Πανεπιστήμιο Πειραιώς – Τμήμα Πληροφορικής
Πρόγραμμα Μεταπτυχιακών Σπουδών
«Πληροφορική»

Μεταπτυχιακή Διατριβή

Τίτλος Διατριβής	**ΓΡΑΦΟΘΕΩΡΗΤΙΚΟΙ ΑΛΓΟΡΙΘΜΟΙ ΗΜΙ-ΕΠΙΤΗΡΟΥΜΕΝΗΣ ΜΑΘΗΣΗΣ. ΕΦΑΡΜΟΓΗ ΣΕ ΔΕΔΟΜΕΝΑ ΤΗΣ ΡΟΔΟΥ.** **GRAPH – BASED SEMI-SUPERVISED ALGORITHMS. IMPLEMENTATION AT RHODES ISLAND DATA**
Ονοματεπώνυμο Φοιτητή	**ΣΑΡΑΣΟΥΑΤΙ – ΔΑΛΙΑΝΗΣ ΓΕΩΡΓΙΟΣ - ΔΑΣΡΑΤ**
Πατρώνυμο	**ΓΙΟΓΚΑΝΑΝΤΑ**
Αριθμός Μητρώου	**ΜΠΠΛ/ 15063**
Επιβλέπων	**ΔΙΟΝΥΣΙΟΣ ΣΩΤΗΡΟΠΟΥΛΟΣ, ΕΠΙΚΟΥΡΟΣ ΚΑΘΗΓΗΤΗΣ**

Ημερομηνία Παράδοσης **12/01/19**

Τριμελής Εξεταστική Επιτροπή

Δ. Σωτηρόπουλο, Επίκ. Καθηγητής	Γ. Τσιχριντζή, Καθηγητής	Ε. Αλέπη, Επίκ. Καθηγητής

Ευχαριστώ τους ανθρώπους με τους οποίους έχω καθημερινή τριβή και είναι αυτοί που με βοήθησαν να φτάσω στην πραγματοποίηση του στόχου μου, να ολοκληρώσω τις σπουδές μου. Εξελίσσομαι σαν προσωπικότητα μαζί τους.

Επίσης ευχαριστώ το Δωδεκανησιακό Ίδρυμα Υποτροφιών καθώς με την στήριξη του αισθάνομαι ότι κατευθύνομαι στο σωστό δρόμο της γνώσης.

Περίληψη

Ο σκοπός της παρούσας έρευνας είναι να βρεθούν οι τοποθεσίες που αναπτύσσουν κοινωνικοοικονομικά την περιοχή της Ρόδου. Αυτό επιτυγχάνεται μέσα από την ανάδειξη των τοποθεσιών με αυξημένη δημοτικότητα και κίνηση στο κοινωνικό δίκτυο της εταιρείας Facebook.

Ερευνητικοί στόχοι της εργασίας είναι να αναλυθούν κάποιες μεθοδολογίες ημι-επιτηρούμενης μάθησης graph based και να παραμετροποιηθούν προσανατολισμένες στα καλύτερα αποτελέσματα. Να παρουσιαστεί η επιστημονική βιβλιογραφία και να επιλεχθεί το κατάλληλο μοντέλο.

Όσον αφορά το αναπτυξιακό θέμα της Ρόδου, στόχος είναι να λειτουργήσουν σαν πρότυπα οι περιοχές - επιχειρήσεις με θετικό αντίκτυπο και να γίνει έναυσμα για υιοθέτηση των πρακτικών που κάνουν από τις άλλες επιχειρήσεις, ώστε να αυξηθεί η δημοτικότητα στο σύνολο της Ρόδου.

Θεωρητικά αναλύεται η ημι-επιτηρούμενη μάθηση και οι αλγόριθμοι ανάλυσης αισθήματος. Για την επιλογή του κατάλληλου αλγορίθμου εξετάστηκαν αλγόριθμοι διάδοσης και μετάδοσης με δεδομένα από υπάρχουσα επιστημονική εργασία. Μετέπειτα καταλήξαμε στο κώδικα που θα χρησιμοποιήσουμε και τον εφαρμόσαμε σε 674 εγγραφές με 3 παραμέτρους η καθεμία.

Στο τέλος μετά την ανάλυση καταλήγουμε σε συμπεράσματα που θα μας κάνουν να συνεχίσουμε την έρευνα μας και να προτυποποιήσουμε τοποθεσίες και οργανισμούς.

ABSTRACT

The purpose of this research is to locate the socio-economically developing sites of Rhodes. This is achieved through the promotion of sites with increased popularity and traffic in Facebook's social network.

The research objectives of the thesis are to analyze some graph based semi-supervised learning methodologies and to parameterize oriented to the best results. Present the scientific literature and select the appropriate model.

Regarding the development theme of Rhodes, the goal is to operate as a region-business-positive business and to stimulate the adoption of practices by other businesses in order to increase popularity throughout Rhodes.

In theory, semi-supervised learning and sense analysis algorithms are analyzed. For the selection of the appropriate algorithm, dissemination and transmission algorithms were examined with data from existing scientific work. Later we ended up using the code we used and applied it to 674 records with 3 parameters each.

At the end of the analysis, we come to conclusions that will make us continue our research and standardize locations and organizations.

Περιεχόμενα

Εισαγωγή

Στο πρώτο κεφάλαιο αναφερόμαστε στην θεωρητική προσέγγιση για τον τρόπο με τον οποίο αναλύουμε και επεξεργαζόμαστε τα δεδομένα μας. Αυτός δεν είναι άλλος από την μηχανική μάθηση και το παρακλάδι της, την ημι-επιτηρούμενη μάθηση. Πιο συγκεκριμένα της διάδοση της ετικέτας του εκάστοτε δεδομένου στις υπόλοιπες μεταβλητές μας.

Στο δεύτερο κεφάλαιο, έχουμε την βιβλιογραφική επισκόπηση στην εξόρυξη γνώσης και ανάλυση αισθήματος, όπως επίσης παρουσιάζονται παραδείγματα για την καλύτερη κατανόηση του αντικειμένου. Αναλύεται βιβλιογραφικά οι αλγόριθμοι ανάλυσης αισθήματος μέσα από τις δημοσιεύσεις της παγκόσμιας επιστημονικής κοινότητας.

Κατόπιν, έχουμε την μεθοδολογία. Όπου είχαμε πρώτα την αξιολόγηση των αποτελεσμάτων των αλγορίθμων σαν αλγόριθμοι για να δούμε ποιον θα υλοποιήσουμε και μετέπειτα την συλλογή των πρωτογενών δεδομένων και την αξιολόγηση των αποτελεσμάτων της έρευνας μας. Αναφέραμε τα βήματα των αλγορίθμων διάδοσης και μετάδοσης και τα υλοποιήσαμε. Αξιολογήσαμε τις υλοποιήσεις μας μέσα από δευτερογενή δεδομένα από δημοσιευμένη επιστημονική εργασία και καταλήξαμε στον κώδικά που θα χρησιμοποιήσουμε για την έρευνα της δημοτικότητας των τοποθεσιών της Ρόδου. Ποιες δηλαδή τοποθεσίες έχουν αυξημένη δημοτικότητα στο κοινωνικό δίκτυο έχοντας ετικέτα με σημείωση +1 και είναι αντίστοιχες της Παλιάς Πόλης της Ρόδου. Επιπλέον παρουσιάζονται τα περιγραφικά στατιστικά με benchmarking.

Στο τελευταίο κεφάλαιο παρουσιάζονται τα συμπεράσματα μας,, τα οποία κατά κύριο λόγο είναι δύο και γίνονται προτάσεις για περαιτέρω έρευνα.

Ημι επιτηρούμενη μάθηση (θεωρητική προσέγγιση)

Η ημι-επιτηρούμενη μάθηση είναι μια κλάση των εργασιών και των τεχνικών της επιτηρούμενης μάθησης, η οποία χρησιμοποιεί δεδομένα χωρίς ετικέτες για την εκπαίδευση τους. Τυπικά, χρησιμοποιεί ένα μικρό αριθμό δεδομένων με ετικέτες και ένα μεγάλο χωρίς ετικέτες. Η ημι-επιτηρούμενη μάθηση βρίσκεται μεταξύ της μάθησης χωρίς επιτήρηση (unlabeled data) και της επιτηρούμενης μάθησης (full labeled training data). Πολλοί ερευνητές μηχανικής μάθησης βρήκαν ότι τα δεδομένα χωρίς ετικέτα, όταν χρησιμοποιούνται σε σύνδεση με ένα μικρό δείγμα δεδομένων με ετικέτα μπορούν να παράγουν σημαντική βελτίωση στην ακρίβεια της εκπαίδευσης – μάθησης. Το κόστος που συνδέεται με την διαδικασία πλήρους επισήμανσης – ετικετοποίησης, μπορεί να καταστήσει μια εκπαίδευση ως μη εφικτή. Σε μια τέτοια περίπτωση θα μπορούσε η ημι-επιτηρούμενη μάθηση να έχει πλούσια πρακτική αξία. Η ημι-επιτηρούμενη μάθηση έχει επίσης θεωρητικό ενδιαφέρον τη μηχανική μάθηση και ως πρότυπο για την ανθρώπινη μάθηση.

Όπως και στο πλαίσιο επιτηρούμενης μάθησης, μας δίνεται ένα σύνολο “L” ετικετών από ανεξάρτητα κατανεμημένα στοιχεία:

$$x_1, \ldots, x_l \in X$$

Τα οποία αντιστοιχούν σε ετικέτες:

$$y_1, \ldots, y_l \in Y.$$

Επιπρόσθετα, δίνεται ένα διάνυσμα **u** στοιχείων χωρίς ετικέτα :

$$x_{l+1}, \ldots, x_{l+u} \in X.$$

Η ημι-επιτηρούμενη μάθηση προσπαθεί να χρησιμοποιήσει αυτές τις συνδυασμένες πληροφορίες για να ξεπεράσει την απόδοση ταξινόμησης που θα μπορούσε να επιτευχθεί είτε με την απόρριψη των μη επισημασμένων δεδομένων και την επίβλεψη της μάθησης είτε με την απόρριψη των ετικετών και την εκμάθηση χωρίς επίβλεψη. Η ημι-επιτηρούμενη μάθηση μπορεί να αναφέρεται είτε στη μεταγωγική μάθηση είτε στην

επαγωγική μάθηση. Ο στόχος της μεταγωγικής μάθησης είναι να συμπεράνει τις σωστές ετικέτες για τα δεδομένα μη επισημασμένα δεδομένα.

$x_{l+1}, \ldots, x_{l+u}$.

Ενώ ο στόχος της επαγωγικής μάθησης είναι να συναγάγουμε τη σωστή χαρτογράφηση από το $\boldsymbol{X}$ στο $\boldsymbol{Y}$.

Προκειμένου να γίνει χρήση τυχόν μη επισημασμένων δεδομένων, πρέπει να υποθέσουμε κάποια δομή στην υποκείμενη κατανομή δεδομένων. Οι ημι-εποπτευόμενοι αλγόριθμοι μάθησης χρησιμοποιούν τουλάχιστον μία από τις ακόλουθες παραδοχές.

Υποθέσεις & προϋποθέσεις

Υπόθεση Συνέχειας

Τα σημεία που βρίσκονται κοντά μεταξύ τους είναι πιθανότερο να μοιράζονται μια ετικέτα. Αυτό είναι γενικά υποτιθέμενο στην εποπτευόμενη μάθηση και αποδίδει προτίμηση στα γεωμετρικά απλά όρια αποφάσεων. Στην περίπτωση της μάθησης με ημι-επιτήρηση, η παραδοχή ομαλότητας αποδίδει επιπλέον προτίμηση στα όρια απόφασης σε περιοχές χαμηλής πυκνότητας, έτσι ώστε να υπάρχουν λιγότερα σημεία κοντά μεταξύ τους αλλά σε διαφορετικές τάξεις.

Υπόθεση Ομαδοποίησης

Τα δεδομένα τείνουν να σχηματίζουν διακριτές συστάδες και τα σημεία στο ίδιο σύμπλεγμα είναι πιο πιθανό να μοιράζονται μια ετικέτα (παρόλο που τα δεδομένα που μοιράζονται μια ετικέτα μπορούν να διαδοθούν σε πολλαπλά σμήνη). Πρόκειται για μια ειδική περίπτωση της υποθετικής ομαλότητας και οδηγεί στην εκμάθηση χαρακτηριστικών με αλγόριθμους ομαδοποίησης.

Υπόθεση Τοπολογίας (Manifold)

Τα δεδομένα βρίσκονται περίπου σε μια τοπολογικά πολύ μικρότερη διάσταση από τον χώρο εισόδου. Σε αυτή την περίπτωση μπορούμε να προσπαθήσουμε να μάθουμε την τοπολογία χρησιμοποιώντας τόσο τα επισημασμένα όσο και τα μη επισημασμένα δεδομένα για να αποφύγουμε την κατάρα των διαστάσεων. Στη συνέχεια, η εκμάθηση μπορεί να προχωρήσει χρησιμοποιώντας αποστάσεις και πυκνότητες που ορίζονται στην τοπολογία.

Η παραδοχή της τοπολογίας είναι πρακτική όταν παράγονται δεδομένα μεγάλης διαστάσεως με κάποια διαδικασία που μπορεί να είναι δύσκολο να μοντελοποιηθεί άμεσα, αλλά η οποία έχει μόνο λίγους βαθμούς ελευθερίας. Για παράδειγμα, η ανθρώπινη φωνή ελέγχεται από μερικές φωνητικές πτυχές και εικόνες διαφόρων εκφράσεων του προσώπου ελέγχονται από λίγους μυς. Θα θέλαμε σε αυτές τις περιπτώσεις να χρησιμοποιούμε αποστάσεις και ομαλότητα στο φυσικό χώρο του προβλήματος δημιουργίας και όχι στο χώρο όλων των πιθανών ακουστικών κυμάτων ή εικόνων αντίστοιχα.

Μέθοδοι

Γενετικά μοντέλα

Οι γενετικές προσεγγίσεις στη στατιστική μάθηση προσπαθούν πρώτα να εκτιμήσουν το:

$$p(x|y)$$

τη κατανομή των σημείων δεδομένων που ανήκουν σε κάθε κατηγορία.

Η πιθανότητα p (y | x) που ένα δεδομένο σημείο x έχει την ετικέτα y τότε είναι ανάλογη του p (x | y) p (y) με τον κανόνα του Bayes. Η ημι-επιτηρούμενη μάθηση με γενετικά μοντέλα μπορεί να θεωρηθεί είτε ως επέκταση της επιτηρούμενης μάθησης (ταξινόμηση συν πληροφορίες σχετικά με p (x)) είτε ως επέκταση της μη επιτηρούμενης μάθησης (ομαδοποίηση συν μερικές ετικέτες).

Τα γενετικά μοντέλα υποθέτουν ότι οι κατανομές παίρνουν κάποια συγκεκριμένη μορφή p (x | y, \ θ) παραμετροποιημένη από τον φορέα «θ». Εάν αυτές οι παραδοχές είναι λανθασμένες, τα μη επισημασμένα δεδομένα ενδέχεται να μειώσουν στην πραγματικότητα την ακρίβεια της λύσης σε σχέση με αυτό που θα προέκυπτε μόνο από τα επισημασμένα δεδομένα. Ωστόσο, εάν οι υποθέσεις είναι σωστές, τότε τα μη επισημασμένα δεδομένα βελτιώνουν κατ 'ανάγκη την απόδοση.

Τα μη επισημασμένα δεδομένα κατανέμονται σύμφωνα με ένα μείγμα διανομών ατομικής κλάσης. Προκειμένου να μάθει τη διανομή του μείγματος από τα μη επισημασμένα δεδομένα, πρέπει να είναι αναγνωρίσιμη, δηλαδή διαφορετικές παράμετροι πρέπει να αποδίδουν διαφορετικές κατανομές αθροισμάτων. Οι κατανομές του μείγματος Gauss είναι αναγνωρίσιμες και χρησιμοποιούνται ευρέως για γενετικά μοντέλα.

Διαχωρισμού χαμηλής πυκνότητας

Μια άλλη σημαντική κατηγορία μεθόδων επιχειρεί να τοποθετήσει όρια σε περιοχές όπου υπάρχουν λίγα σημεία δεδομένων (με ετικέτα ή χωρίς ετικέτα). Ένας από τους συνηθέστερα χρησιμοποιούμενους αλγόριθμους είναι η μεταγωγική διανυσματική υποστήριξη μάθησης ή αλλιώς εν συντομία TSVM (η οποία, παρά το όνομά της, μπορεί να χρησιμοποιηθεί και για την επαγωγική μάθηση). Ενώ το support vector machine (SVM) για επιτηρούμενη μάθηση αναζητά τα όρια της απόφασης πρόβλεψης με μέγιστο

περιθώριο κέρδους πάνω από τα δεδομένα που έχουν επισημανθεί, ο στόχος του TSVM είναι η επισήμανση των μη επισημασμένων δεδομένων, έτσι ώστε το όριο απόφασης να έχει μέγιστο περιθώριο κέρδους σε όλα τα δεδομένα. Επιπρόσθετα, άλλες προσεγγίσεις που εφαρμόζονται σε αυτήν την κατηγορία περιλαμβάνουν τα μοντέλα Γκαουσιανής διεργασίας. Κανονικοποιώντας τις πληροφορίες και ελαχιστοποιώντας την εντροπία.

Βασισμένο σε γραφήματα (Graph-based method)

Αυτή η μέθοδος για ημιεπιτηρούμενη μάθηση χρησιμοποιεί την αναπαράσταση των δεδομένων με κόμβους για κάθε επισημασμένο και μη παράδειγμα. Το γράφημα κατασκευάζεται χρησιμοποιώντας το πεδίο γνώσης ή παρόμοια παραδείγματα, δύο κοινοί μέθοδοι όπου συνδέονται με το εκάστοτε σημείο δεδομένου στο «*k*» κοντινότερο γείτονα ή στα παραδείγματα με μερική απόσταση «*ε*» . Τα βάρη των ακμών μεταξύ των *Xi & Xj* προσαρμόζονται στο:

$$e^{\frac{-\|x_i - x_j\|^2}{\epsilon}}.$$

Στο πλαίσιο της συστηματοποίησης της τοπολογίας (Manifold regularization) η γραφική παράσταση χρησιμεύει ως υποκατάστατο για την τοπολογία. Ένας όρος προστίθεται στο τυποποιημένο πρόβλημα τακτοποίησης του Tikhonov για την επιβολή της ομαλότητας του διαλύματος σε σχέση με την τοπολογία (στον εγγενή χώρο του προβλήματος) καθώς και σε σχέση με τον χώρο εισόδου περιβάλλοντος. Το πρόβλημα ελαχιστοποίησης γίνεται

$$\underset{f \in \mathcal{H}}{\operatorname{argmin}} \left(\frac{1}{l} \sum_{i=1}^{l} V(f(x_i), y_i) + \lambda_A \|f\|_{\mathcal{H}}^2 + \lambda_I \int_{\mathcal{M}} \|\nabla_{\mathcal{M}} f(x)\|^2 dp(x) \right)$$

Όπου *"H"* είναι η αναπαραγωγή του πυρήνα στο χώρο Hilbert και "*M"* είναι η τοπολογία στην οποία βρίσκονται τα δεδομένα. Η συστηματοποίηση των παραμέτρων λ_A και λ_I ελέγχει την ομαλότητα στους περιβάλλοντες και τους εσωτερικούς χώρους αντίστοιχα. Η γραφική μέθοδος χρησιμοποιεί κατά προσέγγιση τον όρο της εσωτερικής συστηματοποίησης. Προσδιορίζει το γράφημα Laplacian $L=D-W$ όπου

$$D_{ii} = \sum_{j=1}^{l+u} W_{ij} \text{ a}$$

Και *f* το διάνυσμα

$$[f(x_1) \dots f(x_{l+u})],$$

Επομένως έχουμε,

$$\mathbf{f}^T L\mathbf{f} = \sum_{i,j=1}^{l+u} W_{ij}(f_i - f_j)^2 \approx \int_{\mathcal{M}} \|\nabla_{\mathcal{M}} f(x)\|^2 dp(x).$$

Η Laplacian μέθοδος μπορεί επίσης να χρησιμοποιηθεί σε προέκταση των αλγορίθμων επιτηρούμενης μάθησης. Συστηματοποιώντας τα ελάχιστα των τετραγώνων και τον SVM στην ημιεπιτηρούμενη μάθηση με Laplacian ελαχιστοποίηση των τετραγώνων και Laplacian SVM.

Ευρετικές προσεγγίσεις

Μερικές μέθοδοι για την ημιεπιτηρούμενη μάθηση δεν είναι εγγενώς προσανατολοσιμένοι να μαθαίνουν και από τα επισημασμένα και μη δεδομένα, αλλά αντ 'αυτού να κάνουν χρήση των μη επισημασμένων δεδομένων μέσα σε ένα εποπτευόμενο πλαίσιο μάθησης. Για παράδειγμα τα επισημασμένα και μη παραδείγματα μπορούν να ενημερώσουν την επιλογή της αναπαράστασης, από τις μετρικές της απόστασης ή από τον πυρήνα των δεδομένων χωρίς επιτήρηση για first step.

Τα αυτοεκπαίδευση είναι μια μέθοδος που περιτυλίγει την ημιεπιτηρούμενη μάθηση. Αρχικά, ένας αλγόριθμος επιτηρούμενης μάθησης είναι εκπαιδευμένος σε επισημασμένα δεδομένα μόνο. Αυτός ο ταξινομητής εφαρμόζεται στη συνέχεια στα μη επισημασμένα δεδομένα για να παράγει περισσότερα επισημασμένα παραδείγματα ως είσοδο για τον αλγόριθμο εποπτευόμενης μάθησης. Γενικά, σε κάθε βήμα προστίθενται μόνο οι ετικέτες που ο ταξινομητής είναι πιο σίγουρος.

Η συνεκπαίδευση είναι μια επέκταση της αυτο-εκπαίδευσης στην οποία πολλαπλοί ταξινομητές εκπαιδεύονται σε διαφορετικά (ιδανικά διαχωρισμένα) σύνολα χαρακτηριστικών και παράγουν επισημασμένα παραδείγματα μεταξύ τους.

Ανθρώπινης γνωστικής λειτουργείας (Human cognition)

Οι απαντήσεις των ανθρώπων σε επίσημα προβλήματα μάθησης οδήγησαν σε διαφορετικά συμπεράσματα σχετικά με τον βαθμό επίδρασης των μη επισημασμένων δεδομένων . Πιο φυσικά προβλήματα μάθησης μπορεί επίσης να θεωρηθούν ως περιπτώσεις ημιεπιτηρούμενης μάθησης. Μεγάλο μέρος της ανθρώπινης έννοιας μάθησης περιλαμβάνει μια μικρή ποσότητα άμεσης διδασκαλίας (π.χ. γονική επισήμανση αντικειμένων κατά τη διάρκεια της παιδικής ηλικίας) σε συνδυασμό με μεγάλες ποσότητες μη επισημασμένης εμπειρίας (π.χ. παρατήρηση αντικειμένων χωρίς ονομασία ή απαρίθμηση αυτών ή τουλάχιστον χωρίς ανατροφοδότηση). Τα ανθρώπινα βρέφη είναι ευαίσθητα στη δομή των μη επισημασμένων φυσικών κατηγοριών όπως εικόνες σκύλων και γάτων ή ανδρών και γυναικών. Πιο πρόσφατες εργασίες έχουν δείξει ότι τα βρέφη και τα παιδιά λαμβάνουν υπόψη όχι μόνο τα διαθέσιμα παραδείγματα που δεν έχουν επισημανθεί, αλλά η διαδικασία δειγματοληψίας από την οποία προκύπτουν επισημασμένα παραδείγματα.

Εξόρυξη γνώσης – Ανάλυση αισθήματος (Sentiment Analysis)

Η ανάλυση αισθήματος ή αλλιώς εξόρυξη γνώσης αναφέρεται στη χρησιμοποίηση της επεξεργασίας της φυσικής γλώσσας, του text analysis, της υπολογιστικής γλωσσολογίας και των βιομετρικών για τον συστηματικό προσδιορισμό τους, εξάγοντας ποσοτικοποιημένα αποτελέσματα και μελέτες. Η ανάλυση αισθήματος εφαρμόζεται ευρέως στις κριτικές και τις απαντήσεις ερευνών , στα διαδικτυακά μέσα ενημέρωσης και κοινωνικοποίησης, καθώς επίσης και στα υγειονομικά θέματα. Γενικότερα εφαρμόζεται για το Μάρκετινγκ μέχρι και την ιατρική.

Η ανάλυση συναισθημάτων έχει ως στόχο να καθορίσει τη στάση ενός ομιλητή – πομπού – συγγραφέα ή κάποιου υποκειμένου σε σχέση με κάποιο θέμα ή τη συνολική εικόνα που απορρέει. Η συμπεριφορά που εκτιμάται μπορεί να είναι μία κρίση ή αξιολόγηση , μια συναισθηματική κατάσταση ή μια επιδιωκόμενη συναισθηματική επικοινωνία κοκ.

Παράδειγματα

Ο στόχος και οι προκλήσεις της ανάλυσης του συναισθήματος μπορούν να αποδειχθούν με μερικά απλά παραδείγματα.

Απλές περιπτώσεις

Το Blue Star 2 είναι το καλύτερο πλοίο

Δεν μου αρέσουν τα παλιά πλοία

Πιο πολύπλοκες

Μου αρέσει το κινητό μου αλλά δεν το συνιστώ σε κανέναν από τους συναδέλφους μου. (Αναγνωρισμένο θετικό συναίσθημα, δύσκολο να κατηγοριοποιηθεί)

Η ταινία προκαλεί έκπληξη με πολλές ανησυχητικές συσπάσεις. (Αρνητικός όρος που χρησιμοποιείται με θετική έννοια σε ορισμένους τομείς).

Βιβλιογραφική επισκόπηση του SVM

Η ανάπτυξη των κοινωνικών μέσων δικτύωσης έχουν γεννήσει ένα τεράστιο όγκο δεδομένων ο οποίος μοχλεύεται από το ανάλυση των ίδιων (Social Media Analysis - SMA). Η ανάλυση με τον χαρακτήρα που έχουν αποκτήσει τα κοινωνικά μέσα δικτύωσης πλέον μπορεί να γίνει και σε επιχειρησιακό επίπεδο. Η Sentiment Analysis είναι ένα σημαντικό πεδίο έρευνας του SMA, το οποίο είναι βασισμένο στα συναισθήματα ή στην μεροληπτική πολικότητα των κειμένων των κοινωνικών δικτύων, δηλαδή στο κοινωνικό συναίσθημα.

Η απόδοση της ανάλυσης συναισθημάτων και εξόρυξης γνώμης μέσω Twitter είναι μια περιοχή που έχει αντλήσει το ενδιαφέρον πολλών ερευνητών. Η πρόκληση με ακρίβεια να προβλέψει την κοινωνική διάθεση με βάση το κείμενο που εξήχθη από το Twitter, παραμένει μια μεγάλη πρόκληση και βρίσκεται σήμερα σε εξέλιξη, η οποία διερευνήθηκε σε διάφορους τομείς της αγοράς και της ακαδημαϊκής κοινότητας. Ο O'Connor σε συνδεδεμένες μετρήσεις της κοινής γνώμης που μετριέται από δημοσκοπήσεις με υπολογίσιμο συναίσθημα από κείμενο και διαπίστωσε ότι οι απόψεις που μετρήθηκαν από δημοσκοπήσεις συσχετίζονται με τις συχνότητες λέξεων συναισθημάτων στο σύγχρονα μηνύματα του Twitter. Η μελέτη καταλήγει στο δυναμικό της χρήσης των ροών κειμένου ως υποκατάστατο και συμπλήρωμα για την παραδοσιακή δημοσκόπηση. Ο Jansen διερεύνησε τη συνολική δομή των καταχωρίσεων μικρο-blog, των τύπων εκφράσεων και τις διακυμάνσεις του συναισθήματος συζητώντας τις συνέπειες για τον οργανισμό στη χρήση του μικρο-blogging ως μέρος της συνολικής στρατηγικής μάρκετινγκ και εκστρατείες της επωνυμίας (branding). Ο Mishne στη μελέτη του, δείχνει ότι, στο τομέα των ταινιών, υπάρχει καλή συσχέτιση μεταξύ αναφορών σε ταινίες σε αναρτήσεις ιστολογίου - και πριν και μετά την απελευθέρωσή τους - και τις οικονομικές επιτυχίες των ταινιών. Επιπλέον, αποδεικνύει όταν γίνεται χρήση της ανάλυσης συναισθήματος στα ιστολόγια μπορεί να βελτιωθεί η συσχέτιση των παραμέτρων. Ο Tumasjan χρησιμοποίησε το πλαίσιο των γερμανικών ομοσπονδιακών εκλογών για να διερευνήσει αν το Twitter χρησιμοποιείται ως φόρουμ πολιτικής συζήτησης και το κατά πόσο τα μηνύματα στο διαδίκτυο στο Twitter αντικατοπτρίζουν έγκυρα πολιτικά αισθήματα. Τελικά, η λεπτομερής μελέτη διαπίστωσε τα μηνύματα αντανακλούν το αποτέλεσμα των εκλογών και έρχονται ακόμη και κοντά στις παραδοσιακές δημοσκοπήσεις. Ο Bollen και η ομάδα του, υποστηρίζουν ότι η ψυχική διάθεση του Twitter προβλέπει τη χρηματιστηριακή αγορά. Σε μελέτη τους καταλήγουν στο συμπέρασμα ότι οι αλλαγές στη κατάσταση διάθεσης του κοινού μπορεί πράγματι να εντοπιστεί από το περιεχόμενο της μεγάλης κλίμακας του Twitter που τροφοδοτεί με τη βοήθεια της μάλλον απλές τεχνικές επεξεργασίας κειμένου και ότι αυτές οι αλλαγές ανταποκρίνονται σε μια ποικιλία κοινωνικο-πολιτιστικών οδηγών με ένα πολύ διαφοροποιημένο τρόπο που με τη σειρά του συσχετίζει ή ακόμα και πρόβλεψη των τιμών χρηματιστηριακών δεικτών.

Η ανάλυση συναισθημάτων του περιεχομένου του online κειμένου είναι πλέον σε ένα ώριμο στάδιο και ένα μεγάλο μέρος των επιχειρήσεων αγοράζει λογισμικό ανάλυσης όπως το Radian6 ή το IBM Cognos Consumer Insight. L.A. Times.

Το εργαστήριο καινοτομίας της USC Annenberg έχει χρησιμοποιήσει την ανάλυση συναισθήματος στο twitter και τροφοδοτείται με δεδομένα για να προβλέψει τα Όσκαρ του 2012. Η IBM μαζί με την USC Annenberg πραγματοποίησαν ανάλυση συναισθήματος για το Super Bowl XLVI , το οποίο αναλύει 600.000 tweets για να καθορίσετε ποιοι παίκτες και ποιες ομάδες έχουν την μεγαλύτερη υποστήριξη. Αν και η έρευνα στο πεδίο των αναλύσεων συναισθημάτων περιλαμβάνει πολλές μελέτες που εκτίμησαν την πολικότητα του κοινωνικού συναισθήματος με μια ποικιλία προσεγγίσεων (π.χ. με βάση την εκμάθηση μηχανών αλγορίθμων ή μεθόδων ταξινόμησης λεξικών) υπάρχουν περιορισμένες γνώσεις σχετικά με τις αιτίες που οδηγούν σε μια ιδιαίτερη κατάσταση κοινωνικού αισθήματος. Πράγματι, το κοινωνικό συναίσθημα μπορεί να θεωρηθεί πολυδιάστατο φαινόμενο. Συλλογικές απόψεις σχετικά με τις βασικές συζητήσεις, τα θέματα που μπορεί να επηρεάσουν θετικά ή αρνητικά την πολικότητα του κοινωνικού συναισθήματος. Αυτή η επεξηγηματική διερεύνηση του κοινωνικού συναισθήματος μπορεί να αποβεί κρίσιμη για τις πληροφορίες και για την απόδοση του υπό μελέτη αντικειμένου.

Οι αναλυτές δεδομένων των κοινωνικών μέσων γνωρίζουν να αποσυνθέτουν τους παράγοντες του κοινωνικού συναισθήματος με τη μορφή διαφόρων σημασιολογικά καθορισμένων ιδιοτήτων. Μια άλλη μελέτη (SVM-Based Sentiment Classification: A Comparative Study against State-of-the-Art Classifiers - 2006) υιοθετεί και αποσκοπεί στην ανάπτυξη ενός πλαισίου που εξηγεί τις αιτίες του κοινωνικού αισθήματος αντί να την καταγράφετε απλά. Συγκεκριμένα, πρόκειται για μια εκτενή πειραματική σύγκριση όπου εξετάσανε τον ταξινομητή SVM έναντι ενός συνόλου σύγχρονων ταξινομητών μηχανικής μάθησης σε σύνολο δεδομένων αναφοράς που προέρχεται από τον τομέα των ελληνικών τραπεζών με τη συλλογή στοιχείων από το streaming API του Twitter που αναφερόταν ρητά στις μεγάλες τράπεζες της Ελλάδας. Τα αποτελέσματά στην παρούσα ακρίβεια ταξινόμησης και τις μετρήσεις χρόνου εκτέλεσης για κάθε ταξινομητή αποκαλύπτουν την υπεροχή του παραδείγματος μάθησης SVM στην εκχώρηση μοτίβων στη σωστή τάξη συναισθημάτων.

Η προσέγγιση που ακολουθήθηκε σε αυτή τη μελέτη πήρε τα δευτερογενή δεδομένα από το paper των Dionisios N. Sotiropoulos, Chris D. Kounavis, George M. Giaglis and Panos Kourouthanassis με τίτλο «What drives social sentiment? An entropic measure-based clustering approach towards identifying factors that influence social sentiment polarity». Στο οποίο αναλύουν το κοινό συναίσθημα πάνω από τα κοινωνικά μέσα και τις ροές που αποτελούν ένα εξαιρετικά απαιτητικό έργο που οφείλεται κυρίως στις δυσκολίες που επιβάλλονται από το ευρύ φάσμα των θεμάτων συζήτησης που αποτελούν τη βάση μιας δεδομένης συλλογής θέσεων. Πιο συγκεκριμένα στη μελέτη τους εξετάζουν το πρόβλημα του προσδιορισμού του υποκείμενου των σημασιολογικών παραγόντων που επηρεάζουν την πολικότητα του κοινωνικού συναισθήματος μέσω της χρήσης εντροπίας

με βάση τις μετρήσεις. Εκτεταμένες μελέτες εξετάζουν τη σημασιολογική δομή των δεδομένων κοινωνικού δικτύου κατά κύριο λόγο μέσα από τη μοντελοποίηση του θέματος ή τις μεθόδους ανάλυσης στο συναίσθημα. Η καινοτομία στη προσέγγισή τους έγκειται στην αξιοποίηση μιας σημασιολογικά ευαίσθητης διαδικασίας ομαδοποίησης που συνδυάζει αποτελεσματικά τη μοντελοποίηση θεμάτων και αλγορίθμων ανάλυσης συναισθήματος. Η προσέγγισή τους επεκτείνει τη θεμελιώδη παραδοχή πίσω από την παραδοσιακή μέθοδο ανάλυσης αισθήσεων, σύμφωνα με την οποία μπορεί να συνδεθεί το συναίσθημα, όπως τα χαρακτηριστικά του εγγράφου χαμηλού επιπέδου που είναι λέξεις, φράσεις ή πεδία. Υποστηρίζουν ότι το συναίσθημα μπορεί να συνδεθεί με οντότητες υψηλότερου επιπέδου όπως οι σημασιολογικοί άξονες που καλύπτουν ένα δεδομένο όγκου των καταστάσεων, πραγματοποιώντας έτσι ανάλυση συναισθημάτων σε επίπεδο θέματος κειμένου. Ο πειραματισμός τους, παρέχει ισχυρές ενδείξεις ότι συνδυάζοντας τα αποτελέσματα της ανάλυσης θεμάτων και των συναισθημάτων με μια σημασιολογικά γνωστή διαδικασία ομαδοποίησης μπορεί να αποκαλύψει την κατανομή του συνολικού δημόσιου κλίματος στο υποκείμενο των σημασιολογικών αξόνων. **Βάσει λοιπόν των ευρημάτων τους κατασκευάστηκαν αλγόριθμοι, συγκρίθηκαν με τα αποτελέσματα τους και αφού μας δείξαν την εγκυρότητας τους εφαρμόστηκαν σε δικά μας δεδομένα**

Οι αλγόριθμοι που υλοποιήσαμε προέρχονται από την δημοσίευση των Yoshua Bengio, Olivier Delalleau και Nicolas Le Roux με τίτλο «Label Propagation and Quadratic Criterion» το έτος 2006. Οι οποίοι χρησιμοποίησαν τους γραφοθεωρητικούς αλγόριθμους ημιεπιτηρούμενης μάθησης ώστε να εξάγουν ικανοποιητικά αποτελέσματα. Παρουσίασαν πως διάφοροι αλγόριθμοι μπορούν να μετατραπούν σε ένα κοινό πλαίσιο και να επιλύσουν ένα πολυδιάστατο πρόβλημα.

Μεθοδολογία

Την μεθοδολογία στην συγκεκριμένη μελέτη την χωρίσαμε σε δύο τμήματα, την ερευνητική και την πειραματική. Αυτό έγινε ώστε να μπορέσουμε να κατασκευάσουμε τον κατάλληλο αλγόριθμο και να σιγουρευτούμε ότι μας παρέχει αξιόπιστα αποτελέσματα και μετέπειτα να το εφαρμόσουμε στο πείραμα μας. Το οποίο εφαρμόζεται σε επιχειρησιακά δεδομένα οργανισμών στη Ρόδο.

Τεχνολογία

Για την υλοποίηση των αλγορίθμων χρησιμοποιήθηκε η γλώσσα προγραμματισμού python με έκδοση 3.5. Οι κύριες βιβλιοθήκες που χρειάστηκαν ήταν οι εξής, η numpy για την αποδοτικότητα της και υπολογιστική της ικανότητα, η pandas για την διαχείριση των δεδομένων και η scipy τις υπολογιστικές παραμέτρους της. Όλες πλαισιώνονται από το anaconda.

Αλγόριθμοι

Η ιδέα των αλγορίθμων βασίζεται στην υπόθεση διάδοσης ετικετών. Δηλαδή, θεωρούμε σαν σημεία στο χώρο τα δεδομένα μας από τα οποία κάποια έχουν ετικέτα +1, κάποια -1 και κάποια δεν έχουν. Αυτές βασίζονται στους παραμέτρους των σημείων αλλά και στην μεταξύ τους απόσταση. Στην συνέχεια βάση του εκάστοτε αλγορίθμου οι τιμές που έχουν πάρει οι κόμβοι, μας δείχνουν την «προτίμηση» ή αλλιώς την «προσέγγιση» στην ετικέτα τους.

Αρχικά, υπολογίσαμε τον πίνακα W , την μήτρα δηλαδή των ευκλείδειων αποστάσεων στο Γκαουσινιανό πυρήνα με πλάτος σ:

$$\mathbf{W}_{ij} = e^{-\frac{\|x_i - x_j\|^2}{2\sigma^2}}.$$

Όπου ο αριθμητής είναι η ευκλείδεια απόσταση των εκάστοτε σημείων.

Βήματα 1ου αλγορίθμου (Διάδοσης –Zhu and Ghahramani [2002])

- Υπολογίζουμε τον W
- Υπολογίζουμε τη διαγώνιο D από το

$$\sum_j W_{ij}$$

- Αρχικοποιούμε το Y_0

$$\hat{Y}^{(0)} \leftarrow (y_1, \ldots, y_l, 0, 0, \ldots, 0)$$

- Βρίσκουμε το

$$\hat{Y}^{(t+1)} \leftarrow \mathbf{D}^{-1}\mathbf{W}\hat{Y}^{(t)}$$

- Αντικαθιστούμε

$$\hat{Y}_l^{(t+1)} \leftarrow Y_l$$

- Επαναλαμβάνουμε τα δύο προηγούμενα βήματα μέχρι ότου διάνυσμα μας να φτάσει στο άπειρο
- Συλλέγουμε τις τιμές

Βήματα 2ου αλγορίθμου (Διάδοσης – Παρόμοιος με του Jacobi – Yoshua Bengio, Olivier Delalleau και Nicolas Le Roux)

- Υπολογίζουμε τον W
- Μηδενίζουμε την διαγώνιο του W
- Υπολογίζουμε τη διαγώνιο D όπως προηγουμένως
- Επιλέγουμε μια παράμετρο *α* η οποία βρίσκεται μεταξύ 0 και 1
- Επιλέγουμε ένα μικρό αριθμό *ε* >0
- Υπολογίζουμε το *μ=α/(α-1)*
- Υπολογίζουμε τον πίνακα A ο οποίος προέρχεται από την παρακάτω παράσταση

$$\mathbf{A}_{ii} \leftarrow I_{[l]}(i) + \mu\mathbf{D}_{ii} + \mu\epsilon$$

  Το I είναι ένας πίνακας που δείχνει πότε οι κόμβοι είναι επισημασμένοι
- Επαναλαμβάνουμε το 3ο βήμα του 1ου αλγορίθμου

- Υπολογίζουμε όπως προηγουμένως, δηλαδή μέχρι το Y να προσεγγίσει το άπειρο την παράσταση

$$\hat{Y}^{(t+1)} \leftarrow \mathbf{A}^{-1}(\mu \mathbf{W}\hat{Y}^{(t)} + \hat{Y}^{(0)})$$

Βήματα 3ου αλγορίθμου (Μετάδοσης – Zhou et al. [2004])

- Υπολογίζουμε τον W
- Μηδενίζουμε την διαγώνιο του W
- Υπολογίζουμε τη διαγώνιο D όπως προηγουμένως
- Υπολογίζουμε το κανονικοποιημένο γράφημα Laplacian

$$\mathcal{L} \leftarrow \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}$$

- Αρχικοποιούμε το Y_0 όπως προηγουμένως
- Επιλέγουμε μια παράμετρο *α* στο διάστημα από 0 έως 1
- Υπολογίζουμε όπως προηγουμένως, δηλαδή μέχρι το Y να προσεγγίσει το άπειρο την παράσταση

$$\hat{Y}^{(t+1)} \leftarrow \alpha\mathcal{L}\hat{Y}^{(t)} + (1-\alpha)\hat{Y}^{(0)}$$

Σε όλες τις περιπτώσεις θα λάβουμε κάποια αποτελέσματα σε μορφή διανυσμάτων τα οποία θα μας λένε τι τιμή προβλέπουν για την ετικέτα του εκάστοτε κόμβου.

Με τα δεδομένα της έρευνας θα γίνει αξιολόγηση συγκρίνοντας τα αποτελέσματα των αλγορίθμων με τα αποτελέσματα της εργασίας που συλλέξαμε τα δευτερογενή δεδομένα. Αν πρόκειται για έγκυρο αλγόριθμο τότε εφαρμόζουμε τους αλγορίθμους στο πείραμα μας.

Αξιολόγηση - Υλοποίηση Κώδικα

Σε αυτό το σκέλος θα γίνει η παρουσίαση της ανάλυσης και των αποτελεσμάτων του πρώτου σκέλους που αφορά την αξιολόγηση των αλγορίθμων μας.

Για το πρώτο αλγόριθμο δεν βρέθηκαν σημαντικά αποτελέσματα επομένως δεν προστίθεται στην έρευνα μας.

Δειγματοληψία Έρευνας

Συλλέξαμε και αναλύσαμε ένα σύνολο πάνω από 135.000 tweets κατά τη χρονική περίοδο μεταξύ 2 Φεβρουαρίου και 13 Φεβρουαρίου 2013, χρησιμοποιώντας το API ροής του Twitter. Η συλλογή δεδομένων επικεντρώθηκε στη συγκέντρωση των tweets που ήταν ρητά αναφερόμενη στους δύο κορυφαίους κινητούς ευρυζωνικούς φορείς που βρίσκονται στην ευρύτερη περιοχή της Βόρειας Αμερικής, δηλαδή στην AT & T και την Verizon. Αυτό το έργο επιτεύχθηκε με την τεχνολογία streaming API του Twitter μέσω της διεύθυνσης λέξεων - κλειδιών στους όρους "AT & T" και "Verizon".

Το σύνολο δεδομένων που προέκυψε από ένα συνολικό αριθμό 66.000 και 70.000 tweets για το AT & T και για το Verizon αντίστοιχα, το οποίο στη συνέχεια υποβλήθηκε σε μια σειρά εργασιών εκκαθάρισης δεδομένων και προεπεξεργασίας. Για τα δεδομένα υπήρξε μια διαδικασία προετοιμασίας. Ειδικότερα, για την μετατροπή κειμένου σε λέξεις, καταργήθηκαν οι αγγλικές λέξεις stopwords (όπως “a”, “an”, “or”) και οι λέξεις με λιγότερο από τρεις χαρακτήρες, και έγινε εξαγωγή της ρίζας από κάθε λέξη. Επομένως, η τελική έκδοση του των κειμένων μας σχηματίστηκε από μια συλλογή καθαρισμένων εγγράφων όπου κάθε έγγραφο περιέλαβε το κείμενο από ένα μόνο tweet.

2ος αλγόριθμος (με αξιολόγηση)

```
#ΕΙΣΑΓΟΥΜΕ ΤΙΣ ΒΙΒΛΙΟΘΗΚΕΣ
import pandas as pd
import numpy as np
from scipy.io import loadmat
from scipy.spatial.distance import pdist, squareform
from scipy.spatial import distance
#ΤΕΛΟΣ ΕΙΣΑΓΩΓΗΣ ΒΙΒΛΙΟΘΗΚΩΝ
#ΔΗΩΝΟΥΜΕ ΤΙΣ ΣΥΝΑΡΤΗΣΕΙΣ ΜΑΣ
#ΓΙΑ ΝΑ ΕΙΣΑΓΟΥΜΕ ΤΑ ΔΕΔΟΜΕΝΑ ΜΑΣ
def Load_Data():
        x=loadmat("sentiment_corpus_tfidf_vectors_400.mat")
        y=loadmat("sentiment_corpus_tfidf_vectors_400_labels.mat")
        x=pd.DataFrame(x['tfidf_vectors'])
        y=pd.DataFrame(y['tfidf_vectors_labels']).T
        y.columns = ['Y']
        y = pd.concat([y, x], axis=1)#load Data to dataframe pandas
        return y
#ΠΡΟΕΠΕΞΕΡΓΑΖΟΜΑΣΤΕ ΤΑ ΔΕΔΟΜΕΝΑ ΜΑΣ
def proepeksergasia(y,k):
        y=y[y.Y != 0]#Deleting DataFrame row in Pandas based on column value==0
        y=y.drop(y.index[[0,1]])#delete two label +1 to have bigger greater divisor
        y=y[0:]
        y = y.reset_index()#anaprosarmogi index
        y['Xi'] = y.index#anti gia index Xi || gia paradeigma X1 X2 ...Xn
        del y['index']#delete old index
        y['Y_eval']=y['Y']#dimiourgia Y gia elegxo sto telos ;oste na ginei evaluation
        y['Y0']=y['Y']#statheres times Y0
```

```
    cols=list(y.columns.values)
    cols = cols[-3:] + cols[:-3]
    y=y[cols]#anakatanomi columns
    return y
```

#ΕΠΙΛΕΓΟΥΜΕ ΤΑ ΔΕΔΟΜΕΝΑ ΠΟΥ ΧΡΕΙΑΖΟΜΑΣΤΕ – ΑΦΑΙΡΟΥΜΕ ΑΥΤΑ ΠΟΥ ΔΕΝ ΘΕΛΟΥΜΕ

```
def Select_Data(n,y):
    X=y.loc[0:,y.columns[3:]]
    y0=X.as_matrix(columns=["Y"])*1
    del X["Y"]#afairoume ta label gia na kanoyme upologismoys aperiskepta
    X=X.as_matrix()#metatrepoyme se numpy matrix
    return X,y0
```

#ΥΠΟΛΟΓΙΖΟΥΜΕ ΤΟΝ ΠΙΝΑΚΑ W

```
def calculate_W(X):
    #pairwise_sq_dists = squareform(pdist(X, 'sqeuclidean'))
    #W = np.exp(-((pairwise_sq_dists))/( 2* (np.std( np.linalg.norm(X, axis=1))**2)))
    W=distance.cdist(X, X, 'euclidean')
    W = np.exp(-(np.power(W,2))/( 2*(np.std( np.linalg.norm(X, axis=1))**2)))
    W[np.diag_indices_from(W)] = 0
    return W
```

#ΥΠΟΛΟΓΙΖΟΥΜΕ ΤΟΝ ΠΙΝΑΚΑ D

```
def calculate_D(n,W):
    D=np.zeros((n, n),float)
    Diagwnios= np.array(W.sum(axis=1))#Wi,j sum
    np.fill_diagonal(D,Diagwnios)
    return D
```

#ΣΟΡΤΑΡΟΥΜΕ ΤΟΥΣ ΕΠΙΣΗΜΑΣΜΕΝΟΥΣ ΚΑΙ ΜΗ ΚΟΜΒΟΥΣ ΜΑΣ

```
def Dataframe_short(y):
```

```
    y = [y.loc[y['Y'] == 1],y.loc[y['Y'] == -1],y.loc[(y['Y'] != 1) & (y['Y'] != -1 )]]
    y = pd.concat(y)
    y=y.reset_index(drop=True)
    return y
```

#ΥΠΟΛΟΓΙΖΟΥΜΕ ΤΟΝ ΠΙΝΑΚΑ Ι

```
def estimate_I(n,y0):
    I=np.zeros((n,n),int)
    np.fill_diagonal(I,np.abs(y0))
    return I
```

#ΥΠΟΛΟΓΙΖΟΥΜΕ ΤΟΝ ΠΙΝΑΚΑ Α

```
def estimate_A(n,I,μ,D,e):
    A=np.dot(μ,D)
    #print (A)
    di = np.diag_indices(n)
    A[di] = A[di] +μ*e
    A[di]=A[di]+I[di]
    return A
```

#ΞΕΚΙΝΑΕΙ ΤΟ ΠΡΟΓΡΑΜΜΑ

#ΘΕΤΟΥΜΕ ΠΑΡΑΜΕΤΡΟΥΣ ΚΑΙ ΚΑΛΟΥΜΕ ΤΙΣ ΣΥΝΑΡΤΗΣΕΙΣ ΣΥΜΦΩΝΑ ΜΕ ΤΗΝ ΣΕΙΡΑ ΤΩΝ #ΑΛΓΟΡΙΘΜΩΝ

```
a=0.5
μ=a/(1-a)
y=Load_Data()
y=proepeksergasia(y,k)
n=len(y.index)
neg,pos= y['Y'].value_counts()
y=Dataframe_short(y)#sortaroume ta stoixeia prwta 1 meta -1 kai telos ta midenika poy dimourgisame
```

```
y_mirror=y*1

k_neg=0

neg,pos= y['Y'].value_counts()
```

#ΥΠΟΛΟΓΙΖΟΥΜΕ ΣΕ ΠΟΣΑ ΤΜΗΜΑΤΑ ΘΑ ΣΠΑΣΟΥΜΕ ΤΗΝ ΑΞΙΟΛΟΓΗΣΗ ΜΑΣ

```
from math import gcd

gcd=gcd(neg, pos)

print (gcd)

neg_fold=int(neg/gcd)

pos_fold=int(pos/gcd)


neg_fold=int(neg*0.1)

pos_fold=int(pos*0.1)

print (neg_fold)

print (pos_fold)
```

#ΚΑΠΟΙΕΣ ΛΙΣΤΕΣ ΩΣΤΕ ΝΑ ΜΠΟΡΕΣΟΥΜΕ ΝΑ ΚΑΤΑΧΩΡΙΣΟΥΜΕ ΑΡΓΟΤΕΡΑ ΤΑ #ΑΠΟΤΕΛΕΣΜΑΤΑ ΤΗΣ ΑΞΙΟΛΟΓΗΣΗ ΜΑΣ

```
q,w,e,r=[],[],[],[]
```

#ΞΕΚΙΝΑΕΙ Η ΕΠΑΝΑΛΗΨΗ ΑΞΙΟΛΟΓΟΓΗΣΗΣ (FOLDING)

```
for k in range(0,pos,pos_fold):

        n=len(y.index)

        y.loc[y.index[k:k+pos_fold],'Y'] = 0#make some Xi unlabeled || Yo and Y==0 !analoga poia theloume na dialekoume[0:10]= ta prwta deka

        y.loc[y.index[pos:pos+neg_fold],'Y'] = 0#make some Xi unlabeled || Yo and Y==0 !analoga poia theloume na dialekoume[0:10]= ta

        y=Dataframe_short(y)

        X,y0=Select_Data(n,y)


        W=calculate_W(X)

        D=calculate_D(n,W)
```

```
I=estimate_I(n,y0[0:n])
A=estimate_A(n,I,μ,D,e1)
yt=np.matrix(np.array(y0[0:n]))
di = np.diag_indices(n)
xx=0
AA=np.linalg.inv(A)
statheros_oros=np.dot(W,μ)
statheros_oros=np.dot(AA,statheros_oros)
y0=np.dot(AA,y0)
e_check = np.longdouble(1.)
```

#ΥΠΟΛΟΓΙΖΟΥΜΕ ΤΑ Υ ΜΕΧΡΙ ΠΟΥ ΠΡΟΣΕΓΓΙΖΟΥΝ ΤΟ ΑΠΕΙΡΟ

```
while  e_check>0.0000000001:
        y1=np.add(np.dot(statheros_oros,yt), y0)      #upologizoyme ta dianismata
        e_check=np.sum(np.abs(np.subtract(y1,yt)))#checkaroume gia to apeiro
        yt=y1*1.
```

#ΚΑΤΑΧΩΡΟΥΜΕ ΤΙΣ ΛΥΣΕΙΣ ΓΙΑ ΝΑ ΓΙΝΕΙ Η ΑΞΙΟΛΟΓΗΣΗ

```
y_eval=y.as_matrix(['Y_eval'])
print ("unlabeled positive percentage")
false_at_y1,false_at_y_eval,= list(np.where( (np.sign(y1[n-neg_fold-pos_fold:n-neg_fold]) == np.sign(y_eval[n-neg_fold-pos_fold:n-neg_fold]))==False))
print (len(y1[false_at_y1])/pos_fold)
superman=superman+ (len(y1[false_at_y1])/pos_fold)
q.append(len(y1[false_at_y1]))
print ("unlabeled negative percentage")
false_at_y1,false_at_y_eval,= list(np.where( (np.sign(y1[n-k_neg:n]) == np.sign(y_eval[n-k_neg:n]))==False))
print (len(y1[false_at_y1])/neg_fold)
superman=superman+ (len(y1[false_at_y1])/pos_fold)
```

```
        w.append(len(y1[false_at_y1]))

        print ("labeled positive percentage")

        false_at_y1,false_at_y_eval,= list(np.where( (np.sign(y1[0:pos-pos_fold]) ==
np.sign(y_eval[0:pos-pos_fold]))==False))

        print (len(y1[false_at_y1])/(pos-pos_fold))

        superman=superman+ (len(y1[false_at_y1])/pos_fold)

        e.append(len(y1[false_at_y1]))

        print ("labeled negative percentage")

        false_at_y1,false_at_y_eval,= list(np.where( (np.sign(y1[pos-pos_fold:n-
(pos_fold+neg_fold)]) == np.sign(y_eval[pos-pos_fold:n-
(pos_fold+neg_fold)]))==False))

        print (len(y1[false_at_y1])/(neg-neg_fold))

        superman=superman+ (len(y1[false_at_y1])/pos_fold)

        r.append(len(y1[false_at_y1]))

        print ("sunolika gia unlabeled")

        false_at_y1,false_at_y_eval,= list(np.where( (np.sign(y1) ==
np.sign(y_eval))==False))

        print (len(false_at_y1)/(pos_fold+neg_fold))

        false_at_y1,false_at_y_eval,= list(np.where( (np.sign(y1) ==
np.sign(y_eval))==False))

        print (len(false_at_y1))


        y=y_mirror*1

        k_neg=int(k_neg+neg_fold)
```

3ος αλγόριθμος (με αξιολόγηση)

#Τα σχόλια είναι τα ίδια με την προηγούμενη υλοποίηση με την διαφορά ότι υπολογίζουμε #άλλες μεταβήτές

```
import pandas as pd
import numpy as np
from scipy.io import loadmat
from scipy.spatial.distance import pdist, squareform
from scipy.spatial import distance


def Load_Data():
        x=loadmat("sentiment_corpus_tfidf_vectors_400.mat")
        y=loadmat("sentiment_corpus_tfidf_vectors_400_labels.mat")
        x=pd.DataFrame(x['tfidf_vectors'])
        y=pd.DataFrame(y['tfidf_vectors_labels']).T
        y.columns = ['Y']
        y = pd.concat([y, x], axis=1)#load Data to dataframe pandas
        return y
def proepeksergasia(y,k):
        y=y[y.Y != 0]#Deleting DataFrame row in Pandas based on column value==0
        y=y.drop(y.index[[0,1]])#delete two label +1 to have bigger greater divisor


        y=y[0:]
        y = y.reset_index()#anaprosarmogi index
        y['Xi'] = y.index#anti gia index Xi || gia paradeigma X1 X2 ...Xn
        del y['index']#delete old index


        y['Y_eval']=y['Y']#dimiourgia Y gia elegxo sto telos ;oste na ginei evaluation


        y['Y0']=y['Y']#statheres times Y0
```

```
    cols=list(y.columns.values)
    cols = cols[-3:] + cols[:-3]
    y=y[cols]#anakatanomi columns
    return y
def Select_Data(n,y):
    X=y.loc[0:,y.columns[3:]]
    y0=X.as_matrix(columns=["Y"])*1
    del X["Y"]#afairoume ta label gia na kanoyme upologismoys aperiskepta
    X=X.as_matrix()#metatrepoyme se numpy matrix
    return X,y0
def calculate_W(X):
    #pairwise_sq_dists = squareform(pdist(X, 'sqeuclidean'))
    #W = np.exp(-((pairwise_sq_dists))/( 2* (np.std( np.linalg.norm(X,
axis=1))**2)))
    W=distance.cdist(X, X, 'euclidean')
    W = np.exp(-(np.power(W,2))/( 2*(np.std( np.linalg.norm(X, axis=1))**2)))
    W[np.diag_indices_from(W)] = 0
    return W
def calculate_D(n,W):
    D=np.zeros((n, n),float)
    Diagwnios= np.array(W.sum(axis=1))#Wi,j sum
    np.fill_diagonal(D,Diagwnios)
    return D


def Dataframe_short(y):
    y = [y.loc[y['Y'] == 1],y.loc[y['Y'] == -1],y.loc[(y['Y'] != 1) & (y['Y'] != -1 )]]
    y = pd.concat(y)
    y=y.reset_index(drop=True)
    return y
```

```
def estimate_L(W,D,n):
        di = np.diag_indices(n)
        D[di]=np.float_power(D[di],-0.5)
        L=np.dot(W,D)
        L=np.dot(D,L)
        return L



a=0.5
k=0#variable to make some data unlabeled
y=Load_Data()
y=proepeksergasia(y,k)
n=len(y.index)
neg,pos= y['Y'].value_counts()
y=Dataframe_short(y)#sortaroume ta stoixeia prwta 1 meta -1 kai telos ta midenika poy dimourgisame
y_mirror=y*1
k_neg=0
neg,pos= y['Y'].value_counts()

from math import gcd
gcd=gcd(neg, pos)
print (gcd)
neg_fold=int(neg/gcd)
pos_fold=int(pos/gcd)

neg_fold=int(neg*0.1)
pos_fold=int(pos*0.1)
```

```
print (neg_fold)
print (pos_fold)
q,w,e,r=[],[],[],[]


for k in range(0,pos,pos_fold):

    y.loc[y.index[k:k+pos_fold],'Y'] = 0#make some Xi unlabeled || Yo and Y==0 !analoga poia theloume na dialekoume[0:10]= ta prwta deka
    y.loc[y.index[pos+k_neg:pos+k_neg+neg_fold],'Y'] = 0#make some Xi unlabeled || Yo and Y==0 !analoga poia theloume na dialekoume[0:10]= ta prwta deka

    y=Dataframe_short(y)#sortaroume ta stoixeia prwta 1 meta -1 kai telos ta midenika poy dimourgisame

    X,y0=Select_Data(n,y)
    W=calculate_W(X)
    D=calculate_D(n,W)
    L=estimate_L(W,D,n)


    yt=np.matrix(np.array(y0[0:n]))
    statheros_oros1=np.dot(L,a)
    statheros_oros2=np.dot(y0,0.5)

    e_check=1
    while e_check>0.000000000000001 and not np.isnan(e_check):
        y1=np.add(np.dot(statheros_oros1,yt),statheros_oros2)
        e_check=np.abs(np.sum(np.subtract(y1,yt)))
```

```
            yt=y1*1.
        y_eval=y.as_matrix(['Y_eval'])


        print ("unlabeled positive percentage")
        false_at_y1,false_at_y_eval,= list(np.where( (np.sign(y1[n-neg_fold-pos_fold:n-neg_fold]) == np.sign(y_eval[n-neg_fold-pos_fold:n-neg_fold]))==False))
        print (len(y1[false_at_y1])/pos_fold)
        superman=superman+ (len(y1[false_at_y1])/pos_fold)
        q.append(len(y1[false_at_y1]))
        print ("unlabeled negative percentage")
        false_at_y1,false_at_y_eval,= list(np.where( (np.sign(y1[n-k_neg:n]) == np.sign(y_eval[n-k_neg:n]))==False))
        print (len(y1[false_at_y1])/neg_fold)
        superman=superman+ (len(y1[false_at_y1])/pos_fold)
        w.append(len(y1[false_at_y1]))
        print ("labeled positive percentage")
        false_at_y1,false_at_y_eval,= list(np.where( (np.sign(y1[0:pos-pos_fold]) == np.sign(y_eval[0:pos-pos_fold]))==False))
        print (len(y1[false_at_y1])/(pos-pos_fold))
        superman=superman+ (len(y1[false_at_y1])/pos_fold)
        e.append(len(y1[false_at_y1]))
        print ("labeled negative percentage")
        false_at_y1,false_at_y_eval,= list(np.where( (np.sign(y1[pos-pos_fold:n-(pos_fold+neg_fold)]) == np.sign(y_eval[pos-pos_fold:n-(pos_fold+neg_fold)]))==False))
        print (len(y1[false_at_y1])/(neg-neg_fold))
        superman=superman+ (len(y1[false_at_y1])/pos_fold)
        r.append(len(y1[false_at_y1]))
        print ("sunolika gia unlabeled")
```

```
        false_at_y1,false_at_y_eval,= list(np.where( (np.sign(y1) ==
np.sign(y_eval))==False))

        print (len(false_at_y1)/(pos_fold+neg_fold))

        false_at_y1,false_at_y_eval,= list(np.where( (np.sign(y1) ==
np.sign(y_eval))==False))

        print (len(false_at_y1))


        y=y_mirror*1

        k_neg=int(k_neg+neg_fold)
```

Αποτελέσματα αξιολόγησης

Στους παρακάτω πίνακες συγκεντρώνονται τα αποτελέσματα από τους έγκυρους αλγορίθμους. Αυτό που παρατηρούμε είναι ότι και στους δύο αλγορίθμους παρουσιάζεται σφάλμα γύρω στο 15%, το οποίο το θεωρούμε ικανοποιητικό ώστε να εφαρμόσουμε στα δεδομένα των οργανισμών της Ρόδου.

Ειδικότερα, σύμφωνα με τα αποτελέσματα θα χρησιμοποιήσουμε τον αλγόριθμο διάδοσης – spreading καθώς είναι λαμβάνει υπόψη του πολύπλευρα το πρόβλημα.

propagation	ΚΟΜΒΟΙ	ΑΓΝΩΣΤΑ LABEL	1η Επανάληψη	2η Επανάληψη	3η Επανάληψη	4η Επανάλ	5η Επανάλ	6η Επανάλ	7η Επανάλ	8η Επανάληψη	9η Επανάληψη	10 Επανάληψη	ΜΟ
ΣΥΝΟΛΟ	9590												
-1	6460	646	0	37	80	96	100	120	101	115	119	153	92.1
1	3130	313	59	80	40	54	51	64	44	55	57	80	58.4
		ΣΥΝΟΛΙΚΑ ΣΦΑΛΜΑΤΑ ΑΓΝΩΣΤΩΝ	59	117	120	150	151	184	145	170	176	233	150.5
		ΠΟΣΟΣΤΟ	6.15%	12.20%	12.51%	15.64%	15.75%	19.19%	15.12%	17.73%	18.35%	24.30%	15.69%
		ΓΝΩΣΤΑ LABEL											
-1		5814	9	30	35	34	32	11	10	32	32	34	25.9
1		2817	41	34	34	41	41	41	41	41	41	41	39.6
		ΣΥΝΟΛΙΚΑ ΣΦΑΛΜΑΤΑ ΓΝΩΣΤΩΝ	50	64	69	75	73	52	51	73	73	75	65.5
		ΠΟΣΟΣΤΟ	0.58%	0.74%	0.80%	0.87%	0.85%	0.60%	0.59%	0.85%	0.85%	0.87%	0.76%

spreading	ΚΟΜΒΟΙ	ΑΓΝΩΣΤΑ LABEL	1η Επανάληψη	2η Επανάληψη	3η Επανάληψη	4η Επανάλ	5η Επανάλ	6η Επανάλ	7η Επανάλ	8η Επανάληψη	9η Επανάληψη	10 Επανάληψη	ΜΟ
ΣΥΝΟΛΟ	9590												
-1	6460	646	0	38	90	97	99	145	105	130	121	181	100.6
1	3130	313	63	82	48	45	53	62	41	45	57	75	57.1
		ΣΥΝΟΛΙΚΑ ΣΦΑΛΜΑΤΑ ΑΓΝΩΣΤΩΝ	63	120	138	142	152	207	146	175	178	256	157.7
		ΠΟΣΟΣΤΟ	6.57%	12.51%	14.39%	14.81%	15.85%	21.58%	15.22%	18.25%	18.56%	26.69%	16.44%
		ΓΝΩΣΤΑ LABEL											
-1		5814	0	0	0	0	0	0	0	0	0	0	0
1		2817	0	0	0	0	0	0	0	0	0	0	0
		ΣΥΝΟΛΙΚΑ ΣΦΑΛΜΑΤΑ ΓΝΩΣΤΩΝ	0	0	0	0	0	0	0	0	0	0	0
		ΠΟΣΟΣΤΟ	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%

Πείραμα – Εφαρμογή σε δεδομένα της Ρόδου

Δειγματοληψία Πειράματος

Συλλέξαμε δεδομένα από το κοινωνικό δίκτυο Facebook. Για την συλλογή του εκτελέσαμε ερώτημα στην βάση δεδομένων της ιστοσελίδας κοινωνικής δικτύωσης μέσω του Facebook API.

Οι εγγραφές μας ήταν 674 τοποθεσίες γύρω από την Ρόδο. Οι παράμετροι της εκάστοτε εγγραφής ήταν το πλήθος των δημοσιεύσεων τοποθεσιών (checkins), το πλήθος των αξιολογήσεων τοποθεσιών (rating count) και η αξιολόγηση των τοποθεσιών (overall rating). Επιπλέον, λήφθηκαν δεδομένα όπως το επίπεδο κόστους της εκάστοτε τοποθεσίας, τα οποία όμως είναι για ποιοτική ανάλυση των υποκειμένων.

Τις παραπάνω εγγραφές καταχωρίσαμε στη Παλιά Πόλη της Ρόδου την ετικέτα +1, λαμβάνοντας υπόψη ότι είναι ένα δημοφιλές μέρος που επισκέπτονται πολλοί παραθεριστές και μη σύμφωνα με το αρχείο των παγκόσμιων πολιτιστικών κληρονομιών της UNESCO. Αρνητική ετικέτα, με την πρόφαση ότι δεν δημιουργεί θετική δημοσιότητα, ούτε βέβαια και αρνητική , δώσαμε σε δύο τυχαίες εγγραφές από τις χειρότερες καταχωρήσεις στο tripadvisor. Πηγαίνοντας μάλιστα στην εγγραφή της Παλιά Πόλης βλέπουμε μια ομοιογένεια των σχετικών παραμέτρων της. Όπως αντίστοιχα και στις άλλες δύο.

Ουσιαστικά θέλαμε να βρούμε τις δημοφιλέστερες περιοχές - επιχειρήσεις με θετικό αντίκτυπο στην κοινωνία της Ρόδου αλλά και στον τουρισμό που είναι αντίστοιχες με αυτές της Παλιά Πόλης της Ρόδου.

Για το ερώτημα στη βάση δεδομένων του κοινωνικού δικτύου, επιλέξαμε λέξεις κλειδιά όπως φαίνονται και στον κώδικα παρακάτω αλλά ταυτόχρονα και το γεωγραφικό μήκος και πλάτος του νησιού με εύρος την απόσταση στο ερώτημα.

Κώδικας Δειγματοληψίας

```
import urllib3

import facebook

import requests

import json

import pandas as pd

keyword=['Ρόδος','kiotari','lindos','sea','seaside','beach','blue
lagoon','tourism','tourist','indian','italiania','pizza','tamam','chinese','2004','swedco','rhodes'
,'rodos','rhodos','cafe','bar','restaurant','rhodes','rhodes','museum','culture','food','beverage','
danish
corner','fish','drink','club','hotel','apartment','apartments','pansion','hostel','accommodation'
,'store','drink','ouzo','musaka','coffee','cofee','kafe','espresso','mojito','capuchino','cappuchi
no','vodka','bacardi','jin','square','barstreet','wine']
```

```
token=
'EAACTxoj7TjgBADZAToWLSZBEBG31pceHj4xsPzJsFLdLLyBsxaqzZAPkgnKmdm
IFV6RvbvpXrlT4bma4SXmSYczwcKLSsabZBOEsoH6FiZC0CxFN0ldvZBFjdWY4aq2
FX3rPmOZCp6HIydouSiZBqDZCjPm8aSOLvBZAuQDJKKAdrZB6Qlw7WD7LWN7R
ALLfsW5ZB4YZD'



graph =facebook.GraphAPI(access_token=token, version = 2.10)


epix = graph.request('/search?type=place&q=
&center=36.446694,28.219757&distance=1000&limit=2000&fields=name,price_range,c
heckins,rating_count,overall_star_rating')

epix1=epix['data']

df=pd.DataFrame(data=epix1)


for i in keyword:

        epix =
graph.request('/search?type=place&q='+i+'&center=36.446694,28.219757&distance=100
0&limit=100&fields=name,price_range,checkins,rating_count,overall_star_rating')

        epix1=epix['data']

        df1=pd.DataFrame(data=epix1)

        df=df.append(df1)


df.to_pickle('dataset', compression='infer')

df=pd.read_pickle('dataset', compression='infer')

df=df.reset_index()

print (df)
```

Επεξεργασία Δεδομένων

Καθώς το ερώτημα μας μπορεί να εμπεριέχει εσφαλμένες εγγραφές προεπεξεργαστήκαμε το αρχείο σύμφωνα με τον παρακάτω κώδικα. Επίσης βρήκαμε τις μέγιστες τιμές της εκάστοτε παραμέτρου ώστε να ομοιογενιοποιήσουμε και μοναδιοποιήσουμε τα δεδομένα μας για εύλογη απόρροια των αποτελεσμάτων μας.

ΚΩΔΙΚΑΣ ΠΡΟΕΠΕΞΕΡΓΑΣΙΑΣ

```
import json

import pandas as pd

#fortosi dedomenwn rodoy

df=pd.read_pickle('dataset', compression='infer')


#proepeksergasia kai ekatharisi dedomenwn

df=df.reset_index()

df= (df.drop_duplicates('id'))

df=df.sort_values('rating_count')

df=df.drop('index', axis=1)

df=df[df.name != 'Tivoli']

df=df[df.id != '104995096203106']

df=df[df.id != '200596543437449']

df=df[df.id != '549561225154313']

df=df.drop('id', axis=1)


#euresi max gia kanonikopoihsh

print (df.loc[df['checkins'].idxmax()])

print (df.loc[df['rating_count'].idxmax()])

print (df.loc[df['overall_star_rating'].idxmax()])

#kanonikopoihsh - monadiopoihsh

df['checkins']= (df['checkins'].divide(463587, axis=0))

df['rating_count']= (df['rating_count'].divide(1066, axis=0))
```

```
df['overall_star_rating']= (df['overall_star_rating'].divide(5, axis=0))



df=df.sort_index()

df = df.reset_index(drop=True)

print (df)

#euresi label basi dimofilotitas apo ta eksis link


#https://www.tripadvisor.com.gr/Restaurant_Review-g189449-d4117148-Reviews-Bar_in_Rhodes_old_town-Rhodes_Dodecanese_South_Aegean.html

#https://www.tripadvisor.com.gr/Restaurant_Review-g635613-d10352972-Reviews-Panorama_Cafe_Bar-Rhodes_Town_Rhodes_Dodecanese_South_Aegean.html

#palia poli +1 - politistiki klironomia

#http://www.rodiaki.gr/article/382047/prwtoboylies-tha-analhfthoyn-gia-thn-anadeixh-ths-palias-polhs

print (df.loc[df['name'] == "Lithos Cocktail Bar"])

print (df.loc[df['name'] == "Panorama Cafe Bar Rhodes"])

print (df.loc[df['name'] == "Παλιά Πόλη , Ρόδος"])

#dinoume to label sumfwna me to index toys - ola ta alla exoyn agnosto

df["label"]= 0

df["label"][455:456]= -1

df["label"][413:414]= -1

df["label"][91:92]= 1


print (df)
```

Υλοποίηση αλγορίθμου για το πείραμα

Παραμετροποιήσαμε τον τελευταίο αλγόριθμο ώστε να λάβουμε τα αποτελέσματα μας.

```
import pandas as pd
import numpy as np
from scipy.io import loadmat
from scipy.spatial.distance import pdist, squareform
from scipy.spatial import distance
import json
import pandas as pd

def Load_Data():
        #fortosi dedomenwn rodoy
        y=pd.read_pickle('dataset', compression='infer')
        return y


def Select_Data(n,y):
        #y=y.drop('price_range', axis=1)
        X = pd.concat([y['checkins'], y['overall_star_rating']], axis=1,
join_axes=[y['checkins'].index])
        X = pd.concat([X, y['rating_count']], axis=1, join_axes=[X.index])
        y0=y["label"]*1
        X=X.as_matrix()#metatrepoyme se numpy matrix
        return X,y0
def calculate_W(X):
        #pairwise_sq_dists = squareform(pdist(X, 'sqeuclidean'))
        #W = np.exp(-((pairwise_sq_dists))/( 2* (np.std( np.linalg.norm(X,
axis=1))**2)))
```

```
    W=distance.cdist(X, X, 'euclidean')
    W = np.exp(-(np.power(W,2))/( 2*(np.std( np.linalg.norm(X, axis=1))**2)))
    W[np.diag_indices_from(W)] = 0
    return W
def calculate_D(n,W):
    D=np.zeros((n, n),float)
    Diagwnios= np.array(W.sum(axis=1))#Wi,j sum
    np.fill_diagonal(D,Diagwnios)
    return D


def Dataframe_short(y):
    y = [y.loc[y['label'] == 1],y.loc[y['label'] == -1],y.loc[(y['label'] != 1) & (y['label']
!= -1 )]]
    y = pd.concat(y)
    y=y.reset_index(drop=True)
    return y


def estimate_L(W,D,n):
    di = np.diag_indices(n)
    D[di]=np.float_power(D[di],-0.5)
    L=np.dot(W,D)
    L=np.dot(D,L)
    return L


#139 / 1591 / 2011



a=0.5
y=Load_Data()
```

```
print (y)

print (y.loc[y['checkins'].idxmax()])

print (y.loc[y['rating_count'].idxmax()])

print (y.loc[y['overall_star_rating'].idxmax()])

y=y.fillna(0)

n=len(y.index)


y=Dataframe_short(y)#sortaroume ta stoixeia prwta 1 meta -1 kai telos ta midenika poy dimourgisame



y_mirror=y*1

X,y0=Select_Data(n,y)

W=calculate_W(X)

D=calculate_D(n,W)

L=estimate_L(W,D,n)




yt=np.matrix(np.array(y0[0:n]))

yt= (yt.T)

statheros_oros1=np.dot(L,a)


statheros_oros2=np.dot(yt,0.5)


e_check=1
```

```
while e_check>0.000000000000001 and not np.isnan(e_check):
        y1=np.add(np.dot(statheros_oros1,yt),statheros_oros2)
        e_check=np.abs(np.sum(np.subtract(y1,yt)))
        yt=y1*1.



labels = (yt > 0).astype(int)

y["prediction_label"]=labels
y_=y.where(y["prediction_label"]>0)

print (y_.dropna())
```

Αποτελέσματα

Αρχικά παρουσιάζουμε τα ανώτατα αποτελέσματα.

- Για το πλήθος των δημοσιεύσεων της παρουσίας των χρηστών σε τοποθεσίας, η «Rodos» έρχεται πρώτη.
- Για το πλήθος αυτών που αξιολόγησαν η εγγραφή «Elli Beach» έρχεται πρώτη.
- Ενώ για την αξιολόγηση της τοποθεσίας έχουμε πολλά διαφορετικά με την μέγιστη τιμή 5 αστέρια.

Περιγραφική στατιστική ομοιογενειοποιημένου δείγματος

	checkins	overall_star_rating	rating_count
count	674.000000	569.000000	674.000000
mean	0.009844	0.955747	0.051343
std	0.053817	0.073558	0.088704
min	0.000000	0.200000	0.000000
25%	0.000067	0.940000	0.003752
50%	0.000888	0.980000	0.018762
75%	0.003876	1.000000	0.058161
max	1.000000	1.000000	1.000000

Παρατηρούμε ότι ο μέσος όρος των checkins είναι αρκετά χαμηλός σε σχέση με τις ακραίες θετικές τιμές και η τυπική απόκλιση πολύ μεγάλη σε σχέση με τον μέσο όρο. Όπως αυτό φαίνεται και από τα τεταρτημόρια. Αντίθετα η αξιολόγηση των περιοχών έχει μέσο όρο πολύ υψηλό και μικρή τυπική απόκλιση, που σημαίνει ότι πολλές περιοχές αξιολογούνται θετικά. Οι εγγραφές με missing value θεωρήθηκαν μηδενικές. Όσον αφορά το πλήθος των αξιολογήσεων ο μέσος όρος βρίσκεται στο 5% της περιοχής με τις περισσότερες αξιολογήσεις.

Εν συνέχεια με θετικό πρότυπο – ετικέτα την Παλιά Πόλη της Ρόδου και αρνητικό δύο τυχαία δείγματα από τις τελευταίες βαθμολογικά καταχωρήσεις του tripadvisor τρέξαμε τον παρακάτω κώδικα - πείραμα.

```
 import pandas as pd
import numpy as np
from scipy.io import loadmat
from scipy.spatial.distance import pdist, squareform
from scipy.spatial import distance
import json
import pandas as pd


def Load_Data():
        #fortosi dedomenwn rodoy
        y=pd.read_pickle('dataset', compression='infer')
        return y


def Select_Data(n,y):
        #y=y.drop('price_range', axis=1)
        X = pd.concat([y['checkins'], y['overall_star_rating']], axis=1,
join_axes=[y['checkins'].index])
        X = pd.concat([X, y['rating_count']], axis=1, join_axes=[X.index])
        y0=y["label"]*1
        X=X.as_matrix()#metatrepoyme se numpy matrix
        return X,y0
def calculate_W(X):
        #pairwise_sq_dists = squareform(pdist(X, 'sqeuclidean'))
        #W = np.exp(-((pairwise_sq_dists))/( 2* (np.std( np.linalg.norm(X,
axis=1))**2)))
        W=distance.cdist(X, X, 'euclidean')
        W = np.exp(-(np.power(W,2))/( 2*(np.std( np.linalg.norm(X, axis=1))**2)))
        W[np.diag_indices_from(W)] = 0
        return W
```

```
def calculate_D(n,W):
        D=np.zeros((n, n),float)
        Diagwnios= np.array(W.sum(axis=1))#Wi,j sum
        np.fill_diagonal(D,Diagwnios)
        return D


def Dataframe_short(y):
        y = [y.loc[y['label'] == 1],y.loc[y['label'] == -1],y.loc[(y['label'] != 1) & (y['label']
!= -1 )]]
        y = pd.concat(y)
        y=y.reset_index(drop=True)
        return y


def estimate_L(W,D,n):
        di = np.diag_indices(n)
        D[di]=np.float_power(D[di],-0.5)
        L=np.dot(W,D)
        L=np.dot(D,L)
        return L


#139 / 1591 / 2011



a=0.5
y=Load_Data()
print (y)
k= (y.describe())
pd.k.to_csv("descriptive.csv")
```

```
print (y.loc[y['checkins'].idxmax()])

print (y.loc[y['rating_count'].idxmax()])

print (y.loc[y['overall_star_rating'].idxmax()])

y=y.fillna(0)

n=len(y.index)


y=Dataframe_short(y)#sortaroume ta stoixeia prwta 1 meta -1 kai telos ta midenika poy dimourgisame



y_mirror=y*1

X,y0=Select_Data(n,y)

W=calculate_W(X)

D=calculate_D(n,W)

L=estimate_L(W,D,n)




yt=np.matrix(np.array(y0[0:n]))

yt= (yt.T)

statheros_oros1=np.dot(L,a)


statheros_oros2=np.dot(yt,0.5)


e_check=1

while e_check>0.000000000000001 and not np.isnan(e_check):
```

```
y1=np.add(np.dot(statheros_oros1,yt),statheros_oros2)
e_check=np.abs(np.sum(np.subtract(y1,yt)))
yt=y1*1.
```

```
labels = (yt > 0).astype(int)

y["prediction_label"]=labels
y_=y.where(y["prediction_label"]>0)

print (y_.dropna())
```

Ο οποίος μας έδωσε τα αποτελέσματα του παρακάτω πίνακα.

A/A	checkins	name	overall star rating	price range	rating count	Start label
1	0.075899454	Παλιά Πόλη , Ρόδος	0.94	nan	0.42682927	1
2	0.159655038	Gazi Club ΡΟΔΟΣ	0.94	$	0.40712946	0
3	0.258521054	OLA ελληνικά Ρόδος	0.96	$	0.55065666	0
4	0.060907661	Macao Lounge Bar	0.94	$$$	0.44090056	0
5	0.275726023	Barra Tres Rhodes	0.94	$	0.44934334	0
6	0.020356481	Tamam Restaurant	0.98	$$	0.62851782	0
7	0.177906197	Elli Beach Rhodes	0.9	nan	1	0
8	0.003738241	Rhodes By Night	0.96	$	0.55628518	0

ΑΝΑΛΥΣΗ ΑΠΟΤΕΛΕΣΜΑΤΩΝ

Έχουμε λοιπόν 8 εγγραφές από το σύνολο των δεδομένων που τελικά πήραν την ετικέτα της δημοφιλούς περιοχής – επιχείρησης όπως αυτή της Παλιάς Πόλης της Ρόδου.

Η Παλιά Πόλη με 7,5% του πλήθους των δημοσιεύσεων τοποθεσιών, με 94% ως προς την άριστη αξιολόγηση και με 42% ως προς το μεγαλύτερο πλήθος τοποθεσίας ξαναπήρε ετικέτα με +1. Αυτό το πρότυπο μας, μας δείχνει ότι έχουμε πάρει μια έγγραφή η οποία δεν έχει να κάνει καθαρά και μόνο με την ποσότητα αλλά υπάρχει μια σχέση ποσότητας και ποιότητας.

Οι επιχειρήσεις υγειονομικού ενδιαφέροντος όπως το $2^{η}$, η $3^{η}$, $4^{η}$ και η $5^{η}$ εγγραφή οφείλονται στην τοπική κοινωνία και είναι ιδιαίτερα δημοφιλείς σε αυτήν. Για όλες αυτές έχουν αξιολογήσει πάνω από το 40% του μέγιστου πλήθους αξιολόγησης τοποθεσίας-επιχειρήσης.

Το νούμερο 6. Η καταχώρηση του Tamam πρόκειται για μια περισσότερο τουριστικά προσανατολισμένη επιχείρηση υγειονομικού ενδιαφέροντος – εστιατόριο. Παρατηρούμε ότι από την 8άδα των αποτελεσμάτων έχει την καλύτερη αξιολόγηση και τους περισσότερους αξιολογητές παρόλο που στην δημοσίευση παρουσίας χρήστη (checkin) βρίσκεται στο χαμηλότερο επίπεδο.

Το Elli Beach, η εγγραφή με τους περισσότερους αξιολογητές, έχει ικανοποιητικό αριθμό checkins, 17% σε σχέση με την εγγραφή με τα περισσότερα checkins. Έχει όμως την χαμηλότερη αξιολόγηση από τις 8 εγγραφές, η οποία και πάλι είναι σχετικά υψηλή. Αυτό μας δείχνει ότι συγκαταλέγεται σε ένα από τους έγκυρα δημοφιλέστερους προορισμούς στην Ρόδο.

Η συμπλήρωση της 8άδας γίνεται με την εγγραφή Rhodes by night. Πρόκειται για μια τοπική επιχείρηση που έχει σκοπό την προβολή της νυκτερινής διασκέδασης της Ρόδου. Κατάφερε να εισέλθει στις εγγραφές με ετικέτα +1 καθώς έχει υψηλή αξιολόγηση και μεγάλο πλήθος αξιολογητών. Πρόκειται για μια εγγραφή που βοηθάει στην προβολή της Ρόδου και στην αύξηση της δημοφιλίας της με τις ενέργειες τις.

Γενικότερα παρατηρούμε ότι δεν υπάρχουν εγγραφές υψηλού κόστους όπως έχουν οριστεί μέσα από το κοινωνικό δίκτυο.

Συμπεράσματα

Για την παρουσίαση των τοποθεσιών και επιχειρήσεων που συμβάλουν στην κοινωνικοοικονομική ανάπτυξη της Ρόδου μέσα από τη προβολή στο νούμερο ένα κοινωνικό δίκτυο στο κόσμο κάποιες μορφές ανάλυσης. Για να βρούμε ένα αξιόπιστο μοντέλο κάναμε επικύρωση μέσω δευτερογενών δεδομένων σε κάποιους δημοσιευμένους αλγορίθμους. Αυτοί οι αλγόριθμοι είχαν σαν βάση την ανάλυση με ημι-επιτηρούμενη μάθηση αλγορίθμων. Έχοντας δηλαδή κάποιες τιμές, προσεγγίζουμε τις τιμές των υπολοίπων εγγραφών του δείγματος μας.

Συμπεραίνουμε λοιπόν ότι:

- Υπάρχει μεγάλη ανησυχία στην επιστημονική κοινότητα για το θέμα των αλγόριθμων μηχανικής μάθησης
- Ο αλγόριθμος Μετάδοσης – Zhou et al. [2004] είναι ένα αξιόπιστο μοντέλο για την ανάλυση μας με πάνω από 75% ορθά αποτελέσματα.
- Σύμφωνα με αυτό τον αλγόριθμο, από τα δεδομένα του κοινωνικού δικτύου της Facebook Inc, υπάρχουν 7 τοποθεσίες – επιχειρήσεις στο ύψος της δημοτικότητας της παγκόσμιας πολιτιστικής κληρονομιάς (Παλιά Πόλη), των οποίων η δημοτικότητα οφείλεται στη ύπαρξη της τοπικής κοινωνίας αλλά και των παραθεριστών. Θα έπρεπε να εξεταστούν και υιοθετηθούν πρακτικές που εφαρμόζονται με γνώμονα αυτές.

Αυτό που γνωρίζουμε είναι ότι υπάρχει περιθώριο αύξησης της δημοτικότητας καθώς υπάρχουν και άλλα μέρη των οποίων οι εγγραφές στο κοινωνικό δίκτυο δεν προωθούνται τόσο πολύ. Είναι σημαντικό να αναγνωριστεί η προσπάθεια μιας επιχείρησης όπως το Rhodes by night να αναδείξει την διασκέδαση στο νησί. Επίσης, να αναγνωριστεί ότι ακόμα μια επιχείρηση ξεπερνάει με την δημοτικότητα της τοπικές επιχειρήσεις (βλέπε Tamam). Ίσως θα πρέπει να υιοθετηθούν οι πρακτικές της και από άλλους οργανισμούς - επιχειρήσεις.

Όσον αφορά το Elli beach Rhodes και την Παλιά Πόλη πρόκειται για τοποθεσίες όπου αφορούν όλους μας και θα έπρεπε να βρίσκονταν και άλλα μέρη όπως αυτά σε τόσο υψηλό επίπεδο δημοτικότητας.

Για περαιτέρω έρευνα θα ήταν φρόνιμο να χωρίζαμε σε κατηγορίες τις εγγραφές, να χρησιμοποιούσαμε περισσότερες εγγραφές και να τμηματοποιήσουμε σε περισσότερες από δύο κατηγορίες τις ετικέτες μας ώστε να εξάγουμε αναλυτικότερα συμπεράσματα. Επιπρόσθετα θα μπορούσαμε να αντλήσουμε δεδομένα και από περισσότερες από μία πηγές.

Επίσης θα ήταν καλό να γινόντουσαν πειραματισμοί σε διάφορους αλγορίθμους και να γινόταν μια προσπάθεια δημιουργίας νέων.

ΒΙΒΛΙΟΓΡΑΦΙΑ

Label Propagation and Quadratic Criterion, Yoshua Bengio, Olivier Delalleau ,Nicolas Le Roux,2006

SVM-Based Sentiment Classification: A Comparative, Study against State-of-the-Art Classifiers, Dionisios N. Sotiropoulos , Demitrios E. Pournarakis , George M. Giaglis , 2017

What drives social sentiment? An entropic measure-based clustering approach towards identifying factors that influence social sentiment polarity, Dionisios N. Sotiropoulos, Chris D. Kounavis, George M. Giaglis and Panos Kourouthanassis, Department of Management Science and Technology, Athens University of Economics and Business ,Department of Informatics, Ionian University,2014

Chapelle, Olivier; Schölkopf, Bernhard; Zien, Alexander (2006). Semi-supervised learning. Cambridge, Mass.: MIT Press. ISBN 978-0-262-03358-9.

Stevens, K.N.(2000), Acoustic Phonetics, MIT Press, ISBN 0-262-69250-3, 978-0-262-69250-2

Scudder, H.J. Probability of Error of Some Adaptive Pattern-Recognition Machines. IEEE Transaction on Information Theory, 11:363–371 (1965). Cited in Chapelle et al. 2006, page 3.

Vapnik, V. and Chervonenkis, A. Theory of Pattern Recognition [in Russian]. Nauka, Moscow (1974). Cited in Chapelle et al. 2006, page 3.

Ratsaby, J. and Venkatesh, S. Learning from a mixture of labeled and unlabeled examples with parametric side information. In Proceedings of the Eighth Annual Conference on Computational Learning Theory, pages 412-417 (1995). Cited in Chapelle et al. 2006, page 4.

ΔΙΑΔΙΚΤΥΑΚΗ

«Διαχείριση της πολιτιστικής κληρονομιάς και τουρισμός: το παράδειγμα της μεσαιωνικής πόλης της Ρόδου» | Η ΡΟΔΙΑΚΗ http://www.rodiaki.gr/article/322858/diaxeirish-ths-politistikhs-klhronomias-kai-toyrismos-to-paradeigma-ths-mesaiwnikhs-polhs-ths-rodoy**,** Ψαρρή Παναγιώτα, Αρχαιολόγος Εφορεία Αρχαιοτήτων Δωδεκανήσου, 16/10/15