζεται (βρίσκουμε τα σαγματικά σημεία).

Ξεκινάμε λοιπόν από την παραγώγιση ως προς β και έχουμε:

$$\frac{d\ln[\mathcal{L}(P;\boldsymbol{y})]}{d\beta} = X'\boldsymbol{y} - \sum_{i=1}^{m} \frac{n_i}{1+\exp\,(\boldsymbol{x_i\beta})} \exp\,(\boldsymbol{x_i\beta})\boldsymbol{x_i}$$

Σύμφωνα με όσα προείπαμε η πιθανότητα ισούται με :

$$P(x_i) = \frac{exp(a + b_1x_{i1} + b_2x_{i2} + b_3x_{i3} + b_4x_{i4} \dots\dots + b_kx_{ik})}{1 + exp(a + b_1x_{i1} + b_2x_{i2} + b_3x_{i3} + b_4x_{i4} \dots\dots + b_kx_{ik})} = \frac{1}{1+e^{-x_i\beta}}$$

$$\Leftrightarrow P(x_i) = \frac{e^{x_i\beta}}{1+e^{x_i\beta}}$$

Άρα αντικαθιστώντας στην τελευταία εξίσωση έχουμε:

$$\frac{d\ln[\mathcal{L}(P;\boldsymbol{y})]}{d\boldsymbol{\beta}} = X'\boldsymbol{y} - \sum_{i=1}^{m} n_i P(x_i) x_i$$

και αφού

$$E(y_i) = n_i P(x_i) = \mu_i$$

προκύπτει ότι

$$\frac{d\ln[\mathcal{L}(P;\boldsymbol{y})]}{d\boldsymbol{\beta}} = X'\boldsymbol{y} - \sum_{i=1}^{m} E(y_i)x_i = X'y - \sum_{i=1}^{m} \mu_i x_i' = X'\boldsymbol{y} - X'\boldsymbol{\mu} = X'(\boldsymbol{y} - \boldsymbol{\mu})$$

Για να βρούμε τα σημεία που μεγιστοποιείται η πιθανοφάνεια πρέπει να βρούμε τα σημεία που η τελευταία σχέση μηδενίζεται:

$$\frac{d\ln[\mathcal{L}(P;\boldsymbol{y})]}{d\boldsymbol{\beta}} = X'(\boldsymbol{y} - \boldsymbol{\mu}) = 0$$

Για την λύση της τελευταίας εξίσωσης χρησιμοποιείται συνήθως η μέθοδος των σταθμισμένων ελάχιστων τετραγώνων (weighted least squares). Ουσιαστικά μέσα από μία επαναληπτική διαδικασία παράγουμε τις εκτιμήσεις των όρων $a, b_1, b_2, b_3, b_4 \ldots\ldots, b_k$ των όρων $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3 \ldots\ldots \hat{\beta}_k$. Πιο συγκεκριμένα θα χρησιμοποιήσουμε τον τύπο:

$$S = \sum_{i=1}^{m} \left[\frac{(y_i - \mu_i)^2}{\sigma_i^2}\right]$$

Όπου με σ_i^2 συμβολίζουμε τη διακύμανση της τυχαίας διωνυμικής κατανομής.

Σε αυτό το σημείο θέλουμε να ελαχιστοποιήσουμε το S ως προς β. Παραγωγίζοντας ως προς β παίρνουμε:

$$\frac{dS}{d\beta} = \frac{d}{d\beta}\sum_{i}^{m}\left[\frac{(y_i - \mu_i)^2}{\sigma_i^2}\right] = \left[\frac{\sum_i^m 2(y_i - \mu_i)}{\sigma_i^2}\right]\frac{d\mu_i}{d\beta} \qquad (1)$$

Όμως γνωρίζουμε για την μέση τιμή ότι:

$$E(y_i) = \mu_i = n_i P(x_i) = n_i \frac{e^{x_i\beta}}{1+e^{x_i\beta}}$$

και παραγωγίζοντας έχουμε:

$$\frac{d\mu_i}{d\beta} = \frac{d}{d\beta}\left(n_i \frac{e^{x_i\beta}}{1+e^{x_i\beta}}\right) = n_i \frac{e^{x_i\beta}x_i(1+e^{x_i\beta}) - e^{x_i\beta}x_i e^{x_i\beta}}{(1+e^{x_i\beta})^2} = \frac{n_i e^{x_i\beta}x_i}{(1+e^{x_i\beta})^2} \quad (2)$$

Ακόμα ισχύει

$$\sigma_i^2 x_i = (n_i P(x_i) - n_i P(x_i)^2)x_i = n_i x_i \left(\frac{e^{x_i\beta}}{1+e^{x_i'\beta}} - \frac{e^{x_i\beta^2}}{1+e^{x_i'\beta^2}}\right) \Leftrightarrow$$

$$\sigma_i^2 x_i = n_i x_i \frac{e^{x_i\beta}(1+e^{x_i'\beta}) - e^{x_i\beta^2}}{1+e^{x_i'\beta^2}} = n_i x_i \frac{e^{x_i\beta}(1+e^{x_i\beta} - e^{x_i\beta})}{1+e^{x_i'\beta^2}} \Leftrightarrow$$

$$\sigma_i^2 x_i = \frac{n_i x_i e^{x_i\beta}}{(1+e^{x_i\beta})^2} \quad (3)$$

Από (2) και (3) έχουμε :

$$\frac{d\mu_i}{d\beta} = \sigma_i^2 x_i$$

Αντικαθιστώντας την τελευταία σχέση στη σχέση (1) και θεωρώντας τη σχέση (1) ίση με το μηδέν αφού θέλουμε βρούμε τα σημεία στα οποία η σχέση ελαχιστοποιείται παίρνουμε:

$$\frac{dS}{d\beta} = \left[\frac{\sum_i^m 2(y_i-\mu_i)}{\sigma_i^2}\right]\sigma_i^2 x_i = 0 \Leftrightarrow \frac{dS}{d\beta} = \sum_i^m (y_i - \mu_i)x_i = X'(y-\mu) = 0$$

Η τελευταία συνθήκη είναι παρόμοια με αυτή που βρήκαμε πιο πάνω. Άρα μία επαναληπτική μέθοδος όπως η παραπάνω μπορεί να χρησιμοποιηθεί για να προσδιορισθούν οι εκτιμητές μέγιστης πιθανοφάνειας.Οι εκτιμήτριες μέγιστης πιθανοφάνειας συμβολίζονται συνήθως με $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3 \ldots\ldots \hat{\beta}_k$ ή αλλιώς :

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{bmatrix}$$

Και υποθέτοντας ότι ο πίνακας $X'X$ είναι αντιστρέψιμος ισχύει ότι :

$$\hat{\beta} = (X'X)^{-1}X'y$$

Τέλος μπορούμε να γράψουμε:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots \ldots + \hat{\beta}_k x_{ik}$$

Ή αλλιώς $\hat{\boldsymbol{Y}} = X\hat{\boldsymbol{\beta}}$

όπου X ο πίνακας σχεδιασμού του μοντέλου.

β. Έλεγχοι Υποθέσεων

Αφού βρήκαμε τις εκτιμήτριες μέγιστης πιθανοφάνειας είναι σημαντικό να διεξάγουμε ελέγχους υποθέσεων προκειμένου να δούμε κατά πόσο είναι σημαντικοί ή όχι στο μοντέλο πρόβλεψης που έχουμε δημιουργήσει και στην συνέχεια να αναζητήσουμε το καλύτερο δυνατό μοντέλο, δηλαδή αυτό με τις λιγότερες παραμέτρους

Ο κανόνας απόφασης για την υπόθεση :

$$H_0: \hat{\beta}_i = 0$$

$$H_1: \hat{\beta}_i \neq 0$$

βασίζεται στην στατιστική συνάρτηση $z_i = \frac{b_i - \hat{\beta}_i}{\hat{\sigma}_{\hat{\beta}_i}}$

Όπου κάθε z_i ακολουθεί την κανονική κατανομή με μέση τίμη ίση με 0 και διακύμανση 1 δηλαδή $z_i \sim N(0,1)$

Ισχύει ακόμα ότι $z_i^2 = \left(\frac{b_i}{\hat{\sigma}_{b_i}}\right)^2$ ακολουθεί ασυμπτωτικά την X_1^2 κατανομή υπό την υπόθεση H_0.

Ο έλεγχος που θα διεξάγουμε είναι ο συνηθισμένος μονοπλευρος ή δίπλευρος έλεγχος και ο υπολογισμός των τιμών X_1^2 και των $p - value$ πραγματοποιείται με χρήση στατιστικών πακέτων. Ανάλογα με το επίπεδο σημαντικότητας που θα επιλέξουμε θα κάνουμε και την αντίστοιχη σύγκριση των $p - values$.

6.3. Support Vector Machine

Μία άλλη μέθοδος πρόβλεψης που θεωρείται πολύ αποτελεσματική είναι η λεγόμενη Support Vector Machines ή αλλιώς SVM. Είναι μία τεχνική μάθησης που σκοπό έχει τη δημιουργία μιας συνάρτησης απόφασης βασιζόμενη σε ένα σύνολο δεδομένων εκπαίδευσης. Τα δεδομένα εκπαίδευσης αποτελούνται από ζεύγη μεταβλητών πρόβλεψης και τιμές στόχους. Κάθε μεταβλητή πρόβλεψης έχει τιμή στόχο. Εάν ο αλγόριθμος μπορεί να προβλέψει ένα χαρακτηριστικό του κειμένου τότε ονομάζεται συνάρτηση ταξινόμησης (classification function) ενώ αν ο αλγόριθμος προβλέπει αριθμητική αξία τότε ονομάζεται παλινδρόμηση (regression).

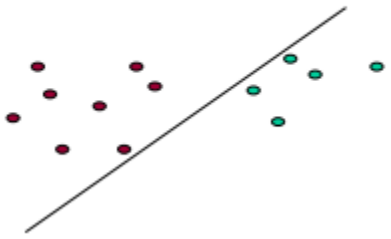
Το SVM είναι μία μηχανή μάθησης στην οποία «δίνεται» ένα σύνολο δεδομένων εκπαίδευσης. Αυτά τα δεδομένα μετατρέπονται σε διανύσματα και αποτελούν ένα υποσύνολο του R^n. Η συνάρτηση απόφασης βέβαια μπορεί να είναι αρκετά πολύπλοκη με αρκετές μεταβλητές. Αλλά για μία δεδομένη είσοδο θα βγάζει πάντα το ίδιο αποτέλεσμα. Εφόσον έχουμε μόνο δύο περιπτώσεις ο στόχος είναι να κατασκευάσουμε ένα δυαδικό ταξινομητή από τα δείγματα εκπαίδευσης. Στην περίπτωση του Text Mining που έχουμε κείμενα, τα διανύσματα αποτελούνται κυρίως από συχνότητες λέξεων κλειδιών.

Οι συλλογές κειμένων περιέχουν εκατομμύρια όρους κάτι το οποίο καθιστά την διαδικασία του Text Mining ιδιαίτερα δύσκολη.

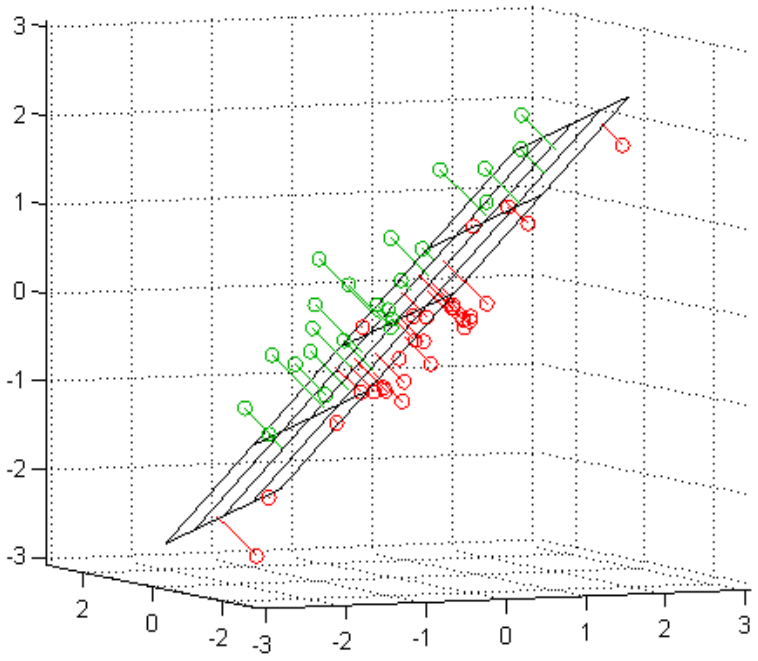
Ένα SVM λαμβάνει δεδομένα-παρατηρήσεις ως είσοδο και εξάγει μια συνάρτηση που μπορεί να χρησιμοποιηθεί για πρόβλεψη σε μελλοντικά δεδομένα. Βασική ιδέα για την περίπτωση των γραμμικών μοντέλων είναι η εύρεση του βέλτιστου υπερεπιπέδου διαχωρισμού και για τα μη γραμμικά γίνεται διαχωρισμός με

χρήση συναρτήσεων πυρήνα οι οποίες μετασχηματίζουν τα αρχικά δεδομένα σε ένα νέο χώρο

Στα SVM χρησιμοποιείται συχνά η έννοια του υπερεπιπέδου. Για να γίνει πιο κατανοητός ο όρος υπερεπίπεδα ας υποθέσουμε ότι έχουμε δύο ομάδες σημείων, τα πράσσινα και τα κόκκινα. Αυτό που προσπαθούμε να επιτύχουμε είναι να διαχωρίσουμε τα κόκκινα από τα πράσινα σημεία. Όπως είναι εμφανές και από το παρακάτω σχήμα υπάρχει ευθεία, η οποία διαχωρίζει τις δύο ομάδες.



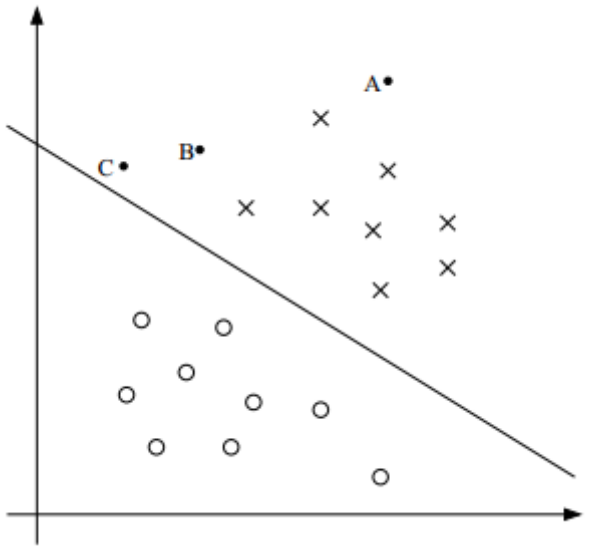
Έτσι λοιπόν σε δύο διαστάσεις τα σημεία μπορούν να χωρισθούν από μία ευθεία, δηλαδή ένα υπερεπίπεδο μίας διάστασης. Σε τρεις διαστάσεις όπως φαίνεται και στο παρακάτω σχήμα μπορούν να χωρισθούν από ένα επίπεδο, δηλαδή ένα υπερεπίπεδο δύο διαστάσεων.



Σε παραπάνω από τρεις διαστάσεις των SVM χωρίζονται από υπερεπίπεδα μίας διάστασης μικρότερης από της διάστασης του χώρου που ορίζονται τα SVM.

Ο κύριος σκοπός αυτής της μεθόδου είναι να βρούμε το βέλτιστο υπερεπίπεδο το οποίο διαχωρίζει καλύτερα τα σημεία μας.

Έστω ότι έχουμε το παρακάτω σχήμα στο οποίο με × παρουσιάζονται τα δεδομένα εκπαίδευσης του μοντέλου τα οποία έχουν θετικά χαρακτηριστικά και με ο τα δεδομένα εκπαίδευσης τα οποία έχουν αρνητικά χαρακτηριστικά. Η γραμμή διαχωρισμού των δύο ομάδων περιγραφεται μαθηματικά από τον τύπο $\theta^T x = 0$ και ονομάζεται διαχωριστικό υπερεπίπεδο.

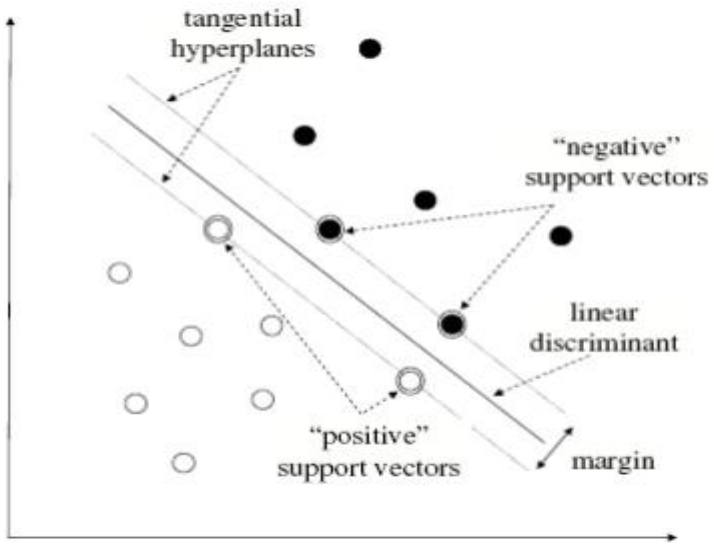


Στο παραπάνω σχήμα υπάρχουν και τα σημεία A, B, C τα οποία δεν ξέρουμε που ανήκουν. Αν έπρεπε να κάνουμε μία πρόβλεψη κατά πόσο το A είναι θετικό, θα μπορούσαμε να πούμε ότι είναι καθώς βρίσκεται αρκετά μακριά από το όριο απόφασης. Όσον αφορά όμως σημείο C, αυτό βρίσκεται πολύ κοντά στο όριο απόφασης και μία μικρή μετατόπιση του ορίου θα μας έδινε άλλο αποτέλεσμα πρόβλεψης. Καταλήγουμε λοιπόν στο συμπέρασμα ότι για το A είμαστε αρκετά πιο σίγουροι από ότι για το C προκειμένου να κάνουμε μία πρόβλεψη.

Είναι σαφές ότι στόχος μας είναι να δημιουργήσουμε το βέλτιστο υπερεπίπεδο με την έννοια αυτό να χωρίζει τα δεδομένα με την μεγαλύτερη δυνατή απόσταση.

Το βέλτιστο υπερεπίπεδο ονομάζεται υπερεπίπεδο μέγιστου εύρους (maximum margin hyperplane). Περιγράφουμε στη συνέχεια τη μέθοδο με την οποία γίνεται η κατασκευή του υπερεπιπέδου μέγιστου εύρους.

Δημιουργούμε δύο παράλληλα υπερεπίπεδα τέτοια ώστε να μην υπάρχουν ανάμεσά τους δεδομένα του συνόλου εκπαίδευσης όπως φαίνεται και στο παρακάτω σχήμα.



Τα σημεία που βρίσκονται πάνω σε αυτά τα δύο υπερεπίπεδα ονομάζονται support vectors και γι' αυτόν τον λόγο και το μοντέλο ονομάζεται Support Vector Machine. Η απόσταση που είναι και η μέγιστη μεταξύ των δύο παράλληλων υπερεπιπέδων ονομάζεται εύρος ($margin$).

Η εξίσωση η οποία συμβολίζει το υπερεπίπεδο έχει την εξής μορφή:

$$\boldsymbol{w} \cdot \boldsymbol{d} + b = 0$$

Με $\boldsymbol{w}$ και b να είναι οι παράμετροι του μοντέλου

Έστω $D = \{d_1, d_2, d_3, \dots\dots, d_n\}$ το σύνολο των δεδομένων εκπαίδευσης και $C = \{c_1, c_2\}$ το σύνολο των κατηγοριών. Πιο συγκεκριμένα $c_i \in \{-1,1\}$ όπου $c_i = 1$ σημαίνει ότι μία δήλωση είναι γνήσια και $c_i = -1$ σημαίνει ότι μία δήλωση είναι απάτη.

Για όσων δεδομένων τα διανύσματα τους βρίσκονται πάνω στο υπερεπίπεδο θα παληθεύουν την εξίσωση $\boldsymbol{w} \cdot \boldsymbol{d} + b = 0$ ενώ τα διανύσματα των υπόλοιπων δεδομένων θα επαληθεύουν την εξίσωση $\boldsymbol{w} \cdot \boldsymbol{d} + b = m$. Συγκεκριμένα όταν ισχύει $\boldsymbol{w} \cdot \boldsymbol{d} + b > 0$ τότε όπως είναι λογικό τα δεδομένα βρίσκονται πάνω από το υπερεπίπεδο-όριο και για όσα ισχύει ότι $\boldsymbol{w} \cdot \boldsymbol{d} + b < 0$ θα είναι κάτω από το

υπερεπίπεδο.Έχουμε επίσης την δυνατότητα να ορίσουμε τις παραμέτρους του μοντέλου έτσι ώστε τα παράλληλα υπερεπίπεδα να εκφράζονται ως (canonical hyperplanes):

$$\boldsymbol{w}\cdot\boldsymbol{d}+b=1$$
$$\boldsymbol{w}\cdot\boldsymbol{d}+b=-1$$

Θεωρώντας λοιπόν ένα σημείο $\boldsymbol{d_1}$ πάνω στο πρώτο υπερεπίπεδο και ένα σημείο $\boldsymbol{d_2}$ πάνω στο δεύτερο υπερεπίπεδο μπορούμε πολύ εύκολα να βρούμε το εύρος ($margin$):

$$\left.\begin{matrix}\boldsymbol{w}\cdot\boldsymbol{d_1}+b=1\\ \boldsymbol{w}\cdot\boldsymbol{d_2}+b=-1\end{matrix}\right\}\Leftrightarrow \boldsymbol{w}(\boldsymbol{d_1}-\boldsymbol{d_2})=2\Leftrightarrow \|\boldsymbol{w}\|\times margin=2\Leftrightarrow margin=\frac{2}{\|\boldsymbol{w}\|}$$

6.4. Γραμμικά διαχωρίσιμα δεδομένα

Ας υποθέσουμε αρχικά ότι το σύνολο εκπαίδευσης είναι γραμμικά διαχωρίσιμο. Το ζητούμενο είναι να βρούμε κατάλληλα $\boldsymbol{w}, \boldsymbol{b}$ προκειμένου πρώτον το εύρος να είναι το μεγαλύτερο δυνατό και δεύτερον οι δηλώσεις της κατηγορίας $\boldsymbol{c_1}$ να βρίσκονται πάνω από το υπερεπίπεδο και οι δηλώσεις της κατηγορίας $\boldsymbol{c_2}$ να βρίσκονται από κάτω.

Οι παραπάνω περιορισμοί εξασφαλίζονται λύνοντας το παρακάτω πρόβλημα βελτιστοποίησης:

$$min\,\frac{\|\boldsymbol{w}\|^2}{2}$$

Και

$$c_i(\boldsymbol{w}\cdot\boldsymbol{d_i}+b)\geq 1 \qquad i=1,2,\ldots,n$$

Για να λύσουμε το συγκεκριμένο πρόβλημα θα κάνουμε χρήση των πολλαπλασιαστών *Lagrange*:

$$L_p = \frac{\|\boldsymbol{w}\|^2}{2} - \sum_{i=1}^{n} \lambda_i (c_i(\boldsymbol{w} \cdot \boldsymbol{d_i} + b) - 1)$$

όπου λ_i κατάλληλες βοηθητικές παράμετροι οι οποίες ονομάζονται πολλαπλασιαστές *Lagrange*. Θέλουμε να ελαχιστοποιήσουμε την παραπάνω ποσότητα οπότε θα πρέπει να την παραγωγίσουμε ως προς τις παραμέτρους και να εξισώσουμε τις παραγώγους ίσες με 0. Έτσι παίρνουμε

$$\frac{d}{d\boldsymbol{w}} L_p = \boldsymbol{w} - \sum_{i=1}^{n} \lambda_i c_i d_i = 0$$

$$\frac{d}{db} L_p = \sum_{i=1}^{n} \lambda_i c_i = 0$$

Και οι δύο εξισώσεις περιλαμβάνουν τους πολλαπλασιαστές *Lagrange* οι οποίοι δεν έχουν γνωστές τιμές.

Οι περιορισμοί που υπάρχουν στο πρόβλημα βελτιστοποίησης εκφράζονται με ανισότητα και όχι με ισότητα και δεν μπορούμε να τις χρησιμοποιήσουμε στην εύρεση της λύσης. Γι' αυτό τον λόγο θα μετατρέψουμε τους περιορισμούς σε ισότητες με την χρήση των συνθηκών *Karush-Kuhn-Tucker (KKT):*

$$\lambda_i \geq 0$$
$$\lambda_i(c_i(\boldsymbol{w} \cdot \boldsymbol{d_i} + b) - 1) = 0$$

Θα πρέπει να επισημάνουμε ότι οι πολλαπλασιαστές *Lagrange* είναι διάφοροι του μηδενός μόνο για τα κείμενα τα οποία βρίσκονται ακριβώς πάνω στα παράλληλα υπερεπίπεδα και αποτελούν τα support vectors.

Τελικά για να επιλύσουμε το σύστημα το μετατρέπουμε στο ισοδύναμο δυικό *(dual problem)* σχηματίζοντας την ποσότητα:

$$L_D = \sum_{i=1}^{n} \lambda_i - \frac{1}{2} \sum_{i,j} \lambda_i \lambda_j c_i c_j \boldsymbol{d_i} \cdot \boldsymbol{d_j}$$

όπου $\lambda_i \geq 0$ *και* $\sum_i \lambda_i c_i = 0$

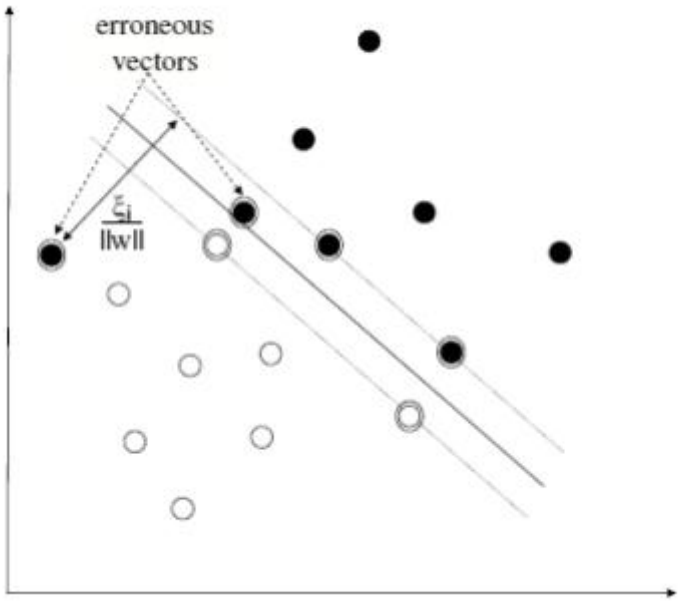
Με την επίλυση του δυικού προβλήματος θα βρούμε τους πολλαπλασιαστές *Lagrange* και με τις συνθήκες *KKT* και των παραγώγων θα βρούμε τις παραμέτρους ***w,b.***.

Η εύρεση της λύσης γίνεται συνήθως με τεχνικές τετραγωνικού προγραμματισμού και η επιφάνεια απόφασης εκφράζεται ως:

$$\left(\sum_{i}^{n} \lambda_i c_i \boldsymbol{d_i} \cdot \boldsymbol{d}\right) + b = 0$$

6.5. Γραμμικά διαχωρίσιμα με λίγα σφάλματα (soft margin)

Στις περισσότερες περιπτώσεις τα δεδομένα μας δεν είναι εύκολα διαχωρίσιμα από ένα υπερεπίπεδο και γι' αυτό πρέπει να αναζητήσουμε άλλους τρόπους διαχωρισμού. Ένας από τους τρόπους που βρίσκουμε λύση σε αυτό το πρόβλημα είναι να τα διαχωρίσουμε γραμμικά επιτρέποντας όμως κάποια σφάλματα με στόχο να επιτύχουμε το καλύτερο εύρος μεταξύ των δύο κατηγοριών.



Η ποσότητα που θέλουμε να ελαχιστοποιήσουμε αλλάζει σε αυτήν την περίπτωση και γίνεται:

$$min\left(\frac{\|\boldsymbol{w}\|^2}{2}+C\sum_{i}^{n}\xi_i\right)$$

Όπου $\xi_i \geq 0$ και $c_i(\boldsymbol{w}\cdot\boldsymbol{d_i}+b) \geq 1-\xi_i \quad i=1,2,\dots\dots,n.$

Τα ξ_i είναι απαραίτητα ώστε να επιτρέψουν να συμβαλινουν σφάλματα στο σύνολο εκπαίδευσης. Για κάθε δήλωση που βρίσκεται στην σωστή πλευρά υπερεπιπέδου με βάση τον διαχωρισμό τα $\xi_i = 0$ ενώ στις άλλες περιπτώσεις μετράνε το σφάλμα και συγκεκριμένα η απόσταση μεταξύ του σημείου και του παράλληλου υπερεπίπεδου που αντιστοιχεί στην κατηγορία αυτού είναι ($\frac{\xi_i}{\|\boldsymbol{w}\|}$).

Η μεταβλητή C καθορίζεται από εμάς και όσο μεγαλύτερη τιμή της δίνουμε τόσο πιο αυστηροί είμαστε στα λάθη.

Στην καινούργια ποσότητα που θα ελαχιστοποιήσουμε θα προστεθούν πολλαπλασιαστές *Lagrange* μ_i για τα $\xi_i \geq 0$. Έτσι φτάνουμε στην ποσότητα:

$$L_p=\frac{\|\boldsymbol{w}\|^2}{2}-C\sum_{i=1}^{n}\xi_i-\sum_{i=1}^{n}\lambda_i(c_i(\boldsymbol{w}\cdot\boldsymbol{d_i}+b)-1+\xi_i)-\sum_{i=1}^{n}\mu_i\xi_i$$

Αφού θέλουμε να ελαχιστοποιήσουμε την παραπάνω συνάρτηση την παραγωγίζουμε και παίρνουμε:

$$\frac{dL_p}{d\boldsymbol{w}}=0\Leftrightarrow \boldsymbol{w}=\textstyle\sum_{i=1}^{n}\lambda_i c_i\boldsymbol{d_i}\,,$$

$$\frac{dL_p}{db}=0\Leftrightarrow 0=\textstyle\sum_{i=1}^{n}\lambda_i c_i\,,$$

$$\frac{dL_p}{d\xi_i}=0\Leftrightarrow C=\lambda_i+\mu_i\,.$$

Με τον ίδιο τρόπο που εργασθήκαμε προηγουμένως και με την χρήση των *KKT* το σύστημα που θέλουμε να λύσουμε παίρνει τη μορφή:

$$\lambda_i \geq 0, \mu_i \geq 0, \xi_i \geq 0,$$

$$\lambda_i(c_i(\boldsymbol{w} \cdot \boldsymbol{d_i} + b) - 1 + \xi_i) = 0,$$

$$\mu_i \xi_i = 0.$$

Στους νέους περιορισμούς ισχύει ότι τα λ_i είναι μη μηδενικά για τα δεδομένα που βρίσκονται πάνω στο παράλληλο υπερεπίπεδο στην κατηγορία τους όπως και προηγουμένως, αλλά στην συγκεκριμένη περίπτωση ισχύει και όταν $\xi_i > 0$, ενώ οι νέες μεταβλητές μ_i είναι ίσες με 0 μόνο όταν $\xi_i > 0$. Η νέα εξίσωση μεγιστοποίησης του δυικού είναι η εξής:

$$L_D = \sum_{i=1}^{n} \lambda_i - \frac{1}{2} \sum_{i,j} \lambda_i \lambda_j c_i c_j \boldsymbol{d_i} \cdot \boldsymbol{d_j}$$

Όπου $0 \leq \lambda_i \leq C$ και $\sum_{i=1}^{n} \lambda_i c_i = 0$

Τέλος η επιφάνεια απόφασης είναι η εξής:

$$\left(\sum_{i=1}^{n} \lambda_i c_i \boldsymbol{d_i} \cdot \boldsymbol{d} \right) + b = 0$$

Εύκολα μπορούμε να παρατηρήσουμε ότι η εξίσωση της επιφάνειας απόφασης είναι ίδια με την προηγούμενη περίπτωση ενώ στο δυικό η μόνη διαφορά είναι ότι υπάρχει άνω όριο στα λ_i το οποία είναι το C.

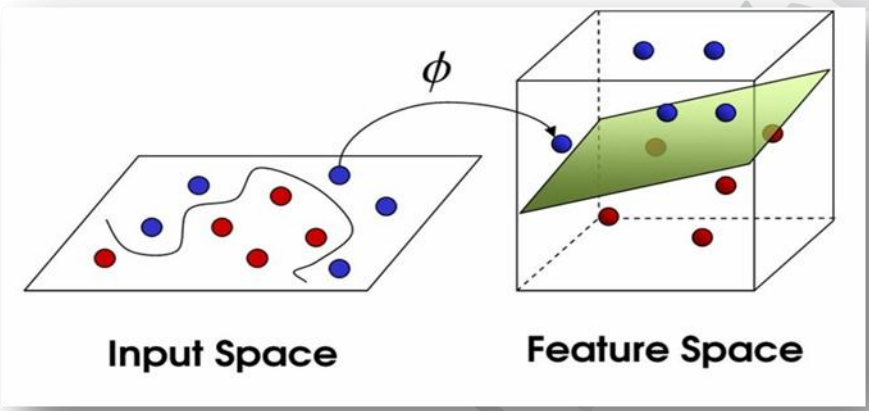
6.6. Μη γραμμικό SVM

Εκτός της παραπάνω προσέγγισης υπάρχει και άλλος τρόπος να διαχωρίσουμε δεδομένα τα οποία δεν είναι γραμμικά διαχωρίσιμα. Αυτό πραγματοποιείται με τις μη γραμμικές *SVM*. Η μέθοδος που χρησιμοποιούμε σε αυτήν την περίπτωση είναι να απεικονίσουμε τα δεδομένα μας σε ένα χώρο μεγαλύτερης διάστασης από τον κανονικό και να τα διαχωρίσουμε γραμμικά σε αυτόν.

Συγκεκριμένα οι δύο βασικές μαθηματικές πράξεις στις οποίες στηρίζονται οι μη γραμμικές *Support Vector Machines* είναι:

1. Μη γραμμικός μετασχηματισμός του κάθε διανύσματος εισόδου σε ένα χώρο υψηλού αριθμού διαστάσεων, ο οποίος είναι κρυφός για τον χώρου εισόδου και εξόδου.

2. Κατασκευή του βέλτιστου υπερεπιπέδου για να διαχωρίζονται τα χαρακτηριστικά του χώρου του βήματος 1.



Το Βήμα 1 στηρίζεται στο θεώρημα του Cover για την δυνατότητα διαχωρισμού των προτύπων. Ας σκεφτούμε ένα χώρο εισόδου ο οποίος δημιουργείται από μη γραμμικά διαχωρίσιμα πρότυπα. Το θεώρημα Cover λέει πως αυτός ο χώρος πολλών διαστάσεων μπορεί να μεταμορφωθεί σε ένα νέο χώρο χαρακτηριστικών, όπου τα πρότυπα είναι γραμμικά διαχωρίσιμα με μεγάλη πιθανότητα, αν ικανοποιούνται δύο συνθήκες. Πρώτον, ο μετασχηματισμός να είναι μη γραμμικός και δεύτερον ο αριθμός των διαστάσεων να είναι αρκετά μεγάλος. Αυτές οι δύο συνθήκες ικανοποιούνται στην δικιά μας περίπτωση.

Εφόσον καταφέρουμε να βρούμε το νέο υπερεπίπεδο μεγαλύτερης διάστασης εργαζόμαστε με τον ίδιο τρόπο όπως στις γραμμικές *SVM*. Ουσιαστικά αυτό που θέλουμε να δούμε είναι με ποιο τρόπο οδηγούμαστε στην μεγαλύτερη διάσταση.

Έστω ότι έχουμε μία νέα συνάρτηση $\Phi(\boldsymbol{d})$ η οποία απεικονίζει τα κείμενα μας στον χώρο μεγαλύτερης διάστασης. Η νέα γραμμική εξίσωση θα είναι της μορφής :

$$\boldsymbol{w} \cdot \Phi(\boldsymbol{d}) + b = 0$$

Εργαζόμενοι με τον ίδιο ακριβώς τρόπο με τις προηγούμενες μεθόδους το πρόβλημα βελτιστοποίησης μας είναι :

$$min\left(\frac{\|\boldsymbol{w}\|^2}{2}\right) + C\sum_{i=1}^{n}\xi_i$$

$$\xi_i \geq 0 \text{ και } c_i(\boldsymbol{w}\cdot\boldsymbol{\Phi}(\boldsymbol{d_i}) + b) \geq 1 - \xi_i \ , \quad i = 1,2,\ldots.n$$

Συνεχίζοντας με την ίδια λογική, η εξίσωση του δυικού προβλήματος θα είναι η εξής:

$$L_D = \sum_{i=1}^{n}\lambda_i - \frac{1}{2}\sum_{i,j}\lambda_i\lambda_j c_i c_j \boldsymbol{\Phi}(\boldsymbol{d_i})\cdot\boldsymbol{\Phi}(\boldsymbol{d_j})$$

$$\text{Όπου } 0 \leq \lambda_i \leq C \text{ και } \textstyle\sum_{i=1}^{n}\lambda_i c_i = 0$$

Αφού βρούμε τα λ_i, για να υπολογίσουμε τις παραμέτρους χρησιμοποιώντας τις $(\boldsymbol{w}, b)$ χρησιμοποιούμε τις συνθήκες *KKT* και παραγωγίζοντας ως προς w έχουμε τις εξής εξισώσεις:

$$\boldsymbol{w} = \sum_{i=1}^{n}\lambda_i c_i \boldsymbol{\Phi}(\boldsymbol{d_i})$$

$$\lambda_i(c_i(\boldsymbol{w}\cdot\boldsymbol{\Phi}(\boldsymbol{d_i}) + b) - 1 + \xi_i) = 0$$

$$\Leftrightarrow$$

$$\lambda_i\left(c_i\left(\sum_{j=1}^{n}\lambda_j c_j \boldsymbol{\Phi}(d_j)\cdot\boldsymbol{\Phi}(d_i) + b\right) - 1 + \xi_i\right) = 0$$

Τελικά η νέα επιφάνεια απόφασης παίρνει την μορφή:

$$\left(\sum_{j=1}^{n}\lambda_j c_j \boldsymbol{\Phi}(d_j)\cdot\boldsymbol{\Phi}(d)\right) + b = 0$$

Η διαφορά με τις άλλες εξισώσεις επιφάνειας απόφασης είναι ότι περιλαμβάνεται και η συνάρτηση $\boldsymbol{\Phi}(\boldsymbol{d})$. Την συγκεκριμένη συνάρτηση δεν την υπολογίζουμε ακριβώς αλλά χρησιμοποιούμε μια έμμεση μέθοδο και συγκεκριμένα την χρήση πυρήνων. Όπως μπορούμε να παρατηρήσουμε στην εξίσωση επιφάνειας απόκρισης απαιτείται όχι ο υπολογισμός της συνάρτησης $\boldsymbol{\Phi}(\boldsymbol{d})$ αλλά του εσωτερικού γινομένου το οποίο εκφράζει την ομοιότητα μεταξύ δύο διανυσμάτων. Σε αυτό ακριβώς μας βοηθά η χρήση πυρήνων.

6.7. Χρήση Πυρήνων

Ένα πυρήνα μπορούμε να τον συμβολίσουμε με τον εξής τρόπο:

$$K(\boldsymbol{u},\boldsymbol{v}) = \Phi(\boldsymbol{u}) \cdot \Phi(\boldsymbol{v})$$

Οι πυρήνες ικανοποιούν το θεώρημα *Mercer* το οποίο μας εξασφαλίζει ότι μπορούν να εκφραστούν ως εσωτερικά γινόμενα σε ένα χώρο μεγαλύτερης διάστασης. Η εξίσωση του υπερεπιπέδου απόφασης σε αυτήν την περίπτωση είναι της μορφής:

$$\left(\sum_{i=1}^{n} \lambda_i \, c_i K(\boldsymbol{d}_i, \boldsymbol{d})\right) + b = 0$$

Οι πιο γνωστοί πυρήνες είναι οι εξής:

Γραμμικός: $K(\boldsymbol{u},\boldsymbol{v}) = \boldsymbol{u} \cdot \boldsymbol{v}$

Πολυωνυμικός : $K(\boldsymbol{u},\boldsymbol{v}) = (\gamma \boldsymbol{u} \cdot \boldsymbol{v} + r)^p, \gamma > 0$

Ακτινωτής συνάρτησης βάσης : $K(\boldsymbol{u},\boldsymbol{v}) = e^{-\gamma \|u-v\|^2}$

Σιγμοειδής : $K(\boldsymbol{u},\boldsymbol{v}) = tanh(\gamma \boldsymbol{u} \cdot \boldsymbol{v} + r)$

Αποτέλεσμα όλων αυτών είναι η δημιουργία μιας συνάρτησης στην οποία αν εισάγουμε μία δήλωση μπορούμε να προβλέψουμε αν είναι απάτη ή όχι. Πιο συγκεκριμένα. Για τη δήλωση d_1 έχουμε για παράδειγμα, τη συνάρτηση *:*

$$f(\boldsymbol{d}_1) = \left[\left(\sum_{j=1}^{n} \lambda_j \, c_j K(\boldsymbol{d}_i, \boldsymbol{d}_1)\right) + b\right]$$

Τότε αν το αποτέλεσμα αυτής είναι θετική τότε $c_i = 1$ άρα η δήλωση είναι γνήσια, ενώ αν είναι αρνητική τότε $c_i = 1$ άρα η δήλωση είναι απάτη.

Στην εξίσωση απόφασης αν έχουμε χρησιμοποιήσει γραμμικό πυρήνα τότε δημιουργούμε ένα γραμμικό *SVM* καθώς ο πυρήνας αυτός δεν κάνει καμία μετατροπή στην αρχική διάσταση των δηλώσεων.

ΚΕΦΑΛΑΙΟ 7

Υλοποίηση εφαρμογών Text Mining με κατάλληλα στατιστικά πακέτα

7.1. Εισαγωγή

Αφού τελειώσαμε με το θεωρητικό κομμάτι και ότι χρειάζεται να γνωρίζουμε γύρω από το Text Mining και τις μεθόδους ομαδοποίησης και πρόβλεψης στο παρών κεφάλαιο θα σας παρουσιάσουμε δύο εφαρμογές σε πραγματικά δεδομένα.

Στην πρώτη εφαρμογή παρουσιάζουμε πως μπορούμε να επεξεργαστούμε έναν όγκο αδόμητων δεδομένων από δηλώσεις ατυχημάτων. Καταρχάς επεξεργαζόμαστε τα δεδομένα σε κατάλληλη μορφή ώστε να μπορούν τα στατιστικά πακέτα να τα επεξεργαστούν. Στην συνέχεια προσπαθούμε να μικρύνουμε το όγκο των δεδομένων με κύριο μέλημα μας να μην να χαθούν σημαντικές πληροφορίες και τελικά δημιουργούμε 8 ομάδες με διαφορετικό περιεχόμενο η κάθε μία, προκειμένου όποιος επιθυμεί μία συγκεκριμένη πληροφορία να μην χρειάζεται να διαβάσει όλα τα κείμενα αλλά να ανατρέξει στην αντίστοιχη ομάδα που τον ενδιαφέρει.

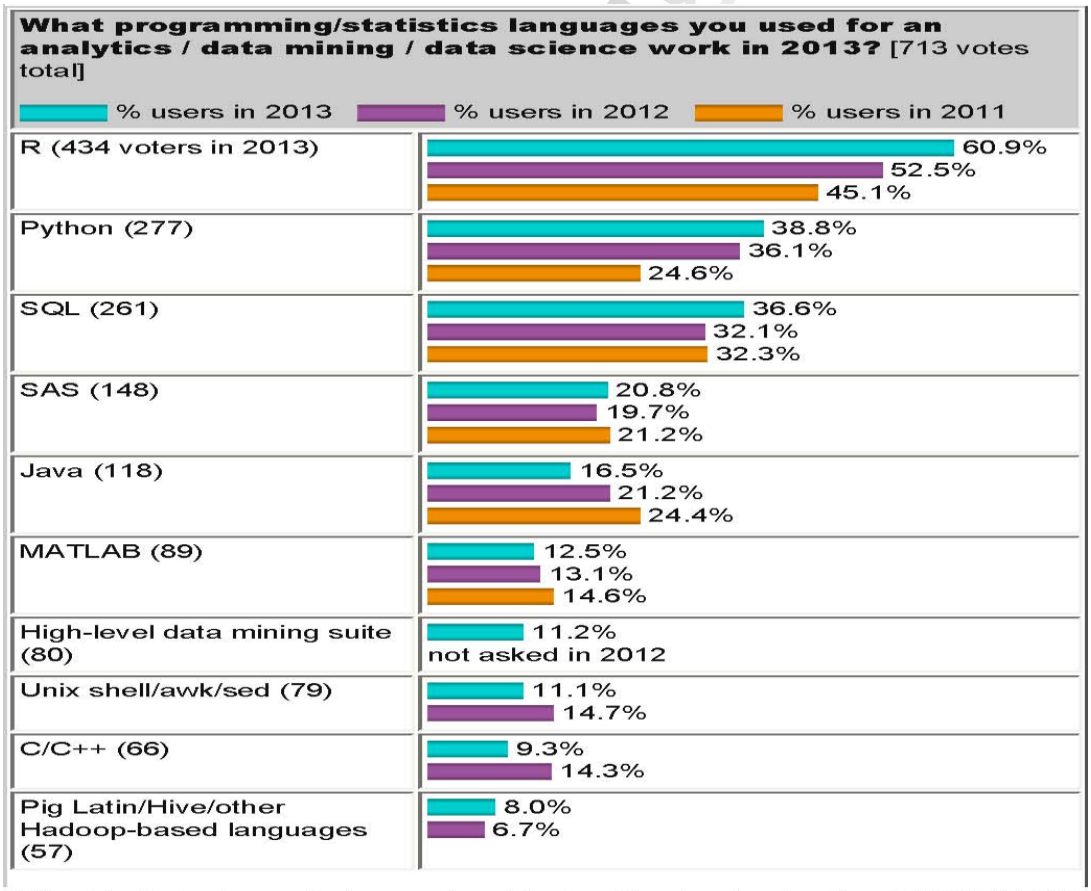
Στην δεύτερη εφαρμογή παρουσιάζουμε μία διαδικασία πρόβλεψης στην οποία παίρνουμε δύο διαφορετικές κατηγορίες κειμένων. Τις επεξεργαζόμαστε και τις χρησιμοποιούμε για την εκπαίδευση ενός μοντέλου που μας προβλέπει σε ποια κατηγορία από τις δύο ανήκει ένα νέο κείμενο.

7.2. Εφαρμογή 1: Ομαδοποίηση δεδομένων με την χρήση της γλώσσας R

Κύριος σκοπός αυτής της παραγράφου, είναι η προεπεξεργασία αδόμητων δεδομένων και η ομαδοποίηση τους. Για τον σκοπό αυτό θα χρησιμοποιήσουμε την R

Η R είναι μία γλώσσα προγραμματισμού,την οποία μπορεί οποιοσδήποτε να κατεβάσει ελεύθερα μέσω internet και ειδικεύεται σε στατιστικούς υπολογισμούς, σε γραφήματα και σε data και text mining.

Δημιουργήθηκε από τους Ross Ihaka και τον Robert Gentleman στο πανεπιστήμιο του Auckland στην Νέα Ζηλανδία και συνεχώς αναπτύσσεται από την R Development Core Team. Είναι μία γλώσσα με αυξημένους ρυθμούς ανάπτυξης με τους χρήστες της χρόνο με το χρόνο να πληθαίνουν. Ειδικά στο text mining αποτελεί την πλέον δημοφιλή γλώσσα προγραμματισμού τα τελευταία χρόνια όπως φαίνεται και στο παρακάτω σχήμα .



Οι δυνατότητες της R επεκτείνονται καθημερινά με τα πακέτα τα οποία δημιουργούνται με ταχείς ρυθμούς και επιτρέπουν σε κάθε χρήστη να χειρίζεται εξειδικευμένες στατιστικές τεχνικές, γραφήματα, εργαλεία αναφοράς κα. Αυτά τα

πακέτα αναπτύσσονται κυρίως στην R αλλά και σε κάποιες περιπτώσεις σε Java, C και Fortran.

Το σύνολο των πακέτων αυτήν την στιγμή είναι πάνω από 5300 και διατίθεται κυρίως από το CRAN (Comprehensive R Archive Network) το Bioconductor αλλά και από άλλα αποθετήρια, Ihaka, Ross (1998).

Η ιστοσελίδα http://cran.r-project.org/web/views/ παραθέτει ένα ευρύ φάσμα εφαρμογών και τα αντίστοιχα διαθέσιμα πακέτα οι οποίες μπορεί να είναι Οικονομικές ,Γενετική, Μηχανική Μάθηση, Κοινωνικές Επιστήμες, Χωρική Στατιστική, Ιατρική Απεικόνιση και High Performance Computing, Robert A. Muenchen (2012). Για την R έχει εντοπιστεί από την Food and Drug Administration (FDA) ότι είναι κατάλληλη για την ερμηνεία δεδομένων από κλινικές έρευνες.

Άλλες πηγές πακέτων για την R είναι η ιστοσελίδα http://crantastic.org/, η οποία αξιολογεί και αναθεωρεί όλα τα πακέτα CRAN καθώς επίσης και η ιστοσελίδα https://r-forge.r-project.org/ η οποία είναι μία κεντρική πλατφόρμα για την συνεργατική ανάπτυξη των R πακέτων, του R λογισμικού και των έργων .Περιλαμβάνει επίσης πολλές αδημοσίευτες εκδόσεις , πακέτα beta καθώς και ανεπτυγμένες εκδόσεις των πακέτων CRAN, Eddelbuettel, Dirk, Francois, Romain (2011).

Η ιστοσελίδα http://www.bioconductor.org/ παρέχει R πακέτα για την ανάλυση δεδομένων γονιδιωματικής όπως Affymetrix, χειρισμός δεδομένων και εργαλεία ανάλυσης cDNA μικροστοιχείων. Τελευταία έχει αρχίσει να παρέχει εργαλεία για την ανάλυση δεδομένων επόμενων γενεών με υψηλής απόδοσης μεθόδους προσδιορισμού αλληλουχίας.

Στο δικό μας παράδειγμα θα ασχοληθούμε κυρίως με το tm (text-mining) πακέτο το οποίο προσφέρει τεράστιες δυνατότητες στην εξόρυξη δεδομένων. Το συγκεκριμένο πακέτο είναι λειτουργικό για την διαχείριση εγγράφων κειμένου, αφαιρεί την διαδικασία χειραγώγησης κειμένου και διευκολύνει την 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. Αυτό συμβαίνει γιατί οι δομές δεδομένων και οι αλγόριθμοι μπορούν να επεκταθούν και να τροποποιηθούν έτσι ώστε να μπορούν να ανταποκριθούν στις εκάστοτε απαιτήσεις αφού έχουν σχεδιαστεί κατά τέτοιον ώστε να είναι εύκολη η

ενσωμάτωση νέων μορφών αρχείων, η ανάγνωση και η επεξεργασία τους, Feinerer (2008). Ας δούμε στην συνέχεια αναλυτικά τα βήματα.

i. Εισαγωγή κειμένου στην R

Στην δικιά μας εφαρμογή έχουμε μαζέψει τις δηλώσεις ατυχημάτων σε ένα αρχείο CSV και έχουμε εγκαταστήσει την R στον υπολογιστή μας μέσω της σελίδας

http://cran.r-project.org/bin/windows/base/

Αφού εγκαταστήσαμε και το πακέτο tm, το πρώτο βήμα ήταν να εισάγουμε τα δεδομένα μας στην πλατφόρμα της R με την εξής εντολή στην οποία αναφέρεται η μορφή του κειμένου που εισάγουμε και η τοποθεσία στην οποία βρίσκεται στον υπολογιστή μας :

txt.csv <- read.csv("C:/Users/arism/Desktop/Book1.csv").

Οι εντολές ανάγνωσης ποικίλουν ανάλογα με τη μορφή του αρχείου και για να δούμε όλες τις εντολές ανάγνωσης μπορούμε να δώσουμε την εντολή getReaders() όπως φαίνεται και στην παρακάτω εικόνα

```
> getReaders()
[1] "readDOC"                "readPDF"
[3] "readReut21578XML"       "readReut21578XMLasPlain"
[5] "readPlain"              "readRCV1"
[7] "readRCV1asPlain"        "readTabular"
[9] "readXML"
>
```

Η κύρια δομή για την επεξεργασία και την διαχείριση των εγγράφων στο tm είναι η λεγόμενη Corpus. Η Corpus είναι μία συλλογή κειμένων που συνήθως αποθηκεύονται με ηλεκτρονικά μέσα και από την οποία μπορούμε να πραγματοποιήσουμε την ανάλυση μας.

Μία Corpus μπορεί να είναι μία συλλογή, παραδείγματος χάριν άρθρα ειδήσεων (Reuters) ή δημοσιευμένα έργα του Σαίξπηρ.

Μέσα από κάθε Corpus μπορούμε να έχουμε χωριστά άρθρα, ιστορίες, δηλώσεις ασφαλιστικών ατυχημάτων κ.α. το καθένα σε επεξεργασία ως ξεχωριστή οντότητα, Feinerer and Hornik (2014).

Στην δημιουργία μιας Corpus είναι απαραίτητη η εντολή της ανάγνωσης του κειμένου που έχουμε ήδη εισαγάγει. Η εντολή ανάγνωσης διαφέρει ανάλογα με το είδος του αρχείου.

Και σε αυτήν την περίπτωση η R μας προσφέρει όλες τις δυνατές εντολές ανάγνωσης.

```
> getSources()
[1] "DataframeSource" "DirSource"       "ReutersSource"   "URISource"
[5] "VectorSource"
>
```

Στην δικιά μας περίπτωση λόγω του τύπου του αρχείου που χρησιμοποιούμε δίνουμε την εντολή:

txt <- Corpus(DataframeSource(txt.csv))

Όλη η παραπάνω διαδικασία γίνεται ουσιαστικά προκειμένου η γλώσσα προγραμματισμού μας να μπορεί να αναγνωρίσει τα κείμενα που θέλουμε να επεξεργαστούμε και αφού γίνει αυτό μπορεί να μας προσφέρει χρήσιμες πληροφορίες για τα δεδομένα μας.

ii. Εξερεύνηση της Corpus

Αφού δημιουργήσαμε λοιπόν το Corpus που θέλαμε, μπορούμε να χρησιμοποιήσουμε διάφορες εντολές που μας παρέχουν κάποια στοιχεία για τα δεδομένα μας

Κατ' αρχάς η εντολή summary() μας παρέχει μια ουσιαστικά περίληψη των δεδομένων μας.

```
> summary(txt)
A corpus with 116 text documents

The metadata consists of 2 tag-value pairs and a data frame
Available tags are:
  create_date creator
Available variables in the data frame are:
  MetaID
>
```

Τα metadata που αναφέρονται κατά την εκτέλεση της εντολής είναι πληροφορίες σχετικά με τα δεδομένα μας. Αυτή είναι ουσιαστικά μια περιγραφή που έχει να κάνει με τους τύπους των μεταβλητών, τις λειτουργίες και τις επιτρεπτές τιμές τους, David Walker, Thomas Scavo (2013).

Στη συνέχεια μπορούμε να δούμε όλες τις δηλώσεις ή τις πέντε πρώτες ή όποιες ακριβώς θέλουμε εμείς με την εντολή inspect(). Όπως φαίνεται στο επόμενο output εμείς έχουμε διαλέξει να μας δείξει τη δεύτερη την τρίτη και την τέταρτη δήλωση.

```
> inspect(txt[2:4])
A corpus with 3 text documents

The metadata consists of 2 tag-value pairs and a data frame
Available tags are:
  create_date creator
Available variables in the data frame are:
  MetaID

$`2`
CAT 345 traveling under a guide wire when the back of the boom
caught the wire. When the wire became taut it caused the power
pole to break and wire to snap.

$`3`
Insd was working and damaged buried service wires at a customer'
residence.

$`4`
I was driving along the motorway when the police pulled me over

> |
```

Η επιλογή inspect είναι μία πολλή χρήσιμη εντολή καθώς μας επιτρέπει να επιθεωρήσουμε ένα υποσύνολο του κειμένου μας, να εξακριβώσουμε την ποιότητά του και αν επιθυμούμε να το εκτυπώσουμε.

Τέλος μπορούμε να δούμε και μεμονωμένα τον αριθμό των δηλώσεων μας με την εντολή

```
> length(txt)
[1] 116
> |
```

iii. Προεπεξεργασία κειμένου

Το επόμενο βήμα του Text Mining είναι η προεπεξεργασία του κειμένου. Σύμφωνα με όσα έχουμε πει, σε αυτό το σημείο μετατρέπουμε τα αδόμητα δεδομένα σε δομημένα, χρήσιμα και ικανά να μας οδηγήσουν μέσω της ανάλυσης τους σε βοηθητικά συμπεράσματα για εμάς και κατά συνέπεια για την ασφαλιστική εταιρεία.

Τα βήματα της προεπεξεργασίας είναι τα εξής:

- Η μετατροπή όλων των γραμμάτων του κειμένου από κεφαλαία σε μικρά μέσω της εντολής

txt<-tm_map(txt,tolower)

```
> txt<-tm_map(txt,tolower)
> inspect(txt[2])
A corpus with 1 text document

The metadata consists of 2 tag-value pairs and a data frame
Available tags are:
  create_date creator
Available variables in the data frame are:
  MetaID

$`2`
cat 345 traveling under a guide wire when the back of the boo
caught the wire. when the wire became taut it caused the powe
pole to break and wire to snap.

> |
```

- Η διαγραφή των stopwords με την εντολή

txt<-tm_map(txt,removeWords,stopwords ('english'))

- Η διαγραφή όλων των αριθμών οι οποίοι γενικά μπορεί να είναι χρήσιμοι σε άλλου είδους εφαρμογές αλλά όχι στη δική μας :

txt<-tm_map(txt,removeNumbers)

- Η διαγραφή όλων των σημείων στίξης:

txt<-tm_map(txt,removePunctuation)

```
> txt<-tm_map(txt,removeNumbers)
> txt<-tm_map(txt,removePunctuation)
> inspect(txt[2])
A corpus with 1 text document

The metadata consists of 2 tag-value pairs and a data frame
Available tags are:
  create_date creator
Available variables in the data frame are:
  MetaID

$`2`
cat  traveling under a guide wire when the back of the boom
caught the wire when the wire became taut it caused the pow
pole to break and wire to snap
```

Η R παρέχει την δυνατότητα να απομακρύνεις από το κείμενό σου όποιες λέξεις ή όποια σύμβολα επιθυμείς:

newstopwords<-c("and ", "for ", "the ", "to ", "in ", "when ","then ","he ","she ","than ")

txt<-tm_map(txt,removeWords,newstopwords)

for(j in 1:length(txt))

txt[[j]]<-gsub("/","",txt[[j]])

Μπορούμε τέλος να απομακρύνουμε από όλες τις λέξεις, καταλήξεις όπως “ed”,” ing”, “es”, “s” επειδή στην ανάλυση των λέξεων ενός κειμένου μας ενδιαφέρει η βάση από την οποία προέρχεται η κάθε λέξη:

txt<-tm_map(txt,stemDocument)

Αφού τελειώσουμε με τον «καθαρισμό» του κειμένου ήρθε η ώρα να προβούμε στην ανάλυσή του. Αυτό θα γίνει μέσα από διαγράμματα και από πίνακες που μας παρέχει η R

iv. Ανάλυση κειμένου με πίνακες και διαγράμματα

Θα ξεκινήσουμε με έναν πίνακα ο οποίος είναι γνωστός ως Document by Term matrix (DTM)

Όπως φαίνεται και στην παρακάτω εικόνα ο πίνακας αυτός μας δείχνει την ακριβή συχνότητα της κάθε λέξης που χρησιμοποιείται συνολικά σε όλα μας τα δεδομένα, για κάθε δήλωση ξεχωριστά. Όπως είναι λογικό θα έχει πολλά μηδενικά αφού όλες οι δηλώσεις δεν γράφονται με τις ίδιες λέξεις αφού δεν γράφονται από τον ίδιο άνθρωπο και ο καθένας έχει διαφορετικό λεξιλόγιο και είδος γραφής.

```
> dtm<-DocumentTermMatrix(txt)
> m<-as.matrix(dtm)
> dtm
A document-term matrix (116 documents, 298 terms)

Non-/sparse entries: 715/33853
Sparsity           : 98%
Maximal term length: 10
Weighting          : term frequency (tf)
> m[1:10,1:13]
    Terms
Docs accid acciden admit alert along anoth appear appli approach
   1     0       0     0     0     0     0      0     0        0
   2     0       0     0     0     0     0      0     0        0
   3     0       0     0     0     0     0      0     0        0
   4     0       0     0     0     1     1      0     0        0
   5     0       0     0     0     0     0      0     0        0
   6     0       0     0     0     1     0      0     0        0
   7     1       0     0     0     0     0      0     0        0
   8     0       0     0     0     0     0      0     0        0
   9     0       0     0     0     0     0      0     0        0
```

Οι λέξεις είναι με αλφαβητική σειρά και οι δηλώσεις συμβολίζονται με αύξοντα αριθμό ανάλογα με την θέση που είχαν στο αρχείο που έγινε η πρώτη εισαγωγή.

Μας δίνεται επίσης η δυνατότητα να δούμε τον αριθμό των σειρών (rows) και των στηλών (columns) του παραπάνω πίνακα

```
> nrow(dtm);ncol(dtm)
[1] 116
[1] 341
>
```

Επιπλέον μπορούμε να δούμε λέξεις με ανώτερο και κατώτερο όριο συχνότητας π.χ. από 2 εώς 56 :

```
> findFreqTerms(dtm,2,56)[1:80]
 [1] "accid"     "alert"     "along"     "anoth"     "appear"
 [7] "asleep"    "attempt"   "avoid"     "away"      "back"
[13] "big"       "blame"     "bounc"     "bring"     "bumper"
[19] "came"      "car"       "caus"      "collid"    "collis"
[25] "corner"    "cow"       "crash"     "damag"     "day"
[31] "direct"    "ditch"     "doctor"    "dog"       "door"
[37] "drive"     "driver"    "driveway"  "drove"     "earli"
[43] "end"       "enough"    "ever"      "eye"       "face"
[49] "fellow"    "fire"      "first"     "fli"       "found"
[55] "front"     "gas"       "gave"      "gentleman" "give"
[61] "got"       "guy"       "happen"    "hard"      "hat"
```

Το μειονέκτημα του πίνακα συχνοτήτων είναι ότι έχει πολλά μηδενικά αφού οι πιο πολλοί όροι εμφανίζονται σε λίγες δηλώσεις, Feinerer, Hornik and Meyer (2008). Για αυτό τον λόγο θα χρησιμοποιήσουμε μία εντολή που μας επιτρέπει να μειώσουμε το μέγεθός του πίνακα χωρίς να χάσουμε κάποια σημαντική πληροφορία από αυτόν.

Η εντολή είναι η removeSparseTerms() και απ' ότι φαίνεται και στο σχήμα μειώνει τους όρους από 298 που ήταν πριν σε 69. Η παράμετρος 0,98 λέει να αφαιρεθούν όλοι οι όροι από το DTM με μηδενικά, στο 98 % των εγγράφων.
Η μείωση μπορεί να είναι μεγαλύτερη αν μειώσουμε αυτόν τον αριθμό π.χ. 0,94.

```
> dtm3<-removeSparseTerms(dtm,0.98)
> dtm3
A document-term matrix (116 documents, 69 terms)

Non-/sparse entries: 380/7624
Sparsity           : 95%
Maximal term length: 10
Weighting          : term frequency (tf)
```

Μπορούμε επίσης να δούμε απλά τις συχνότητες κάθε λέξης ανεξάρτητα με τις δηλώσεις που έχουμε, με την ακόλουθη εντολή

```
> v<-sort(colSums(m),decreasing=TRUE)
> v[1:20]
       car      accid     struck        way      drive        hit
        42         13         13         13         10         10
pedestrian       road       back      found      front       stop
         9          9          8          8          8          7
    vehicl    attempt     collid      drove        end        saw
         7          6          6          6          6          6
>
```

Επιπρόσθετα μπορούμε να δούμε τις συχνότητες για ορισμένες λέξεις για παράδειγμα για τις 14 λέξεις με την μεγαλύτερη συχνότητα στα δεδομένα μας.

```
> head(v,14)
       car      accid     struck        way      drive        hit
        42         13         13         13         10         10
pedestrian       road       back      found      front       stop
         9          9          8          8          8          7
>
```

Πολύ σημαντικό για την ανάλυση του κειμένου μας είναι και ο δείκτης συσχέτισης. Στο παράδειγμά μας θα βρούμε συσχετίσεις λέξεων με ελάχιστο δείκτη συσχέτισης 0,37 με την λέξη accid (ατύχημα). Όπως φαίνεται και παρακάτω η λέξη που εμφανίζεται πιο συχνά κοντά στην λέξη accid(ατύχημα) είναι η λέξη happen (συνέβη) το οποίο είναι άκρως λογικό.

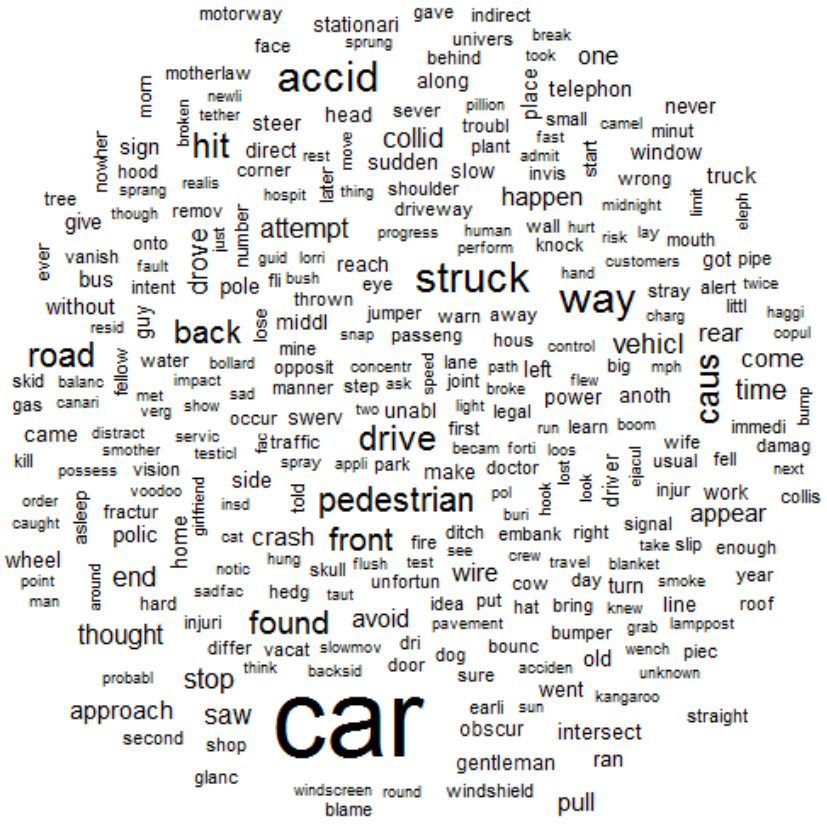
```
> myTdm<-TermDocumentMatrix(txt,control=list(minWordLength=1))
> findAssocs(myTdm,"accid",0.37)
         accid
happen    0.46
asleep    0.37
big       0.37
blame     0.37
bring     0.37
corner    0.37
doctor    0.37
door      0.37
fell      0.37
gave      0.37
indirect  0.37
joint     0.37
littl     0.37
mouth     0.37
occur     0.37
```

Η R μας δίνει επίσης τη δυνατότητα απεικόνισης των αποτελεσμάτων με πολλούς και διαφορετικούς τρόπους με την βοήθεια γραφημάτων

Στον παρακάτω κώδικα πρέπει πρώτα να μετατρέψουμε το term-document matrix σε normal matrix (απλό πίνακα) και στη συνέχεια να υπολογίσουμε τις συχνότητες των λέξεων. Στην συνέχεια είναι απαραίτητη η εγκατάσταση του πακέτου wordcloud. Με την εντολή wordcloud() οι δύο πρώτοι παράμετροι δίνουν μια λίστα των λέξεων και των συχνοτήτων. Λέξεις με συχνότητα μικρότερη του 1 δεν απεικονίζονται, με άλλα λόγια χρησιμοποιούμε όλες τις λέξεις (min.freq=1) κάτι το οποίο μπορούμε να αλλάξουμε, αν το επιθυμούμε

```
> words<-names(v)
> d<-data.frame(words=words,freq=v)
> wordcloud(d$word,d$freq,min.freq=1)
>
```

Το αποτέλεσμα αυτού είναι η δημιουργία ενός γραφήματος στο οποίο όσο μεγαλύτερη είναι η συχνότητα μιας λέξεως τόσο μεγαλύτερη παρουσιάζεται αυτή η λέξη. Όπως είναι εμφανές στο επόμενο σχήμα, η πιο συχνή λέξη είναι το car το οποίο επιβεβαιώνεται και από τους προηγούμενους πίνακες αφού εμφανίζεται 42 φορές συνολικά. Το γεγονός αυτό είναι πολύ λογικό αφού τα δεδομένα προέρχονται από δηλώσεις ατυχημάτων μιας ασφαλιστικής εταιρείας.

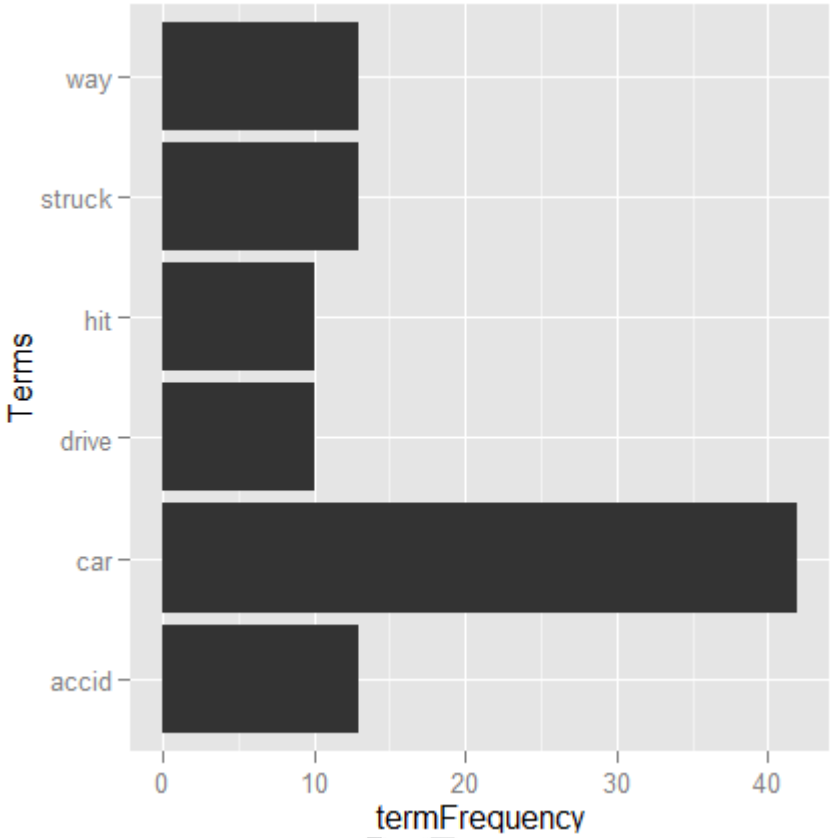


Το επόμενο γράφημα που θα χρησιμοποιήσουμε για να δείξουμε τις λέξεις με την μεγαλύτερη συχνότητα είναι το barplot.

Από τον πίνακα term-document matrix που δημιουργήσαμε πιο πάνω μπορούμε να αντλήσουμε τη συχνότητα των όρων με την εντολή rowSums(). Στην συνέχεια επιλέγουμε τους όρους οι οποίοι εμφανίζονται σε δέκα ή περισσότερες δηλώσεις και τις εμφανίζουμε γραφικά με το barplot από το πακέτο ggplot, Wickham (2009) .

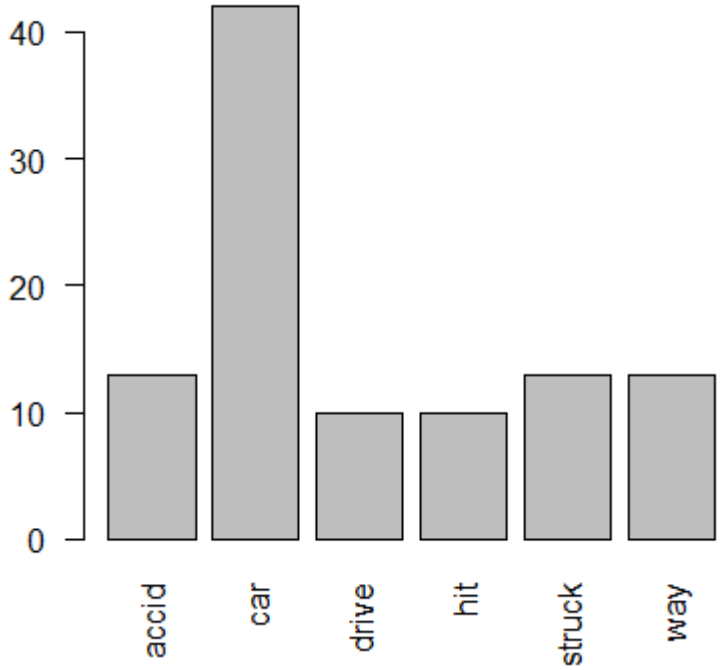
```
> termFrequency<-rowSums(as.matrix(myTdm))
> library(ggplot2)
> qplot(names(termFrequency),termFrequency,geom="bar",xlab="Terms"+coord_flit()
+ )
Error in qplot(names(termFrequency), termFrequency, geom = "bar", xlab = "Terms$
  could not find function "coord_flit"
> termFrequency<-rowSums(as.matrix(myTdm))
> termFrequency<-subset(termFrequency,termFrequency>=10)
> library(ggplot2)
> qplot(names(termFrequency),termFrequency,geom="bar",xlab="Terms")+coord_flip()
Mapping a variable to y and also using stat="bin".
  With stat="bin", it will attempt to set the y value to the count of cases in $
  This can result in unexpected behavior and will not be allowed in a future ve$
  If you want y to represent counts of cases, use stat="bin" and don't map a va$
  If you want y to represent values in the data, use stat="identity".
  See ?geom_bar for examples. (Deprecated; last used in version 0.9.2)
>
```

Το αποτέλεσμα είναι ένα γράφημα με μπάρες που μας δείχνουν ποιες λέξεις εμφανίζονται σε περισσότερα κείμενα και σε πόσα συγκεκριμένα



Υπάρχει και εναλλακτικός τρόπος απεικόνισης δηλαδή να μην είναι οριζόντιες οι μπάρες μας αλλά κάθετες:

```
> barplot(termFrequency,las=2)
>
```



v. Ομαδοποίηση δεδομένων

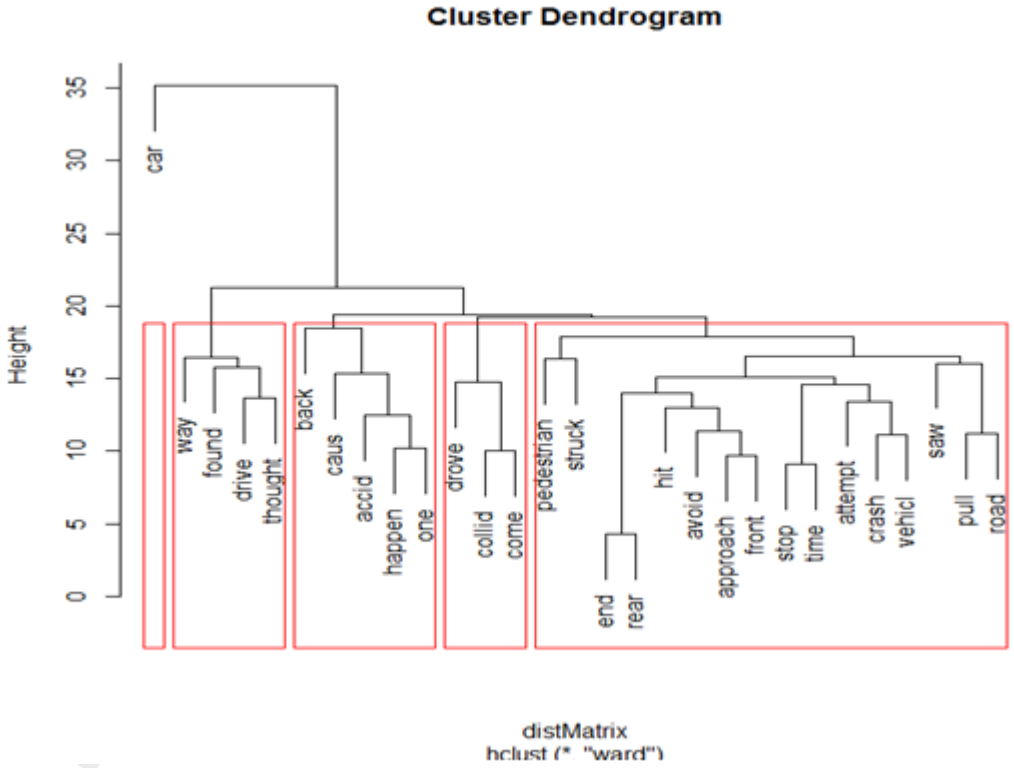
Έχοντας πραγματοποιήσει την προεπεξεργασία των δηλώσεων και μια ανάλυση των περιεχομένων τους σε αυτό το σημείο θα παρουσιάσουμε τρόπους με τους οποίους ομαδοποιούνται τα δεδομένα που έχουν το ίδιο αντικείμενο ή διαφορετικά αναφέρονται στο ίδιο θέμα.

Θα ξεκινήσουμε με την ομαδοποίηση των δεδομένων με την ιεραρχική ομαδοποίηση (hierarchical clustering). Η ιεραρχική ομαδοποίηση όπως προαναφέραμε είναι μία διαδικασία η οποία ξεκινά με πολλές ομάδες και στη συνέχεια προσπαθεί να ενώσει ομάδες με κοινά στοιχεία μέχρις ότου να καταλήξει σε μία ομάδα. Στο παράδειγμά μας οι αραιοί όροι έχουν αφαιρεθεί προκειμένου να μην είναι το διάγραμμα γεμάτο από λέξεις χωρίς ιδιαίτερη αναλυτική σημασία.

Στην συνέχεια υπολογίζονται, με την εντολή dist(), οι αποστάσεις μεταξύ των όρων μετά την κλιμάκωσή τους. Μετά από αυτό οι όροι ομαδοποιούνται με την hclust() και το δενδρόγραμμα «κόβεται» σε 5 συμπλέγματα. Η μέθοδος (method) έχει ορισθεί ως ward στην οποία πραγματοποείται αύξηση της διακύμανσης όταν οι δύο ομάδες συγχωνευθούν, Kamber (2000).

```
> myTdm2<-removeSparseTerms(myTdm,sparse=0.98)
> m2<-as.matrix(myTdm2)
> distMatrix<-dist(scale(m2))
> fit<-hclust(distMatrix,method="ward")
> plot(fit)
> rect.hclust(fit,k=5)
> (groups<-cutree(fit,k=5))
     accid      along      anoth      appear   approach    attempt
         1          2          2          2          2          3
      back       came        car       caus     collid       come
         3          1          4          3          5          5
    direct      drive     driver      drove        end      found
         2          2          1          5          2          3
 gentleman       give        got        guy     happen       head
         2          1          2          2          1          3
      home  intersect       left       make      middl      never
         5          2          2          2          2          1
```

Η R μέσω της ιεραρχικής διαδικασίας δίνει την δυνατότητα να παράγουμε και δενδρογράμματα ή δενδροειδείς απεικονίσεις. Στο επόμενο σχήμα δημιουργήθηκε ένα δενδρόγραμμα το οποίο χωρίστηκε σε 5 ομάδες και οι ομάδες διαχωρίζοντα μεταξύ τους από ένα κόκκινο πλαίσιο



Η επόμενη ομαδοποίηση που θα χρησιμοποιήσουμε είναι η ομαδοποίηση k-means. Πρώτο βήμα αυτής της ομαδοποίησης στην R είναι να μετατρέψουμε το term-document matrix σε document-term. Στη συνέχεια οι δηλώσεις ομαδοποιούνται

σε 5 ομάδες. Αμέσως μετά θα ελέγξουμε τις 3 πιο δημοφιλείς λέξεις κάθε ομάδας καθώς επίσης και τα κέντρα διασποράς τους. Είναι σημαντικό να ρυθμίσουμε την μεταβλητή set.seed() πριν την εκτέλεση του kmeans() έτσι ώστε το αποτέλεσμα να μπορεί να αναπαραχθεί ξανά με τα ίδια αποτελέσματα.

```
> m3<-t(m2)
> set.seed(122)
> k<-5
> kmeansResult<-kmeans(m3,k)
> round(kmeansResult$centers,digits=3)[1:5,1:10]
  accid along anoth appear approach attempt avoid  back came   car
1 0.800     0 0.000   0.00    0.000   0.000 0.000 0.000 0.40 0.400
2 0.000     0 0.000   0.50    0.125   0.000 0.125 0.125 0.00 0.500
3 0.092     0 0.010   0.01    0.041   0.061 0.041 0.071 0.02 0.347
4 0.000     1 0.667   0.00    0.000   0.000 0.000 0.000 0.00 0.667
5 0.000     0 0.000   0.00    0.000   0.000 0.000 0.000 0.00 0.000
> 
```

Για να να μπορέσουμε να κατανοήσουμε σε τι αναφέρεται η κάθε ομάδα θα ελέγξουμε, όπως προείπαμε, τις 3 πρώτες λέξεις της κάθε ομάδας:

```
> for(i in 1:k){
+ cat(paste("cluster",i,":",seq=""))
+ s<-sort(kmeansResult$centers[i,],decreasing=T)
+ cat(names(s)[1:3],"\n")
+ }
cluster 1 : car hit pedestrian
cluster 2 : car vehicl accid
cluster 3 :  way accid caus
cluster 4 : struck car end
cluster 5 : appear stop approach
> 
```

Όπως φαίνεται στην τελευταία εικόνα οι δηλώσεις μας έχουν χωρισθεί σε 5 ομάδες από τις οποίες έχουμε πάρει τις 3 λέξεις με τη μεγαλύτερη συχνότητα Παρατηρείστε ότι οι λέξεις της κάθε ομάδας βγάζουν κάποιο νόημα και μας βοηθάνε να διαπιστώσουμε σε τι αναφέρεται η κάθε ομάδα. Δηλαδή η πρώτη ομάδα μας βγάζει τις λέξεις «car hit pedestrian» στην οποία αν κάνουμε την μετάφραση «αυτοκίνητο χτύπησε πεζό» μπορούμε να προβούμε στο συμπέρασμα ότι αυτή η ομάδα αναφέρεται σε ατυχήματα με πεζούς. Αν προχωρήσουμε στην επόμενη οι πιο συχνές λέξεις είναι οι «car vehicl accid» στην οποία αν κάνουμε πάλι την μετάφραση λέει «αυτοκίνητο όχημα ατύχημα». Έτσι μπορούμε να καταλήξουμε στο συμπέρασμα ότι αναφέρεται σε ατυχήματα μεταξύ οχημάτων. Η τρίτη ομάδα έχει σαν «κορυφαίες» λέξεις τις «way accid caus» δηλαδή «τρόπος ατύχημα αιτία».Και εδώ εύκολα

μπορούμε να καταλάβουμε ότι η συγκεκριμένη ομάδα αναφέρεται κυρίως σε αίτια ατυχημάτων.

Σύμφωνα με όσα προείπαμε γίνεται αντιληπτό πόσο χρήσιμη μπορεί να είναι μια εφαρμογή Text Mining ομαδοποίησης σε μια ασφαλιστική καθώς θα μπορεί σε σύντομο χρόνο να ομαδοποιήσει τεράστιο όγκο αδόμητων δεδομένων. Με αυτόν τον τρόπο θα κατανοήσει το νόημα των κειμένων, θα μπορέσει να τα αρχειοθετήσει και να τα έχει ανά πάσα στιγμή χωρισμένα και έτοιμα για ανάλυση ή για οποιοδήποτε άλλο σκοπό.

7.3. Εφαρμογή 2: Εκπαίδευση μοντέλου πρόβλεψης με το πρόγραμμα Rapid Miner

i. Rapid Miner

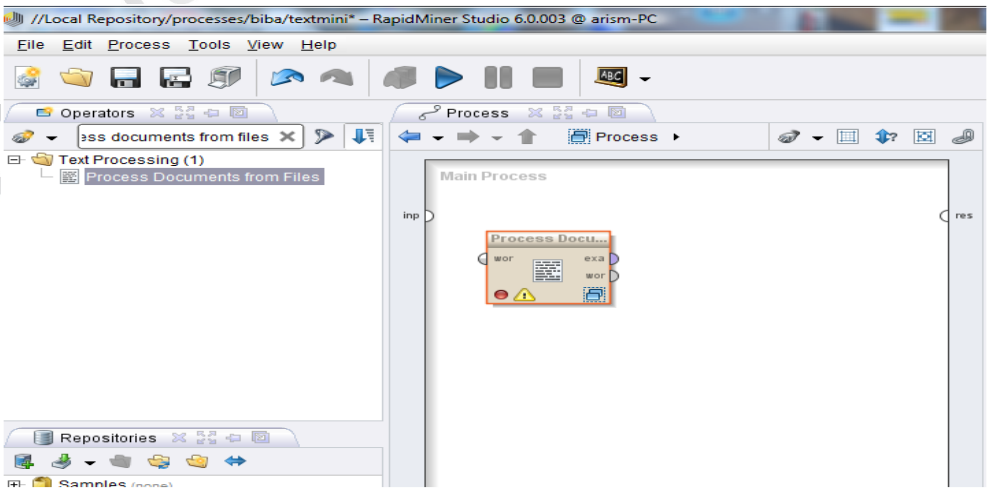
Σε αυτήν την ενότητα της εργασίας θα παρουσιάσουμε μία εφαρμογή την οποία μπορεί να χρησιμοποιήσει κάθε ασφαλιστική εταιρεία προκειμένου να δει ποιες δηλώσεις είναι επικίνδυνες να αποδειχθούν απάτη. Με αυτόν τον τρόπο πρώτον θα αντιμετωπίσει με περισσότερη καχυποψία τις εν λόγω δηλώσεις και κατά δεύτερον θα αναμένει πιθανώς υψηλές απαιτήσεις γεγονός που θα την βοηθήσει στον πιο ακριβή υπολογισμό των Κεφαλαιακών Αποθεμάτων της. Λόγω έλλειψης ποιοτικών δεδομένων από δηλώσεις ζημιών η εφαρμογή θα χρησιμοποιεί δεδομένα από ταινίες. Πιο συγκεκριμένα θα επιδείξουμε ένα μοντέλο το οποίο εκτιμάει αν μία κινηματογραφική κριτική είναι θετική ή αρνητική. Όπως θα φανεί και στην συνέχεια το μόνο που έχει να κάνει ο ενδιαφερόμενος από τον ασφαλιστικό χώρο είναι να αντικαταστήσει τα αρχεία που χρησιμοποιούμε (τα οποία περιέχουν τις καλές και τις κακές κριτικές ταινιών) με δηλώσεις που αποδείχθηκαν απάτη και δηλώσεις που δεν αποδείχθηκαν.

Για αυτήν την εφαρμογή θα χρησιμοποιήσουμε ως εργαλείο το Rapid Miner το οποίο όπως και στην περίπτωση της R εγκαθίσταται ελεύθερα μέσω internet.Το RapidMiner είναι μια πλατφόρμα λογισμικού που αναπτύχθηκε από την εταιρεία με το ίδιο όνομα και παρέχει ένα ολοκληρωμένο περιβάλλον για μηχανική μάθηση, εξόρυξη δεδομένων, κειμένων, predictive analytics και business analytics. Χρησιμοποιείται για επιχειρήσεις και για βιομηχανικές εφαρμογές, καθώς και για την

έρευνα, την εκπαίδευση, την κατάρτιση, την ταχεία προτυποποίηση και ανάπτυξη εφαρμογών και υποστηρίζει όλα τα στάδια της διαδικασίας εξόρυξης δεδομένων, συμπεριλαμβανομένων των αποτελεσμάτων απεικόνισης, την επικύρωση και τη βελτιστοποίηση. Το RapidMiner, ήταν παλαιότερα γνωστό ως YALE (Yet Another Learning Environment). Αναπτύχθηκε το 2001 από τους Ralf Klinkenberg, Ingo Mierswa και Simon Fischer στη Μονάδα Τεχνητής Νοημοσύνης του Πανεπιστημίου του Dortmund. Ξεκινώντας το 2006, την ανάπτυξή του ανέλαβε η Rapid-I, μια εταιρεία που ιδρύθηκε από τον Ingo Mierswa και Ralf Klinkenberg κατά το ίδιο έτος. Το 2007, το όνομα του λογισμικού άλλαξε από YALE σε RapidMiner, Markus Hofmann, Ralf Klinkenberg (2013) .Το Rapid Miner είναι μέσα στα δημοφιλέστερα προγράμματα που χρησιμοποιούνται για text mining όπως φαίνεται και στο παρακάτω σχήμα. Εμφανής είναι ότι τους τελευταίου 12 μήνες είναι το νούμερο ένα στις προτιμήσεις των χρηστών για text mining. Ο πίνακας δείχνει ακόμα μία ισορροπία μεταξύ των ελεύθερων μέσω internet εργαλείων και των εμπορικών γεγονός που δείχνει πόσο πολύ έχει ανέβει η ποιότητα των πρώτων καθώς τα προηγούμενα χρόνια το ποσοστό υπέρ των εμπορικών ήταν συντριπτική. Τέλος με τον όρο alone εννοούμε πόσοι από τους χρήστες χρησιμοποιούν μόνο αυτό το εργαλείο χωρίς την υποστήριξη παρεμφερών εφαρμογών. Τα αποτελέσματα αυτά προέρχονται από μία δημοσκόπηση της εταιρείας KDnuggets Software Poll.

What Analytics, Big Data, Data mining, Data Science software you used in the past 12 months for a real project?[1880 voters]	
Legend: Red: Free/Open Source tools Green: Commercial tools	% users in 2013 % users in 2012
Rapid-I RapidMiner/RapidAnalytics free edition(737), 30.9% alone	39.2% 26.7%
R(704), 6.5% alone	37.4% 30.7%
Excel(527), 0.9% alone	28.0% 29.8%
Weka / Pentaho(269), 5.6% alone	14.3% 14.8%
Python with any of numpy/scipy/pandas/iPython... packages(250), 0% alone	13.3% 14.9%
Rapid-I RapidAnalytics/RapidMiner Commercial Edition(225), 52.4% alone	12.0%
SAS(202), 2.0% alone	10.7% 12.7%
MATLAB(186), 1.6% alone	9.9% 10.0%
StatSoft Statistica(170), 45.9% alone	9.0% 14.0%
IBM SPSS Statistics(164), 1.8% alone	8.7% 7.8%

Αφού εγκαταστήσουμε το Rapid Miner μέσω της ιστοσελίδας http://rapidminer.com/products/rapidminer-studio/ το πρώτο πράγμα που έχουμε να κάνουμε προκειμένου να προβούμε σε ανάλυση Text Mining είναι να εγκαταστήσουμε το Text Mining Extension. Έτσι λοιπόν από την επιφάνεια εργασίας θα επιλέξουμε το Help Updates and Extensions (Marketplace) – Top Downloads. Επιλέγουμε Text Mining Extension και Install. Αμέσως μετά θα πάμε στους operators και θα διαλέξουμε το Process Documents from Files, David Norris (2013).

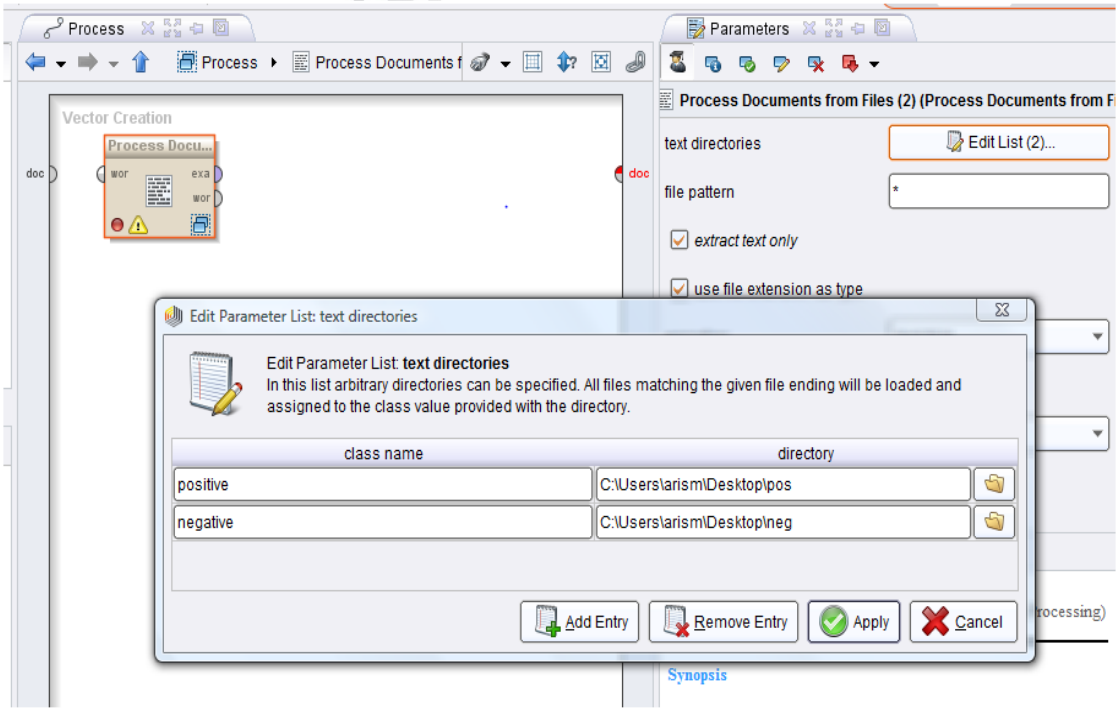


Τα δεδομένα μας για την εφαρμογή έχουν συγκεντωθεί και έχουν αποθηκευθεί σε ένα σημείο στον σκληρό μας δίσκο. Αποτελούνται από 1000 θετικά και 1000 αρνητικά σχόλια για κινηματογραφικές ταινίες Στο μόνο σημείο που θα άλλαζε αυτή η διαδικασία αν την εφαρμόζαμε στον ασφαλιστικό τομέα θα ήταν να αποθηκεύαμε 1000 δηλώσεις που αποδείχθηκαν απάτη και 1000 που δεν αποδείχθηκαν. Σε κανένα άλλο σημείο δεν θα κάνουμε στην συνέχεια διαφορετικό από αυτό που θα κάναμε αν είχαμε δηλώσεις.

ii. Εισαγωγή δεδομένων

Το πρώτο μας βήμα είναι να εισάγουμε τα δεδομένα μας στην πλατφόρμα του Rapid Miner σε δύο χωριστούς φακέλους οι οποίοι περιέχουν ο ένας τα θετικά και ο άλλος τα αρνητικά. Έτσι η ανάλυση που θα γίνει στα μεν και στα δε θα είναι διαφορετική.

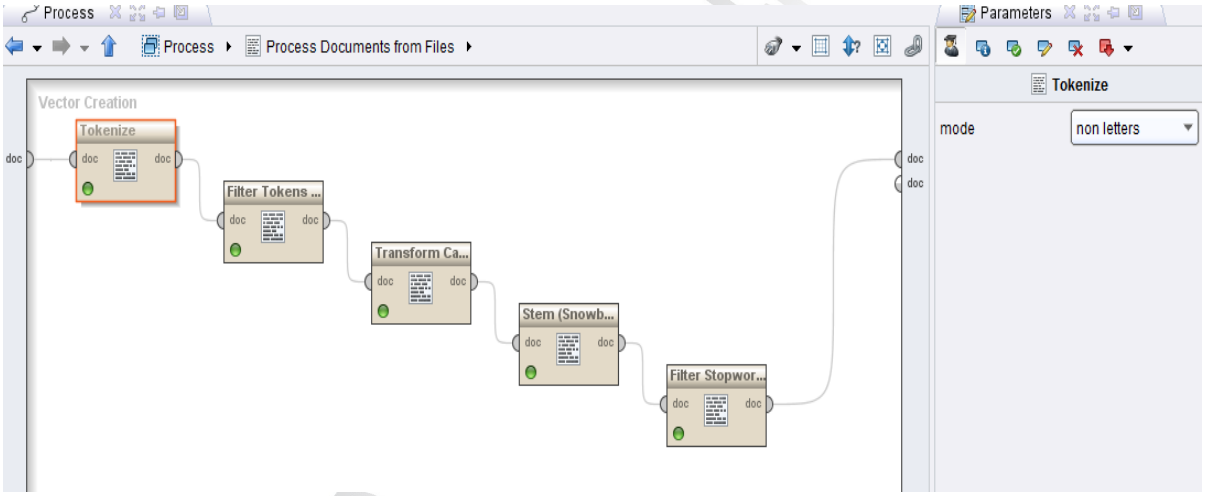
Επιλέγουμε λοιπόν το Process Documents from Files και στο δεξί μέρος υπάρχει η επιλογή Edit List απ' όπου μπορούμε με ευκολία να εισάγουμε τα δεδομένα μας. Άπλα τα βρίσκουμε από το σημείο που τα έχουμε αποθηκεύσει στον υπολογιστή μας και τους δίνουμε όνομα (positive=θετικά, negative= αρνητικά) .



iii. Προεπεξεργασία Δεδομένων

Το επόμενο μας βήμα είναι η προεπεξεργασία του κειμένου. Αυτό το βήμα θα γίνει μέσω του Process Document from Files. Επιλέγοντάς το μας ανοίγει ένα νέο φύλλο εργασίας στο οποίο θα γίνει όλος ο «καθαρισμός» των κειμένων.

Πρώτα από όλα με την εφαρμογή Tokenize θα χωρίσουμε τα κείμενα σε λέξεις και φράσεις. Στην συνέχεια θα χρησιμοποιήσουμε το Filter Tokens (by length) προκειμένου όλες οι λέξεις να έχουν ένα συγκεκριμένο εύρος στο μήκος τους το οποίο θα είναι μεταξύ 4 έως 25 χαρακτήρων. Αμέσως μετά εισάγουμε το Transform Cases το οποίο μετατρέπει όλα τα κεφαλαία γράμματα σε μικρά και ύστερα με το Stem(Snowball) αφαιρούμε από όλες τις λέξεις μας τις καταλήξεις όπως ed, ing, en



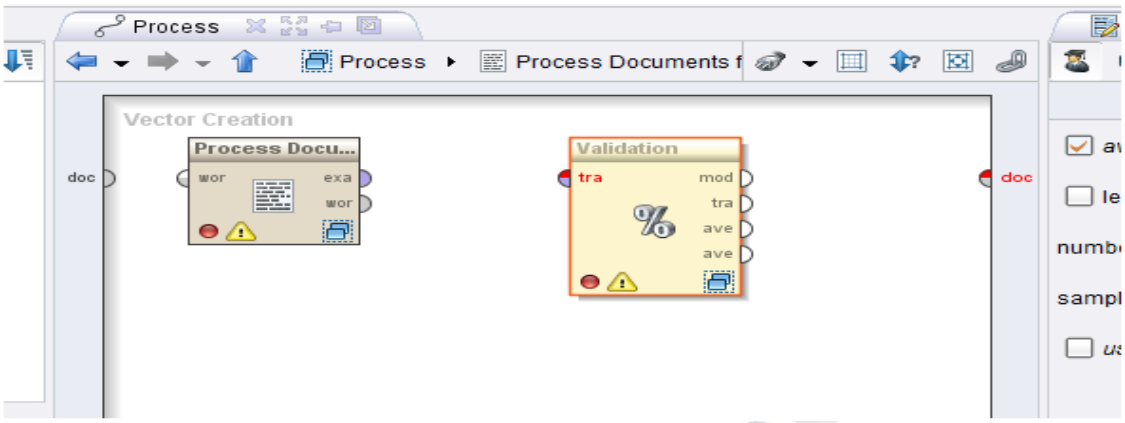
προκειμένου να μείνει μόνο η βάση από την οποία προέρχεται η κάθε λέξη. Τέλος θα αφαιρέσουμε από τα κείμενα λέξεις όπως and , a, the κλπ. Ουσιαστικά όλη η διαδικασία είναι σχεδόν πανομοιότυπη με αυτήν που χρησιμοποιήσαμε προηγουμένως στο παράδειγμά μας με την R.

Αφού ολοκληρώσουμε αυτά τα βήματα το μόνο που μας έχει μείνει είναι να ενώσουμε τους τελεστές μεταξύ τους όπως φαίνεται στο σχήμα.

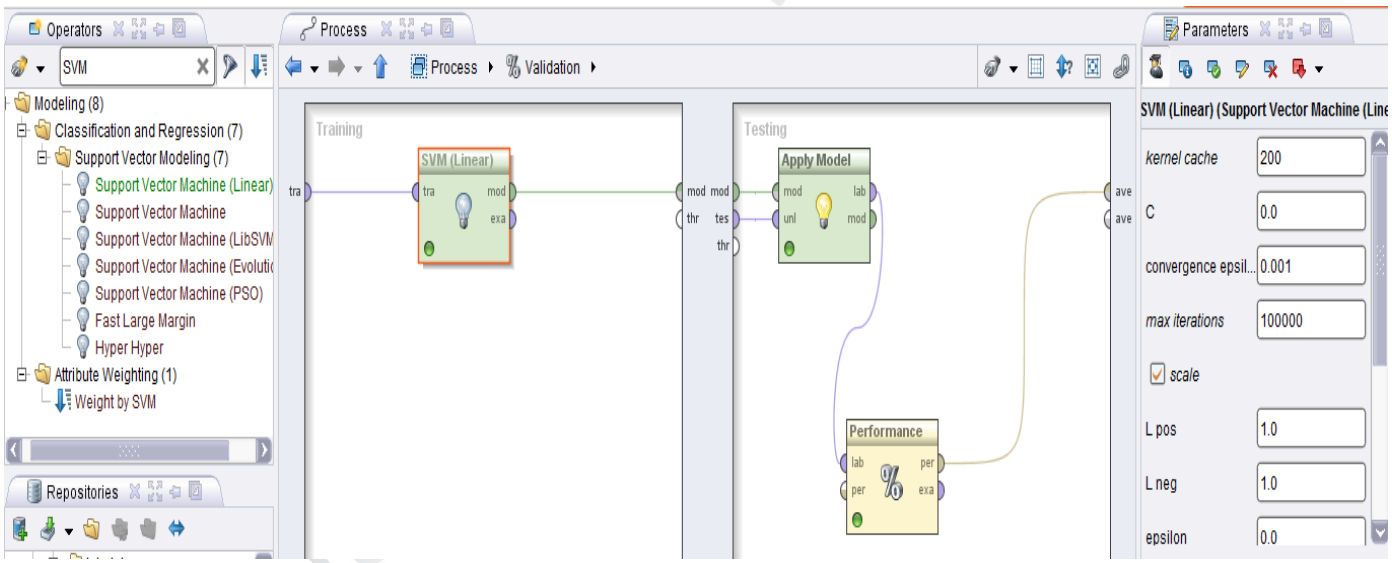
iv. Εκπαίδευση και Αξιολόγηση του Μοντέλου

Επιστρέφοντας στην αρχική επιφάνεια εργασία μας θα προσθέσουμε έναν ακόμα τελεστή ο οποίος θα δημιουργήσει ουσιαστικά το μοντέλο που θα κρίνει αν μία κριτική είναι λάθος ή όχι (μία δήλωση είναι απάτη) αλλά συγχρόνως θα ελέγχει

και θα προβλέπει την αξιοπιστία του μοντέλου. Όλα αυτά θα γίνουν με το % X-Validation, Evan Quinn (2013).



Επιλέγοντάς τον όπως και πριν μας μεταφέρει σε μία νέα επιφάνεια εργασίας η οποία έχει δύο κομμάτια. Το πρώτο είναι το Training και το άλλο είναι το Testing δηλαδή

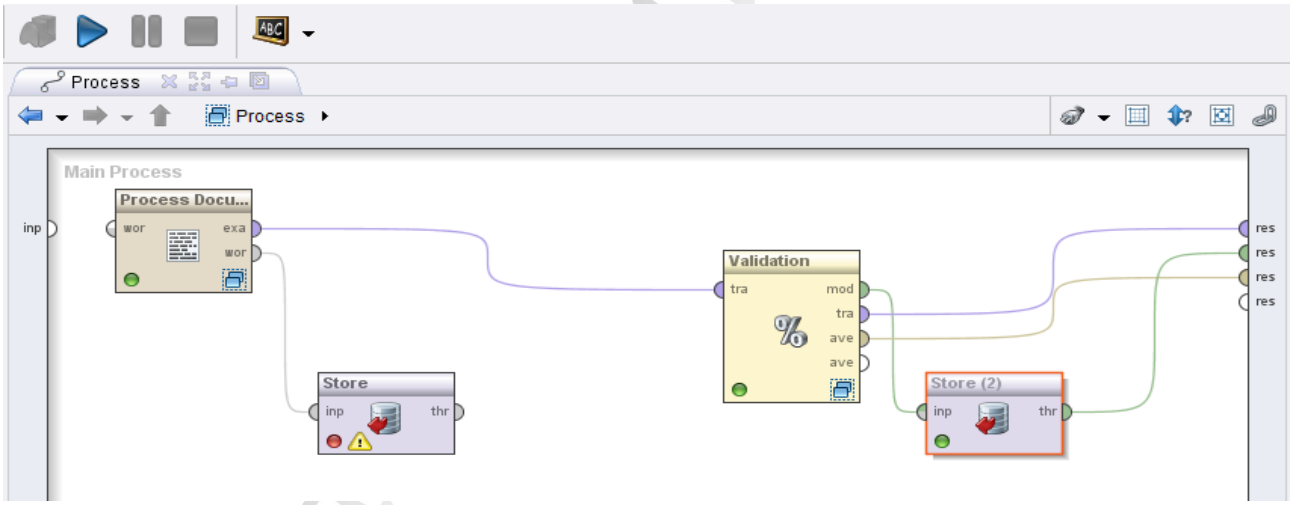


είναι οι δύο εργασίες που προείπαμε .Στο Training το Rapid Miner «εκπαιδεύει» το μοντέλο για να είναι σε θέση να προβλέψει ότι του ζητήσουμε και στο Testing ελέγχει την αξιοπιστία του μοντέλου που μόλις δημιουργήσαμε.

Στο μέρος Training θα προσθέσουμε τον τελεστή Support Vector Machine – SVM(Linear) ο οποίος θα δημιουργήσει ένα μοντέλο με τη μέθοδο που περιγράψαμε στο προηγούμενο κεφάλαιο. Στο μέρος του Testing θα βάλουμε το Apply Model ο οποίος είναι ένας τελεστής που δέχεται δύο εισόδους. Η μία είσοδος

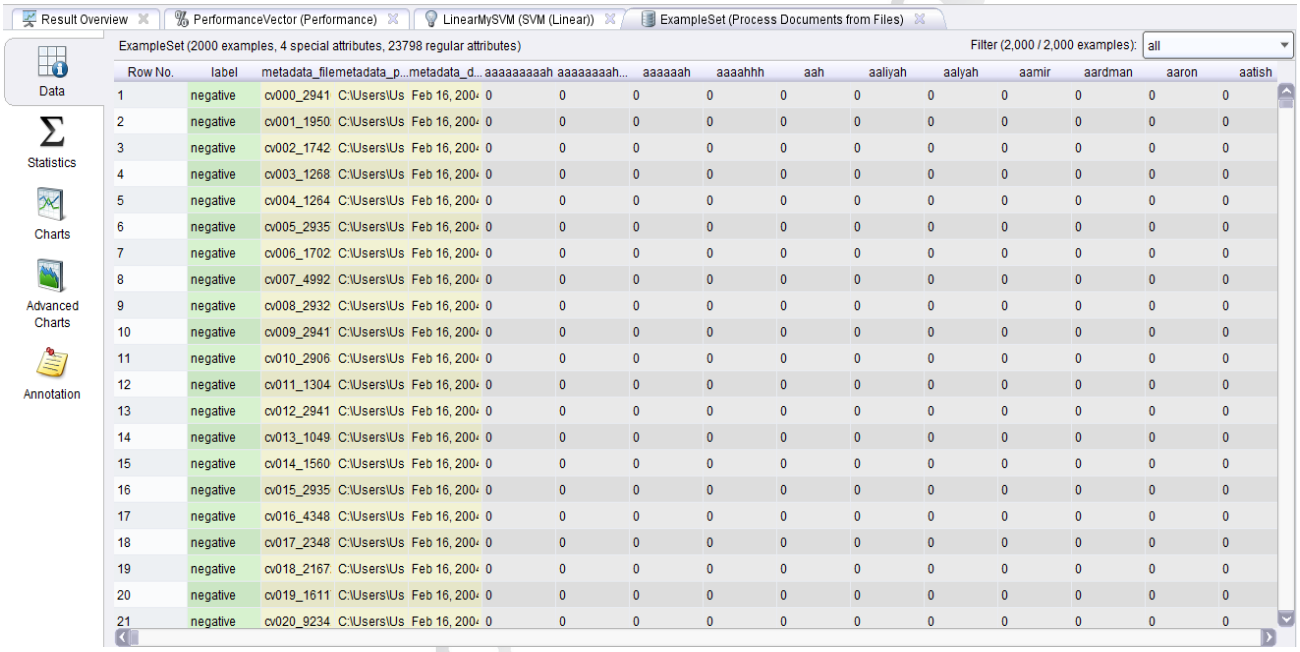
αφορά το μοντέλο και η δεύτερη αφορά τα δεδομένα βάσει των οποίων υπολογίστηκε το μοντέλο για έλεγχο, Ajay Ohri (2011). Τέλος προσθέτουμε και τη λειτουργία Performance η οποία θα κάνει τον έλεγχο αξιοπιστίας του μοντέλου. Το μόνο που μας έχει μείνει είναι όπως και προηγουμένως να ενώσουμε τις ροές μεταξύ τους.

Επόμενο βήμα της διαδικασίας είναι να προσθέσουμε δύο τελεστές Store οι οποίοι χρησιμοποιούνται κυρίως για την αποθήκευση των αποτελεσμάτων μας. Ο πρώτος θα αποθηκεύσει τις λέξεις που χρησιμοποιούνται στα αρνητικά και θετικά σχόλια (δηλώσεις που αποδείχθηκαν απάτη και δηλώσεις που δεν αποδείχθηκαν) καθώς και τις συσχετίσεις μεταξύ των λέξεων. Ο δεύτερος θα αποθηκεύσει το μοντέλο που θέλουμε να δημιουργήσουμε προκειμένου να προβούμε σε πρόβλεψη. Ενώνοντας και τις ροές θα ξεκινήσει η εφαρμογή όπως φαίνεται στο παρακάτω σχήμα.



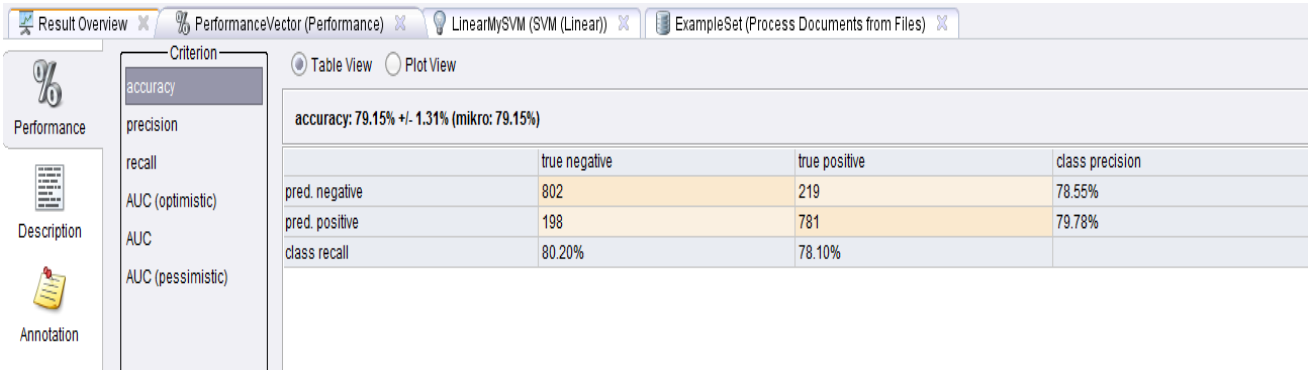
v. Αποτελέσματα

Τα αποτελέσματα αφού πατήσουμε το play θα είναι η δημιουργία κάποιων φακέλων. Ο πρώτος αποτελείται από έναν πίνακα με τις συχνότητες κάθε λέξης ξεχωριστά σε κάθε σχόλιο όπως ο πίνακας DTM που είδαμε στην εφαρμογή της R. Υπάρχει βέβαια ξεχωριστός πίνακας για τα θετικά και ξεχωριστός για τα αρνητικά σχόλια.



Row No.	label	metadata_file	metadata_p...	metadata_d...	aaaaaaaaah	aaaaaaaah...	aaaaaah	aaaahhh	aah	aaliyah	aalyah	aamir	aardman	aaron	aatish
1	negative	cv000_2941	C:\Users\Us	Feb 16, 200-	0	0	0	0	0	0	0	0	0	0	0
2	negative	cv001_1950	C:\Users\Us	Feb 16, 200-	0	0	0	0	0	0	0	0	0	0	0
3	negative	cv002_1742	C:\Users\Us	Feb 16, 200-	0	0	0	0	0	0	0	0	0	0	0
4	negative	cv003_1268	C:\Users\Us	Feb 16, 200-	0	0	0	0	0	0	0	0	0	0	0
5	negative	cv004_1264	C:\Users\Us	Feb 16, 200-	0	0	0	0	0	0	0	0	0	0	0
6	negative	cv005_2935	C:\Users\Us	Feb 16, 200-	0	0	0	0	0	0	0	0	0	0	0
7	negative	cv006_1702	C:\Users\Us	Feb 16, 200-	0	0	0	0	0	0	0	0	0	0	0
8	negative	cv007_4992	C:\Users\Us	Feb 16, 200-	0	0	0	0	0	0	0	0	0	0	0
9	negative	cv008_2932	C:\Users\Us	Feb 16, 200-	0	0	0	0	0	0	0	0	0	0	0
10	negative	cv009_2941	C:\Users\Us	Feb 16, 200-	0	0	0	0	0	0	0	0	0	0	0
11	negative	cv010_2906	C:\Users\Us	Feb 16, 200-	0	0	0	0	0	0	0	0	0	0	0
12	negative	cv011_1304	C:\Users\Us	Feb 16, 200-	0	0	0	0	0	0	0	0	0	0	0
13	negative	cv012_2941	C:\Users\Us	Feb 16, 200-	0	0	0	0	0	0	0	0	0	0	0
14	negative	cv013_1049	C:\Users\Us	Feb 16, 200-	0	0	0	0	0	0	0	0	0	0	0
15	negative	cv014_1560	C:\Users\Us	Feb 16, 200-	0	0	0	0	0	0	0	0	0	0	0
16	negative	cv015_2935	C:\Users\Us	Feb 16, 200-	0	0	0	0	0	0	0	0	0	0	0
17	negative	cv016_4348	C:\Users\Us	Feb 16, 200-	0	0	0	0	0	0	0	0	0	0	0
18	negative	cv017_2348	C:\Users\Us	Feb 16, 200-	0	0	0	0	0	0	0	0	0	0	0
19	negative	cv018_2167	C:\Users\Us	Feb 16, 200-	0	0	0	0	0	0	0	0	0	0	0
20	negative	cv019_1611	C:\Users\Us	Feb 16, 200-	0	0	0	0	0	0	0	0	0	0	0
21	negative	cv020_9234	C:\Users\Us	Feb 16, 200-	0	0	0	0	0	0	0	0	0	0	0

Ένας άλλος φάκελος μας δείχνει την αξιοπιστία του μοντέλου. Αυτό που ουσιαστικά κάνει το Rapid Miner είναι αφού δημιουργήσει το μοντέλο πάει και το δοκιμάζει στα ήδη υπάρχοντα δεδομένα τα οποία είναι χωρισμένα σε θετικά και αρνητικά προκειμένου να βρει το ποσοστό ακρίβειας του. Έτσι λοιπόν όπως φαίνεται και στο παρακάτω σχήμα το ποσοστό ακρίβειας είναι κάτι περισσότερο από ικανοποιητικό αφού και στην περίπτωση του μοντέλου για τα αρνητικά σχόλια και στην περίπτωση του μοντέλου για τα θετικά σχόλια κυμαίνεται γύρω στο 80%

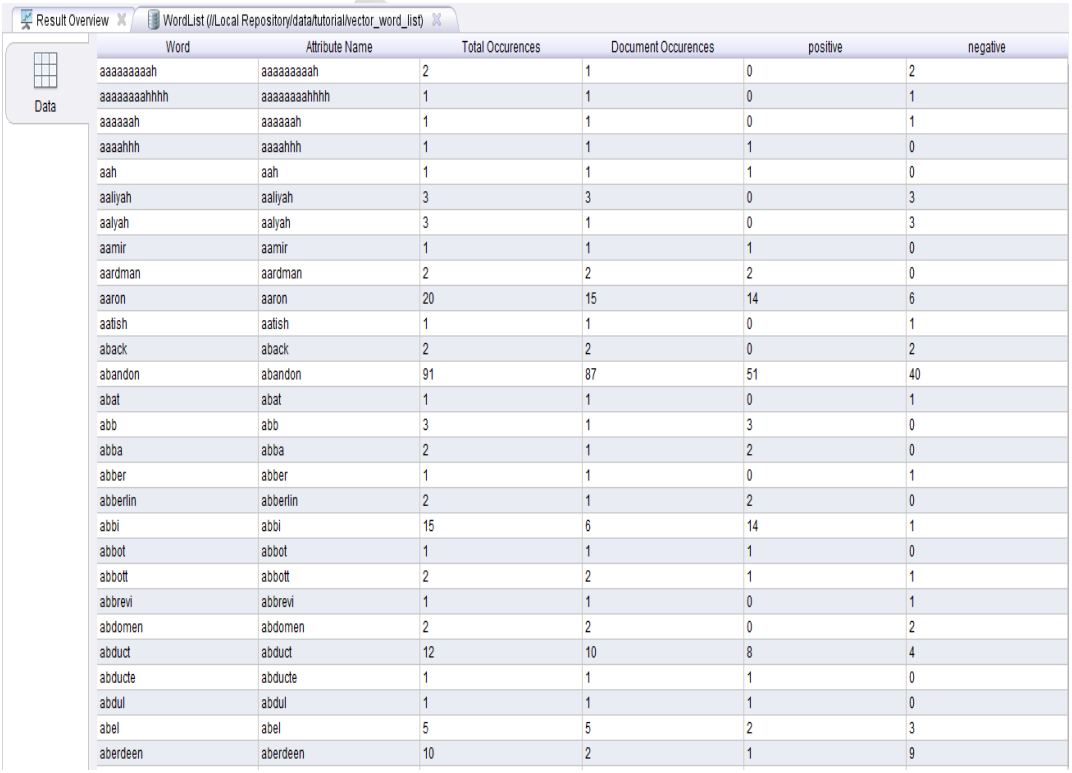


	true negative	true positive	class precision
pred. negative	802	219	78.55%
pred. positive	198	781	79.78%
class recall	80.20%	78.10%	

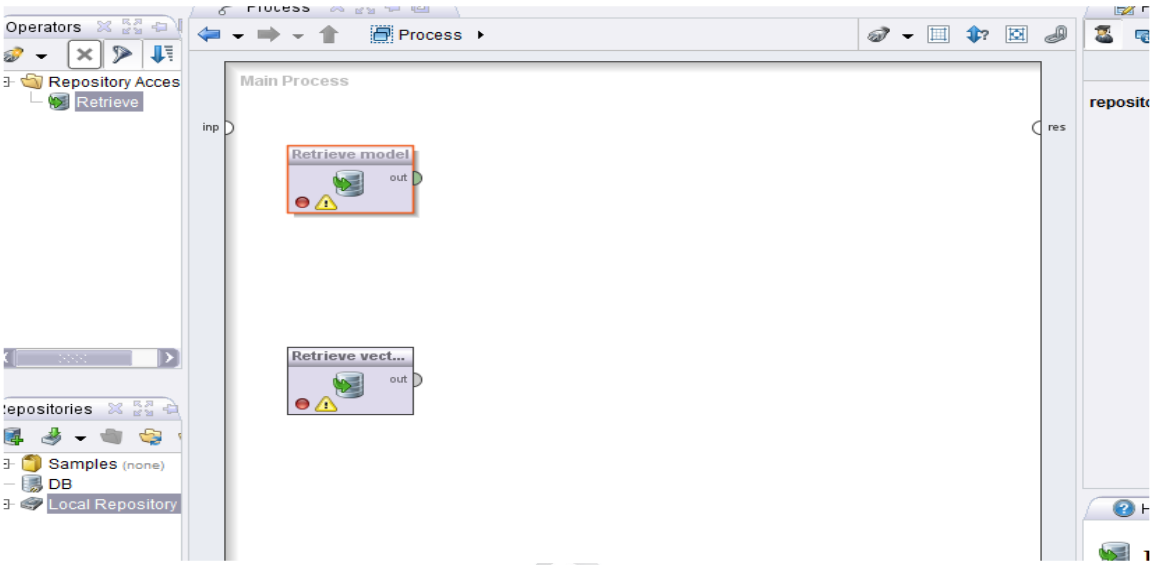
vi. Δοκιμή του μοντέλου σε νέα κείμενα

Μπαίνοντας στο τελικό στάδιο αυτό που μας μένει είναι να δοκιμάσουμε το μοντέλο μας σε νέα δεδομένα και συγκεκριμένα σε 6 σχόλια μιας ταινίας προκειμένου να δούμε αν όντως είναι επιτυχημένο ή όχι. Για τον σκοπό αυτό είναι απαραίτητη η δημιουργία άλλης μίας ροής στο Rapid Miner.

Σε αυτήν την ροή θα εισάγουμε τα δύο αρχεία που δημιουργήθηκαν από την προηγούμενη διαδικασία και τα οποία έχουν αποθηκευθεί στο αριστερό μέρος της οθόνης στα αποθετήρια (Repositories). Το πρώτο είναι το μοντέλο μας με την ονομασία model και το δεύτερο είναι το αρχείο vector_word_list το οποίο περιλαμβάνει έναν πίνακα με τις συχνότητες των λέξεων συνολικά και ξεχωριστά ανάλογα με το αν εμφανίζονται στα θετικά ή στα αρνητικά σχόλια.

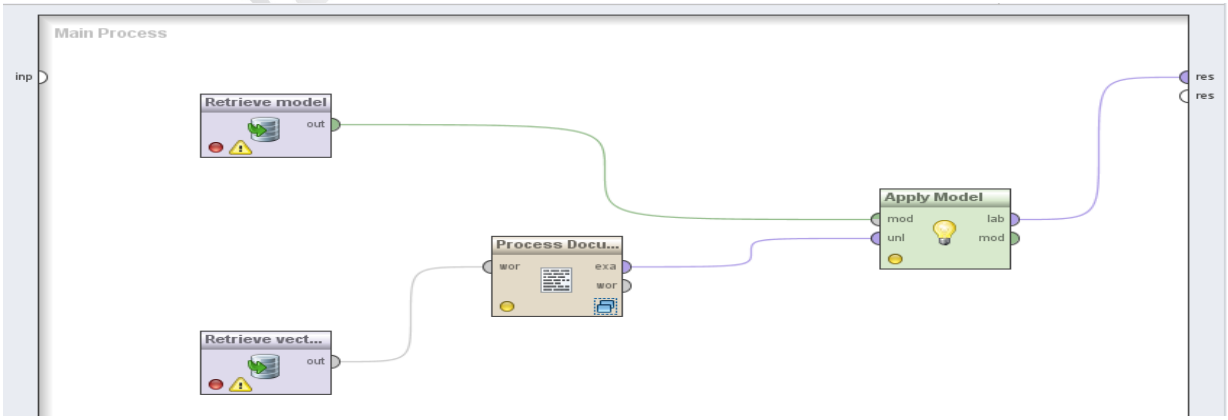


Word	Attribute Name	Total Occurences	Document Occurences	positive	negative
aaaaaaaaah	aaaaaaaaah	2	1	0	2
aaaaaaaahhhh	aaaaaaaahhhh	1	1	0	1
aaaaaah	aaaaaah	1	1	0	1
aaaahhh	aaaahhh	1	1	1	0
aah	aah	1	1	1	0
aaliyah	aaliyah	3	3	0	3
aalyah	aalyah	3	1	0	3
aamir	aamir	1	1	1	0
aardman	aardman	2	2	2	0
aaron	aaron	20	15	14	6
aatish	aatish	1	1	0	1
aback	aback	2	2	0	2
abandon	abandon	91	87	51	40
abat	abat	1	1	0	1
abb	abb	3	1	3	0
abba	abba	2	1	2	0
abber	abber	1	1	0	1
abberlin	abberlin	2	1	2	0
abbi	abbi	15	6	14	1
abbot	abbot	1	1	1	0
abbott	abbott	2	2	1	1
abbrevi	abbrevi	1	1	0	1
abdomen	abdomen	2	2	0	2
abduct	abduct	12	10	8	4
abducte	abducte	1	1	1	0
abdul	abdul	1	1	1	0
abel	abel	5	5	2	3
aberdeen	aberdeen	10	2	1	9



Στην συνέχεια θα εισάγουμε και θα επεξεργασθούμε τα νέα δεδομένα με τον ίδιο τρόπο όπως και πριν, δηλαδή με τον τελεστή Process Documents from Files.

Τέλος θα προσθέσουμε την εντολή Apply Model η οποία όπως προείπαμε έχει δύο εισόδους. Η μία είναι για το μοντέλο και η άλλη για τα δεδομένα. Το μόνο που μας μένει είναι να συνδέσουμε τις ροές και να δούμε τα αποτελέσματα.



Όπως φαίνεται από τις έξη κριτικές, οι 2 χαρακτηρίζοντα ως αρνητικές και οι 4 ως θετικές. Αν διαβάσουμε αυτές τις κριτικές θα συνηδητοποιήσουμε ότι περίπου στο 80% χαρακτηρίζονται σωστά, δηλαδή το ποσοστό που μας έδινε και κατά την αξιολόγηση που έκανε το ίδιο το Rapid Miner στο μοντέλο του.

ExampleSet (6 examples, 7 special attributes, 23798 regular attributes) Filter (6 / 6 examples): all

Row No.	label	prediction(label)	confidence(negative)	confidence(positive)	metadata_file	metadata_p...	metadata_d...	aaaaaaaaah	aaaaaaaah...	aaaaaah	aaaahhh	aah	aaliyah
1	unlabeled	positive	0.393	0.607	Berardinelli.txt	C:\Users\Us	Apr 23, 2014	0	0	0	0	0	0
2	unlabeled	positive	0.334	0.666	Brussat.txt	C:\Users\Us	Apr 23, 2014	0	0	0	0	0	0
3	unlabeled	positive	0.478	0.522	Clifford.txt	C:\Users\Us	Apr 23, 2014	0	0	0	0	0	0
4	unlabeled	negative	0.552	0.448	Johanson.txt	C:\Users\Us	Apr 23, 2014	0	0	0	0	0	0
5	unlabeled	positive	0.496	0.504	Maltin.txt	C:\Users\Us	Apr 23, 2014	0	0	0	0	0	0
6	unlabeled	negative	0.673	0.327	Nashawaty.txt	C:\Users\Us	Apr 23, 2014	0	0	0	0	0	0

Με αυτήν την διαδικασία λοιπόν μπορεί κάθε ασφαλιστική να προβεί σε τέτοιου είδους ελέγχους των δηλώσεων της προκειμένου να μπορεί να προβλέπει καλύτερα τυχών οικονομικές απαιτήσεις που μπορούν να προκύψουν από αυτές ή και επανέλεγχο κάποιων περιπτώσεων.

7.4. Συμπερασματολογία

Παρουσιάσθηκαν παραπάνω δύο εφαρμογές Text Mining από τις οποίες μπορούν να εξαχθούν χρήσιμα συμπεράσματα και να χρησιμοποιηθούν άμεσα από κάθε ασφαλιστική.

Πρώτα παρουσιάσθηκαν εφαρμογές ομαδοποίησης αδόμητων δηλώσεων. Με αυτή επιτεύχθηκε η δημιουργία ομάδων με ίδια νοήματα και χαρακτηριστικά προκειμένου να αρχειοθετηθούν ξεχωριστά και να αποτελούν από εδώ και στο εξής χρήσιμη πληροφορία για τις ασφαλιστικές και όχι δεδομένα που απλά γεμίζουν τις βάσεις δεδομένων τους.

Δεύτερον παρουσιάσθηκε μια εφαρμογή όπου προβλέπει, από ένα μεγάλο όγκο κειμένων, με ιδιαίτερα υψηλή αποτελεσματικότητα, αν ένα σχόλιο είναι θετικό ή αρνητικό. Όπως προαναφέραμε την ίδια ακριβώς διαδικασία μπορούν να ακολουθήσουν οι ασφαλιστικές προκειμένου να προβλέπουν την πιθανότητα μιας δήλωσης να είναι απάτη. Με αυτόν τον τρόπο θα εκπαιδεύσουν ένα μοντέλο το οποίο θα προσφέρει μια ένδειξη ή ένα συναγερμό για συγκεκριμένες δηλώσεις προκειμένου

να ελεγχθούν εκτενέστερα και με μεγαλύτερη ίσως προσοχή από τις υπόλοιπες. Έτσι θα προσθέσουν στην φαρέτρα τους ένα σημαντικό όπλο για την αντιμετώπιση αυτού του τόσο ζημιογόνου φαινομένου για αυτές όπως είναι η απάτη

ΕΛΛΗΝΙΚΗ ΒΙΒΛΙΟΓΡΑΦΙΑ

Ηλιόπουλος, Γ.Πολυμεταβλητή Ανάλυση.Πανεπιστημιακές Σημειώσεις. Πανεπιστήμιο Πειραιώς, Πειραιάς, 2008.

Καρλής, Δ.Πολυμεταβλητή Στατιστική Ανάλυση. Εκδόσεις Σταμούλης Α., Αθήνα, 2008.

Κούτρας, Μ.Εφαρμοσμένη Πολυμεταβλητή Ανάλυση-Ανάλυση κατά συστάδες. Πανεπιστημιακές Σημειώσεις. Πανεπιστήμιο Πειραιώς, Πειραιάς, 2008

Πανάρετος, Ι. και Ξεκαλάκη, Ε.Εισαγωγή στην Πολυμεταβλητή Στατιστική Ανάλυση. Εκδόσεις: Πανάρετος Ι., Αθήνα, 1995.

Σιώμκος, Γ. Ι. και Βασιλακοπούλου, Α. Ι.Εφαρμογή Μεθόδων Ανάλυσης στην Έρευνα Αγοράς. Εκδόσεις: Σταμούλης Α., Αθήνα, 2005.

ΞΕΝΗ ΒΙΒΛΙΟΓΡΑΦΙΑ

Burges, C., 1998. *A Tutorial on Support Vector Machines for Pattern Recognition, data mining and Pattern Recognition. Kluwer Academic Publishers.*

Craven M., DiPasquo D., McCallum A., Mitchell T., Nigam K. and Slattery S.1998. *Learning to extract symbolic knowledge from the world wide web.*

Dingsoyr T. and Lidal E. M., 1997. *An Evaluation of Data Mining Methods and Tools*, Norway, *Norwegian University of Science and Technology (NTNU).*

Elkan C, 2011. *Nearest Neighbor Classification. Morgan Kaufman Publishers, San Mateo CA*

Ester M., Kriegel H.-P., Sander J., and Xu. X., 1996. *A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of 2nd International Conference on KDD. Portland, OR, AAAI Press.*

Friburger N. and Maurel D., 2002. *Textual similarity based on proper names. In Proceedings of Workshop on Mathematical Formal Methods in Information Retrieval at th 25th ACM SIGIR Conference.*

Ghahramani Z., 2001. *An Introduction to Hidden Markov Models and Bayesian Networks. International Journal of Pattern Recognition and Artificial Intelligence, vol. 15, no. 1.*

Greevy, E., Smeaton, A., 2004. *Text Categorization of Racist Texts Using a Support Vector Machine. In annual International ACM SIGIR Conference on Research and Development in Infarmation Retrieval. ACM, New York.*

Han E., Boley D., Gini M., Gross R., Hastings K.,Karypis G., Kumar V., Mobasher B., and Moore J., 1998. Webace: *A web agent for document categorization and exploartion. In Proceedings of the 2nd International Conference on Autonomous Agents.*

Hotho A., Staab S. and Stumme G., 2003. *Wordnet improv*es *text document clustering. Toronto, In Proceedings of the SIGIR Semantic Web Workshop.*

Hu L. and Bentler M. P., 1999. *Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternative. Structural Equation Modelling: A Multidiscriplinary Journal.*

Jain A. K., Murty M. N., and Flynn P. J.,1999. *Data clustering: a review. ACM Computing Surveys* .

Johnson, D. E.,1998. *Applied Multivariate Methods for Data Analysis.Pacific Grove: Duxbury Press.*

Johnson R. A. and Wichern W. D.,1998. *Applied Multivariate Statistical Analysis 4rth ed. New Jersey: Prentice Hall.*

Larsen B. and Aone C., 1999. *Fast and effective text mining using linear-time document clustering. In Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.*

Lee, D., Seung, H.S., 1999. *Learning the Parts of Objects by Non-negative matrix factorization. Nature.*

Lin J., 1991. *Divergence measures based on the shannon entropy. IEEE Transaction on Information Theory.*

M. F. A. Gadi, X. Wang and A. Pereira do Lago, 2008. *Credit Card Fraud Detection with Artificial Immune System. Springer-Verlag Berlin Heidelberg.*

Mardia, K.V. , Kent, J.T. & Bibby, J.M., 1979. *Multivariate Analysis. London: Academic Press.*

Milne D., Medelyan O., and Witten I. H., 2006. *Mining domain-specific thesauri from wikipedia: A case study. In Proceedings of the International Conference on Web Intelligence.*

Neuhaus J. M. and Kalbfleisch J. D., 1998. *Between- and within-cluster covariate effects in the analysis of clustered data. Biometrics.*

Porter M. F.,1980. *An algorithm for suffix stripping. Morgan Kaufmann Publishers Inc.*

Rencher C. A.,2002. *Methods of Multivariate Analysis. 2nd ed. New Jersey: John Wiley & Sons, Inc.*

Salton G., 1989. *Automatic Text Processing. New York: Addison-Wesley.*

Sharma S.,1995. *Applied Multivariate Techniques. New Jersey: John Wiley & Sons, Inc.*

Steinbach M., Karypis G., and Kumar V., 2000. *A comparison of document clustering techniques. In KDD Workshop on Text Mining*

Strehl A., Ghosh J., and Mooney R., 2000. *Impact of similarity measures on web-page clustering. In AAAI-2000: Workshop on Artificial Intelligence for Web Search.*

Theodoridis, S. and Koutroumpas, K.,1999. *Pattern recognition. New York: American Press.*

Tishby N. Z., Pereira F., and Bialek W., 1999. *The information bottleneck method. In Proceedings of the 37th Allerton Conference on Communication, Control and Computing*.

Tsiptsis P. and Chorianopoulos A.,2010. *Data Mining Techniques in CRM. New Jersey: John Wiley & Sons.*

Vapnik, V., 1982. *Estimation of Dependencies Based on Empirical Data. New York, Springer Verlag.*

Voorhees and Harman D., 1998. *Overview of the fifth text retrieval conference. In Proceeding of the Fifth Text REtrieval Conference.*

Willett P., 1988. *Recent trends in hierarchic document clustering: a critical review. Information Processing and Management. An International Journal.*

Yates R. B. and Neto B. R., 1999. *Modern Information Retrieval. New York: ADDISON-WESLEY.*

Zhao Y. and Karypis G., 2002. *Evaluation of hierarchical clustering algorithms for document datasets. In Proceedings of the International Conference on Information and Knowledge Management.*

Zhao Y. and Karypis G., 2004. *Empirical and theoretical comparisons of selected criterion functions for document clustering. Machine Learning.*

Ιστοσελίδες

Amy, L., Carl, M., 2006. *ALS Algorithms Nonnegative Matrix Factorization Text Mining. http://meyer.math.ncsu.edu/Meyer/Talks/SAS_6_9_05_NmfWorkshop.pdf.*

Boutella, M., Shena, X., Luob, J., Browna1, C., 2006. *Multi-label Semantic Scene Classification.http://www.cs.rochester.edu/u/xshen/Publications/TR813.pdf.*

Brank, J., Grobelnik, M., 2006. *Training text classifiers with SVM on very few positive examples, April 2003.ftp://ftp.research.microsoft.com/pub/tr/tr-2003-34.pdf.*

Brian L., 2006. *Non Negative Matrix Factorization, Multidimensional Digital Signal Processing.http://www.ece.utexas.edu/~bevans/courses/ee381k/projects/spring03/*

Lewis D. D., 1999. *Reuters-21578 text categorization test collection distribution 1.0. http://www.research.att.com/ lewis.*

Pradhan, S.Ward, W, Hacioglu, K., Martin, J.H., 2006. *Shallow semantic parsing using support vector machines.http://www.stanford.edu/~jurafsky/hlt-2004-verb.pdf.*

Überarbeitung, J., 2004. *Text mining in the Life* Sciences.http://www.coling.unifreiburg.de /research/projects/TextMining/WhitePaperV