Πρόβλεψη του Θέματος Μικρών Κειμένων σε Διαδικτυακά Κοινωνικά Δίκτυα Βασισμένα σε Τοποθεσία

Η ΜΕΤΑΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ ΕΞΕΙΔΙΚΕΥΣΗΣ

υποβάλλεται στην

ορισθείσα από την Γενική Συνέλευση Ειδικής Σύνθεσης

του Τμήματος Μηχανικών Η/Υ και Πληροφορικής Εξεταστική Επιτροπή

από τον

Ιωάννη Κοτρώτσιο

ως μέρος των Υποχρεώσεων για τη λήψη του

ΜΕΤΑΠΤΥΧΙΑΚΟΥ ΔΙΠΛΩΜΑΤΟΣ ΣΤΗΝ ΠΛΗΡΟΦΟΡΙΚΗ

ΜΕ ΕΞΕΙΔΙΚΕΥΣΗ

ΣΤΟ ΛΟΓΙΣΜΙΚΟ

Φεβρουάριος 2015

Αφιέρωση

Στην οικογένεια μου και σε όλους όσους στέκονται δίπλα μου και με στηρίζουν με οποιονδήποτε τρόπο...

Ευχαριστίες

Θα ήθελα να ευχαριστήσω θερμά την επιβλέπουσα μου, καθηγήτρια κ. Ευαγγελία Πιτουρά για την καθοδήγηση, τις γνώσεις και την βοήθεια που μου προσέφερε, το χρόνο που διέθεσε και κυρίως την υπομονή που έδειξε σε όλη τη διάρκεια της συνεργασίας μας.

Επίσης θα ήθελα να ευχαριστήσω τον συν-επιβλέποντά μου, επίκουρο καθηγητή κ. Παναγιώτη Τσαπάρα για τη βοήθεια και την παροχή γνώσεων και ιδεών κατά τη διάρκεια της συνεργασία μας, καθώς και τον αναπληρωτή καθηγητή κ. Παναγιώτη Βασιλειάδη για τις χρήσιμες παρατηρήσεις του σχετικά με την εργασία αυτή.

Τέλος, θα ήθελα να ευχαριστήσω τους συναδέλφους και φίλους μου Κώστα Σεμερτζίδη, Πέτρο Μανούση και Πάρη Τσανταρλιώτη για την υποστήριξη τους κατά τη διάρκεια των μεταπτυχιακών μου σπουδών.

Περιεχόμενα

Ευρετήριο Σχημάτων

Ευρετήριο Πινάκων

Επεξηγήσεις Συμβολισμών

ΚΔΒΤ: Κοινωνικά Δίκτυα Βασισμένα σε Τοποθεσία

API: Application Programming Interface
JSON: JavaScript Object Notation
LBSN: Location-Based Social Networks
LDA: Latent Dirichlet Allocation

Περίληψη

Ιωάννης Κοτρώτσιος του Παντελή και της Χρυσούλας. MSc, Τμήμα Μηχανικών Η/Υ και Πληροφορικής, Πανεπιστήμιο Ιωαννίνων, Φεβρουάριος, 2015. Πρόβλεψη του Θέματος Μικρών Κειμένων σε Διαδικτυακά Κοινωνικά Δίκτυα Βασισμένα σε Τοποθεσία. Επιβλέπουσα: Ευαγγελία Πιτουρά.

Τα διαδικτυακά κοινωνικά δίκτυα έχουν γίνει ιδιαίτερα δημοφιλή τα τελευταία χρόνια, με πολύ μεγάλο ποσοστό ανθρώπων να τα χρησιμοποιεί καθημερινά, οπουδήποτε και αν βρίσκεται και οποιαδήποτε ώρα της ημέρας. Οι διαδικτυακές αυτές πλατφόρμες επιτρέπουν την αλληλεπίδραση μεταξύ των χρηστών και προσφέρουν μία σειρά από βασικές και συνήθως δωρεάν υπηρεσίες. Με την ανάπτυξη των έξυπνων κινητών συσκευών που διαθέτουν τεχνολογίες εντοπισμού θέσης μέσω χρήσης ασύρματων δικτύων και GPS και την ραγδαία εξάπλωσή τους, οι κλασσικές υπηρεσίες που παρέχουν τα διαδικτυακά κοινωνικά δίκτυα έχουν εμπλουτιστεί με το χαρακτηριστικό της τοποθεσίας με διάφορους τρόπους. Τα κοινωνικά δίκτυα που εμπεριέχουν την έννοια της τοποθεσίας και επίσης η αλληλεπίδραση των χρηστών τους εξαρτάται από την τοποθεσία ονομάζονται *Location-Based Social Networks (LBSN)*. Λόγω της φύσης τους αποτελούν μια πολύ καλή πηγή ολοένα αυξανόμενων δεδομένων που μπορούν να χρησιμοποιηθούν για μελέτη πληθώρας ερευνητικών θεμάτων, από απλών μέχρι αρκετά σύνθετων.

Στην εργασία αυτή ερευνούμε τα μικρά κείμενα (tips) που αφήνουν οι χρήστες στις τοποθεσίες που επισκέπτονται προκειμένου να τις αξιολογήσουν και να μοιραστούν τις εμπειρίες τους με τους υπόλοιπους χρήστες. Μελετάμε τα ιδιαίτερα χαρακτηριστικά τους και τα χρησιμοποιούμε για να αναλύσουμε τις σχέσεις μεταξύ χρηστών και τοποθεσιών. Εξετάζουμε αν είναι εφικτό με το περιεχόμενο αυτό και με διάφορες προσεγγίσεις να προβλέψουμε το περιεχόμενο (θεματολογία) ενός κειμένου που θα αφήσει ένας δεδομένος χρήστης σε μια δεδομένη τοποθεσία. Προσεγγίζουμε το πρόβλημα με χρήση Πιθανολογικών Θεματικών Μοντέλων αλλά και με τη χρήση Κατηγοριοποιητών και στη συνέχεια εφαρμόζουμε τις προσεγγίσεις μας σε πραγματικά δεδομένα που έχουμε συλλέξει από το δίκτυο του Foursquare. Αξιολογούμε τα ποσοστά επιτυχίας τους και δείχνουμε ότι η πρόβλεψη θέματος είναι εφικτή ως ένα βαθμό, ειδικά αν συνυπολογίσουμε την ιδιαιτερότητα των δεδομένων μας.

Abstract in English

Kotrotsios P. Ioannis, MSc, Department of Computer Science and Engineering, University of Ioannina, Greece. February, 2015. Topic Prediction for Small Texts in Location-Based Online Social Networks. Thesis Supervisor: Evaggelia Pitoura.

Online Social Networks have experienced a great growth in popularity in the last years. Most people use them every day, in any place and time of the day. These online platforms allow users to interact with each other by providing a lot interesting services and features, usually free of charge.

With the evolution of smart mobile devices that incorporate position aquisition technologies by using wireless networks and GPS receivers, the ordinary services provided by online social networks have been enriched with location-relevant features.

The social networks that incorporate the sense of location and the interaction between their users depend on the location, are known as *Location-Based Social Networks (LBSN)*. These networks, due to their nature, constitute a great source of endlessly growing data that can be used for a variety of scientific research reasons, from very simple to very complex.

In this work, we make use of small texts called 'tips', that social network users leave to the venues that they visit. We study their characteristics and use them to analyse the relationship between the entities of these social networks, users and venues. We examine if it is possible to predict the content (in terms of topic) of the tip that a given user leave to a given venue. We approach the problem by using Topic Modeling tools and Classification methods. We collect real data from the Foursquare Social Network and evaluate experimentally our approaches in order to find out if it is possible to make such a prediction, taking into account the peculiarity of the data.

Κεφάλαιο 1

Εισαγωγή

1.1 Κοινωνικά Δίκτυα Βασισμένα σε Τοποθεσία

Τα διαδικτυακά κοινωνικά δίκτυα, όπως για παράδειγμα το Facebook[1] και το Twitter[2] έχουν γίνει ιδιαίτερα δημοφιλή τα τελευταία χρόνια, με πολύ μεγάλο ποσοστό ανθρώπων να τα χρησιμοποιεί καθημερινά, οπουδήποτε και αν βρίσκεται και οποιαδήποτε ώρα της ημέρας. Στην πραγματική ζωή τα κοινωνικά δίκτυα είναι σύνολα αλληλεπιδράσεων και διαπροσωπικών σχέσεων. Είναι δομές που αποτελούνται από άτομα που συνδέονται μεταξύ τους για διάφορους λόγους όπως για παράδειγμα φιλία, κοινά ενδιαφέροντα ή κοινή γνώση.

Στις μέρες μας όταν αναφερόμαστε στα κοινωνικά δίκτυα εννοούμε σχεδόν πάντα τις διαδικτυακές πλατφόρμες που επιτρέπουν την αλληλεπίδραση μεταξύ των χρηστών και προσφέρουν μία σειρά από βασικές και συνήθως δωρεάν υπηρεσίες όπως τη δημιουργία προφίλ, τον διαμοιρασμό εικόνων και βίντεο, τον σχολιασμό σε ενέργειες που γίνονται από άλλα μέλη του δικτύου ή μίας ομάδας, την άμεση ανταλλαγή μηνυμάτων και πολλά άλλα.

[1] www.facebook.com

[2] www.twitter.com

Με την ανάπτυξη των έξυπνων κινητών συσκευών που διαθέτουν τεχνολογίες εντοπισμού θέσης μέσω χρήσης ασύρματων δικτύων και GPS και την ραγδαία εξάπλωσή τους, οι κλασσικές υπηρεσίες που παρέχουν τα διαδικτυακά κοινωνικά δίκτυα έχουν εμπλουτιστεί με το χαρακτηριστικό της τοποθεσίας με διάφορους τρόπους. Για παράδειγμα οι χρήστες μπορούν να κοινοποιήσουν την τοποθεσία τους, να προσθέσουν πληροφορίες τοποθεσίας στις φωτογραφίες και το κείμενο που μοιράζονται, να καταγράψουν τα μέρη που επισκέπτονται και τις δραστηριότητες τους και πολλά άλλα.

Η τοποθεσία μπορεί να αναπαρασταθεί είτε με τις ακριβείς συντεταγμένες (γεωγραφικό μήκος και πλάτος), είτε με τοποθεσίες (venues) των οποίων η θέση είναι ήδη καταγεγραμμένη στο κοινωνικό δίκτυο. Με την ευρύτερη έννοια, η τοποθεσία θα μπορούσε να είναι και μια σειρά από τοποθεσίες ή να περιέχει και χρονική πληροφορία, για παράδειγμα όταν αναφερόμαστε σε ένα δρώμενο (event). Το χαρακτηριστικό της τοποθεσίας έρχεται για να γεφυρώσει το κενό ανάμεσα στον πραγματικό κόσμο και τα διαδικτυακά κοινωνικά δίκτυα. Για παράδειγμα, ένας χρήστης που βρίσκεται σε ένα εστιατόριο έχει τη δυνατότητα με το έξυπνο κινητό του τηλέφωνο να αφήσει ένα μικρό κείμενο (tip) για την εμπειρία του από το συγκεκριμένο εστιατόριο σε ένα δίκτυο όπως το Foursquare[3] που του δίνει αυτή τη δυνατότητα. Έτσι οι φίλοι του και γενικότερα χρήστες του ίδιου δικτύου έχουν τη δυνατότητα να δουν το σχόλιο αυτό όταν βρεθούν στο ίδιο εστιατόριο ή ακόμα και κοντά σε αυτό.

Η προσθήκη της τοποθεσίας σε ένα κοινωνικό δίκτυο δεν είναι απλά μία επιπλέον πληροφορία περιεχομένου ανάμεσα στις υπόλοιπες, αλλά δημιουργεί και ένα νέο τύπο αλληλεπίδρασης μεταξύ των χρηστών. Η πληροφορία της τοποθεσίας μπορεί να μας δώσει πολύτιμη γνώση σχετικά με τα ενδιαφέροντα και τη συμπεριφορά του κάθε χρήστη. Στο παραπάνω παράδειγμα με το εστιατόριο, άνθρωποι που το επισκέπτονται και στα σχόλια τους αναφέρουν τα ίδια θέματα θα μπορούσαμε να πούμε ότι συνδέονται μεταξύ τους και πιθανόν θα μπορούσαν να είναι φίλοι και στην πραγματική ζωή. Άλλο παράδειγμα με κοινά ενδιαφέροντα θα μπορούσαν να είναι χρήστες που τους αρέσει η ορειβασία και ανεβαίνουν στο ίδιο βουνό. Η τοποθεσία επίσης δεν είναι ανάγκη να είναι ακριβής, αλλά να είναι συγκεκριμένος τύπος τοποθεσίας. Για παράδειγμα ιταλικά εστιατόρια, γυμναστήρια, παραλίες, πάρκα κτλ. Επιπρόσθετα, σημαντική πληροφορία μπορεί να μας δώσει και η αλληλουχία από τοποθεσίες ενός χρήστη σε συνδυασμό με το περιεχόμενο που δημιουργεί κατά τη μετακίνηση του αυτή.

Τα κοινωνικά δίκτυα που έχουν τα παραπάνω χαρακτηριστικά, δηλαδή εμπεριέχουν την έννοια της τοποθεσίας και επίσης η αλληλεπίδραση των χρηστών τους εξαρτάται από την τοποθεσία, ονομάζονται *Κοινωνικά Δίκτυα Βασισμένα σε Τοποθεσία (Location-Based Social Networks)* [17]. Θα αναφερόμαστε σε αυτά με την συντομογραφία ΚΔΒΤ (LBSN). Τα δίκτυα

[3] www.foursquare.com

αυτά λόγω της φύσης τους αποτελούν μια πολύ καλή πηγή ολοένα αυξανόμενων δεδομένων που μπορούν να χρησιμοποιηθούν για μελέτη πληθώρας ερευνητικών θεμάτων, από απλών μέχρι αρκετά σύνθετων.

Τα είδη των Κοινωνικών Δικτύων Βασισμένων σε Τοποθεσία που υπάρχουν αυτή τη στιγμή θα μπορούσαμε να τα κατατάξουμε σύμφωνα με το [18] σε τρεις γενικές κατηγορίες ως εξής:

- *Εστιασμένα στο περιεχόμενο με γεωγραφική πληροφορία:* Τα δίκτυα αυτού του είδους επιτρέπουν στους χρήστες να προσθέτουν πληροφορίες τοποθεσίας στο περιεχόμενο που δημιουργούν, όπως για παράδειγμα φωτογραφίες, βίντεο και κείμενο. Αυτό μπορεί να γίνει είτε απευθείας με την προσθήκη του περιεχομένου όταν γίνεται από κινητή συσκευή είτε σε δεύτερο χρόνο. Έτσι οι χρήστες έχουν τη δυνατότητα να βλέπουν το περιεχόμενο στην τοποθεσία που δημιουργήθηκε. Αυτό μπορεί να αναπαρασταθεί πάνω σε ένα χάρτη ή αν βρεθούν στο συγκεκριμένο σημείο με μια κινητή συσκευή.

  Αντιπροσωπευτικό παράδειγμα τέτοιου δικτύου είναι το Google Panoramio[4] όμως ολοένα και περισσότερα δίκτυα υιοθετούν την δυνατότητα προσθήκης τοποθεσίας με κυριότερα παραδείγματα το Twitter, το Facebook, Instagram[5] και Flickr[6]. Ωστόσο, το επίκεντρο του ενδιαφέροντος σε αυτές τις υπηρεσίες παραμένει το περιεχόμενο.

- *Εστιασμένα στην τοποθεσία:* Αυτές οι υπηρεσίες επιτρέπουν στο χρήστη να κοινοποιεί (check-in) την τρέχουσα τοποθεσία του, συσχετίζοντας τη με ένα συγκεκριμένο μέρος, όπως ένα εστιατόριο, ένα γήπεδο, ένα κατάστημα κτλ. Κύριος εκπρόσωπος τέτοιας υπηρεσίας είναι το Foursquare στο οποίο θα εστιάσουμε αρκετά σε αυτή την εργασία.

  Το Foursquare ξεκίνησε με έντονο το στοιχείο του παιχνιδιού, προτρέποντας τους χρήστες να κάνουν check-in στις τοποθεσίες (venues) που βρίσκονται προκειμένου να κερδίσουν πόντους (points) και κονκάρδες (badges), ενώ αν κάποιος είναι συχνός θαμώνας μιας τοποθεσίας μπορεί να ανακηρυχθεί δήμαρχος (mayor) της τοποθεσίας αυτής. Οι χρήστες μπορούν σε πραγματικό χρόνο να βλέπουν που βρίσκονται οι φίλοι τους αλλά και αν υπάρχουν άλλοι χρήστες του δικτύου στην τοποθεσία που βρίσκονται εκείνη τη στιγμή. Από τις σημαντικότερες δυνατότητες του Foursquare είναι ότι οι χρήστες μπορούν να αφήσουν μικρά σχόλια (tips) στις τοποθεσίες που επισκέπτονται προκειμένου να μοιραστούν την εμπειρία τους και να αξιολογήσουν την κάθε τοποθεσία. Με τον καιρό, το Foursquare εξελίχθηκε σε μια πολύ δυνατή πλατφόρμα προτάσεων, σχεδόν εξαλείφο-

[4]www.panoramio.com
[5]www.instagram.com
[6]www.flickr.com

ντας το στοιχείο του παιχνιδιού. Είναι προφανές ότι στα δίκτυα αυτής της κατηγορίας το επίκεντρο του ενδιαφέροντος είναι η τοποθεσία και αυτή καθορίζει την αλληλεπίδραση των χρηστών.

- *Εστιασμένα στις τροχιές:* Αυτή η κατηγορία περιλαμβάνει κάπως λιγότερο δημοφιλείς υπηρεσίες, οι οποίες ολοένα και αυξάνονται, στις οποίες σημασία δεν έχουν μόνο οι τοποθεσίες, αλλά και οι διαδρομές που τις συνδέουν, όπως προκύπτουν από την κινητικότητα των χρηστών. Τέτοιου είδους δεδομένα για παράδειγμα μπορεί να παράγονται από χρήστες που ασχολούνται με αθλητικές δραστηριότητες όπως ποδηλασία, τρέξιμο, ορειβασία ή απλά ταξιδεύουν.

  Εκτός από τις πληροφορίες της διαδρομής όπως απόσταση, διάρκεια και ταχύτητα που καταγράφονται από τις κινητές συσκευές, οι χρήστες μπορούν να προσθέσουν κείμενο, εικόνα και βίντεο. Η πληροφορία της διαδρομής δίνει πλουσιότερη και πιο ”ζωντανή” εμπειρία και δυνατότητες στους υπόλοιπους χρήστες του δικτύου. Τέτοιου είδους δίκτυο είναι το Geolife project[7] της Microsoft αλλά και διάφορες υπηρεσίες καταγραφής αθλητικής δραστηριότητας που τείνουν να μετατραπούν σε κοινωνικά δίκτυα.

1.2 Στόχος της Εργασίας

Στην παρούσα εργασία θα επικεντρωθούμε στο δίκτυο του Foursquare και πιο συγκεκριμένα στην δυνατότητα που δίνει στους χρήστες του να αφήνουν μικρά κείμενα στις τοποθεσίες που επισκέπτονται. Τα μικρά αυτά κείμενα (μέχρι 200 χαρακτήρες) αποτελούν μίνι αξιολογήσεις για τις τοποθεσίες και είναι αρκετά σημαντικά για τους υπόλοιπους χρήστες. Για παράδειγμα, σχόλια για την αργή εξυπηρέτηση σε ένα εστιατόριο ή την καθαριότητα σε ένα ξενοδοχείο είναι πιθανό να επηρεάσουν σημαντικά τις επιλογές ενός χρήστη που θα τα διαβάσει. Τα μικρά αυτά κείμενα θα μπορούσαμε να πούμε ότι δημιουργούν μία σχέση ανάμεσα στους χρήστες και στις τοποθεσίες και ιδιαίτερα στις επιχειρήσεις. Οι τελευταίες μπορούν να επωφεληθούν από αυτά, από τη μία γιατί μπορούν να αποκομίσουν πληροφορίες για τα ενδιαφέροντα και τις προτιμήσεις των πελατών τους με αποτέλεσμα την βελτίωση των υπηρεσιών τους και από την άλλη γιατί μπορούν να αποτελέσουν σημαντικό μέσο για προσέλκυση νέων πελατών εφόσον καλά σχόλια λειτουργούν σαν διαφήμιση προς υποψήφιους χρήστες.

Από τα παραπάνω προκύπτει η ανάγκη για εύκολη λήψη από τον χρήστη της πληροφορίας από τα μικρά αυτά κείμενα. Προβλήματα που υπάρχουν συχνά είναι η παρουσία μεγάλου αριθμού από σχόλια μέσα στα οποία κρύβονται οι χρήσιμες πληροφορίες. Επίσης συχνή είναι

[7] http://research.microsoft.com/en-us/projects/geolife/

η παρουσία ανεπιθύμητων, διαφημιστικών ή παραπλανητικών σχολίων. Η δυνατότητα αυτοματοποιημένης εύρεσης θεμάτων μέσα από τα κείμενα αυτά θα διευκόλυνε πολύ στην αντιμετώπιση των προβλημάτων αυτών.

Η εύρεση θέματος (topic identification) [35, 36] είναι η διαδικασία αναγνώρισης των κεντρικών ιδεών μέσα από κείμενα. Η εύρεση θέματος είναι χρήσιμη για πολλές εφαρμογές, ιδιαίτερα στα κοινωνικά δίκτυα που η ποσότητα της πληροφορίας είναι πολύ μεγάλη και αυξάνεται διαρκώς. Μερικές ενδιαφέρουσες εφαρμογές είναι η περίληψη περιεχομένου [39], η εξαγωγή ενδιαφερόντων για χρήστες και παροχή εξατομικευμένων προτάσεων και περιεχομένου [8, 9], αλλά και η χρήση τους από επιχειρήσεις για βελτίωση των υπηρεσιών τους και παροχή στοχευμένης διαφήμισης σε κοινωνικά δίκτυα.

Σε σύγκριση με τα υπόλοιπα συστήματα αξιολογήσεων τοποθεσιών όπως το TripAdvisor[8] και το Yelp[9] στα οποία οι χρήστες αφήνουν πιο ολοκληρωμένες αξιολογήσεις, τα σχόλια του Foursquare είναι πολύ πιο συνοπτικά και περιέχουν πιο υποκειμενικό και "ανεπίσημο" περιεχόμενο. Επιπλέον, οι χρήστες συνήθως όταν αφήνουν σχόλια από τις κινητές συσκευές τους είναι πιο άμεσοι και αποφεύγουν να αναφέρουν λεπτομέρειες και συγκεκριμένα χαρακτηριστικά των τοποθεσιών που βρίσκονται. Οι προσεγγίσεις που θα ασχοληθούμε σε αυτή την εργασία και κυρίως τα θεματικά μοντέλα θα μπορούσαν να εφαρμοστούν και στις ολοκληρωμένες αξιολογήσεις του TripAdvisor και του Yelp, ωστόσο αναμένουμε από τα σχόλια να είναι πιο ειλικρινή και αυθόρμητα οπότε και η πληροφορία τους θα είναι πιο ενδιαφέρουσα.

Στην εργασία αυτή θα χρησιμοποιήσουμε τα μικρά κείμενα (tips) που αφήνουν οι χρήστες στις τοποθεσίες που επισκέπτονται. Θα αναλύσουμε τα χαρακτηριστικά τους και θα τα χρησιμοποιήσουμε για να αναλύσουμε τις τις σχέσεις μεταξύ χρηστών, τις σχέσεις μεταξύ τοποθεσιών, αλλά και μεταξύ χρήστη-τοποθεσίας. Θα μελετήσουμε διαφορετικές προσεγγίσεις για να προβλέψουμε το περιεχόμενο, σε επίπεδο θέματος, ενός μικρού κειμένου που θα αφήσει ένας δεδομένος χρήστης σε δεδομένη τοποθεσία. Θα προσεγγίσουμε το πρόβλημα με χρήση πιθανολογικών θεματικών μοντέλων αλλά και με τη χρήση κατηγοριοποιητών και θα δούμε πειραματικά αν μπορεί να γίνει μια τέτοια πρόβλεψη δεδομένης της ιδιαιτερότητας των δεδομένων και πόσο καλή μπορεί να χαρακτηριστεί.

[8] www.tripadvisor.com

[9] www.yelp.com

1.3 Δομή της Εργασίας

Η εργασία αυτή αποτελείται από 5 Κεφάλαια. Στο Κεφάλαιο αυτό κάναμε μια εισαγωγή στα κοινωνικά δίκτυα βασισμένα σε τοποθεσία, στα χαρακτηριστικά τους και στις ευκαιρίες που μας παρέχουν για να τα αναλύσουμε. Επίσης είδαμε τα αντικείμενα και τους στόχους της παρούσας εργασίας.

Στο Κεφάλαιο 2 θα δούμε σχετικές εργασίες που έχουν εκπονηθεί πάνω σε δεδομένα κοινωνικών δικτύων και ιδιαίτερα με χρήση κειμένου, όπως για παράδειγμα συστήματα συστάσεων, ανάθεση ετικετών σε πόρους (resource tagging) και πιθανολογικά θεματικά μοντέλα (probabilistic topic models).

Στο Κεφάλαιο 3 θα αναλύσουμε το πρόβλημα της πρόβλεψης της θεματολογίας των κειμένων που αφήνουν οι χρήστες σε τοποθεσίες. Θα προσεγγίσουμε το πρόβλημα με τη χρήση θεματικών μοντέλων καθώς και με χρήση κατηγοριοποιητών (classifiers). Θα παρουσιάσουμε επίσης τα αντίστοιχα εργαλεία και μετρικές.

Στο Κεφάλαιο 4 θα δούμε Πειραματικά Αποτελέσματα από τις προσεγγίσεις που αναφέραμε παραπάνω. Θα δώσουμε πληροφορίες και στατιστικά για τα δεδομένα που χρησιμοποιούμε καθώς και τον τρόπο συλλογής τους. Θα δούμε ποσοστά επιτυχίας των μεθόδων και συγκρίσεις μεταξύ τους.

Τέλος στο Κεφάλαιο 5 κάνουμε μια ανακεφαλαίωση, τονίζουμε τα σημαντικότερα συμπεράσματα που προέκυψαν και θέτουμε νέους στόχους για μελλοντικές επεκτάσεις της εργασίας.

Κεφάλαιο 2

Σχετικές Εργασίες

Στο Κεφάλαιο αυτό θα παρουσιάσουμε εργασίες που εκμεταλλεύονται τη δύναμη των κοινωνικών δικτύων βασισμένων σε τοποθεσία και ιδιαίτερα του κειμένου που δημιουργούν οι χρήστες για μια πληθώρα ερευνητικών θεμάτων. Αρχικά θα αναφερθούμε συνοπτικά σε γενικότερα ερευνητικά θέματα που έχουν προκύψει από τα ΚΔΒΤ και σε εργασίες που έχουν ασχοληθεί με αυτά. Στη συνέχεια θα δούμε εργασίες που χρησιμοποιούν το κείμενο των ΚΔΒΤ και πραγματεύονται τεχνικές για εξαγωγή πληροφορίας από αυτό για διάφορες χρήσεις. Θα δούμε σχετική δουλειά με εξαγωγή θεμάτων από κείμενο που σχετίζεται άμεσα με την παρούσα εργασία όπως και προσπάθειες χρήσης των πιθανολογικών θεματικών μοντέλων σε μικρά κείμενα. Τέλος, θα δούμε σχετικές εργασίες για τα προβλήματα της ανάθεσης ετικετών σε κείμενο και παροχή εξατομικευμένων προτάσεων βασισμένες σε κείμενο.

2.1 Γενικά Ερευνητικά Θέματα ΚΔΒΤ

Παραθέτουμε μερικά από τα θέματα που απασχολούν τους ερευνητές των ΚΔΒΤ:

- *Εκτίμηση Ομοιότητας μεταξύ χρηστών:* Το ιστορικό των τοποθεσιών ενός χρήστη μπορεί να μας δώσει χρήσιμες πληροφορίες για τον ίδιο. Χρήστες που επισκέπτονται τα ίδια μέρη είναι πολύ πιθανό να έχουν ομοιότητες στη συμπεριφορά τους και τα ενδιαφέροντά τους. Η πληροφορία της ομοιότητας μεταξύ των χρηστών έχει πολλές εφαρμογές με ίσως πιο χαρακτηριστική τις προτάσεις φιλίας. Οι εργασίες που έχουν ασχοληθεί με το θέμα της ομοιότητας χρηστών με βάση την τοποθεσία αλλά και με τις προτάσεις φιλίας είναι πολλές. Κάποια παραδείγματα είναι οι [21], [22] και [23].

- *Ανίχνευση Κοινοτήτων:* Χρησιμοποιώντας την πληροφορία ομοιότητας των χρηστών μπορούμε να ομαδοποιήσουμε τους χρήστες σε κοινότητες με βάση τις ομοιότητες τους. Η ανίχνευση κοινοτήτων έχει μεγάλη πρακτική εφαρμογή. Για παράδειγμα ένας χρήστης που θέλει να οργανώσει μια ομαδική δραστηριότητα θα μπορέσει εύκολα να βρει και άλλους χρήστες με ενδιαφέρον στο ίδιο αντικείμενο και να το πραγματοποιήσει. Μερικά παραδείγματα εργασιών που ασχολούνται με ανίχνεση κοινοτήτων είναι οι [24] και [25].

- *Ανίχνευση ειδικών σε κάθε περιοχή:* Από το ιστορικό των τοποθεσιών των χρηστών μπορούμε να καταλάβουμε για κάθε περιοχή ποιοι χρήστες είναι πιο έμπειροι στη συγκεκριμένη περιοχή. Έτσι, αν για παράδειγμα έχουμε μια υπηρεσία προτάσεων, η γνώμη και οι αξιολογήσεις των συγκεκριμένων χρηστών θα έχουν μεγαλύτερη βαρύτητα από των υπολοίπων στη συγκεκριμένη περιοχή. Στην [26] μεταξύ των υπολοίπων γίνεται χρήση αυτής της πληροφορίας.

- *Προτάσεις τοποθεσιών σε χρήστες:* Η εύρεση περιοχών που έχουν μεγάλο ενδιαφέρον μέσα σε μια πόλη καθώς και οι διαδρομές που συνήθως ακολουθούν οι χρήστες που τις επισκέπτονται έχει μεγάλη πρακτική χρησιμότητα. Είναι βασική πληροφορία για ένα χρήστη που επισκέπτεται μια άγνωστη πόλη. Στην [26] προτείνεται ένα μοντέλο βασισμένο στον HITS αλγόριθμο για την εύρεση των περιοχών και διαδρομών.
  Άλλο ένα ενδιαφέρον ερευνητικό θέμα είναι η παροχή προγραμματισμού διαδρομής στους χρήστες ανάλογα με τις προτιμήσεις τους. Οι διαδρομές αυτές μπορεί εκτός από τοποθεσίες και διαδρομές να παρέχουν και χρονικό προγραμματισμό για την ώρα και τη διάρκεια της επίσκεψης της κάθε τοποθεσίας. Στην [27] οργανώνουν ιστορικά δεδομένα τοποθεσίας χρηστών σε ένα γράφο και προτείνουν μια αρχιτεκτονική για παροχή προτάσεων διαδρομών.
  Στην [28] προτείνεται ένα σύστημα που παρέχει στον χρήστη διπλές προτάσεις: α) Τις πιο δημοφιλείς δραστηριότητες που μπορεί να κάνει σε μία περιοχή και β) Τις πιο δημοφιλείς περιοχές για κάθε συγκεκριμένη δραστηριότητα. Για να το πετύχει αυτό χρησιμοποιεί και τα σχόλια των χρηστών για κάθε τοποθεσία, επιπρόσθετα στις διαδρομές τους.

- *Εξατομικευμένες προτάσεις τοποθεσιών:* Στην [23] παρουσιάζεται ένα σύστημα προτάσεων τοποθεσιών οι οποίες είναι εξατομικευμένες ανάλογα με τις προτιμήσεις του χρήστη. Προκειμένου να δουλέψει αυτό, οι ομοιότητες μεταξύ των χρηστών ενσωματώνονται σε ένα μοντέλο συνεργατικού φιλτραρίσματος [51]. Η γενική ιδέα του μοντέλου αυτού, το οποίο θα δούμε και στη συνέχεια αυτής της εργασίας, είναι πως όμοιοι χρήστες αξιολογούν όμοια αντικείμενα με τον παρόμοιο τρόπο. Παρόλο που βασίζεται στην ομοιότητα χρηστών και έχει πολύ καλά αποτελέσματα στη πρόβλεψη της συμπεριφοράς ενός χρήστη, έχει το μειονέκτημα ότι πρέπει να υπολογιστεί η ομοιότητα μεταξύ όλων των συνδυασμών χρηστών. Σε ένα πραγματικό σύστημα αυτό είναι αρκετά χρονοβόρο και δύσκολο. Στην [29] προτείνουν ένα μοντέλο συνεργατικού φιλτραρίσματος που βασίζεται στις τοποθεσίες. Στο μοντέλο αυτό υπολογίζουν τις συσχετίσεις μεταξύ των τοποθεσιών όπως προκύπτουν από τα δεδομένα των χρηστών. Δεδομένου ότι ο αριθμός των τοποθεσιών είναι πολύ μικρότερος από τον αριθμό των χρηστών, η μέθοδος αυτή είναι πιο εύκολη να εφαρμοστεί σε ένα πραγματικό σύστημα.

- *Ανίχνευση συμβάντων μέσω κοινωνικών δικτύων:* Κάποιες εργασίες έχουν ασχοληθεί με το θέμα της ανίχνευσης συμβάντων όπως εκδηλώσεις, συναυλίες, ατυχήματα, φυσικά φαινόμενα κ.α σε πραγματικό χρόνο χρησιμοποιώντας δεδομένα που ανεβάζουν οι χρήστες στα ΚΔΒΤ όταν αντιμετωπίζουν τέτοια συμβάντα από την περιοχή που συμβαίνουν. Στην [30] παρουσιάζεται μια μέθοδος εντοπισμού συμβάντων χρησιμοποιώντας το Twitter. Αυτό λειτουργεί παρατηρώντας τυχόν αύξηση στον αριθμό των χρηστών και των tweets από μια περιοχή που είναι μεγαλύτερη από το φυσιολογικό για τη συγκεκριμένη περιοχή.

2.2 Χρήση Μικρών Κειμένων των ΚΔΒΤ

Τα μικρά κείμενα που αφήνουν οι χρήστες των ΚΔΒΤ αξιολογώντας τις τοποθεσίες που επισκέπτονται αποτελούν αντικείμενο μελέτης σε πολλές εργασίες που διαπραγματεύονται μια πληθώρα ερευνητικών θεμάτων και εφαρμογών.

Στην [31] οι συγγραφείς χρησιμοποιούν μεταξύ άλλων τα σχόλια του Foursquare για να κατηγοριοποιήσουν τους χρήστες με βάση τη συμπεριφορά τους το περιεχόμενο και το είδος των σχολίων, το αν είναι σχετικά αυτά που γράφουν σε σχέση με την τοποθεσία κ.α. Παρόμοια δουλεία υπάρχει και στην [32], όπου οι συγγραφείς χρησιμοποιούν τα χαρακτηριστικά των σχολίων για να ανιχνεύσουν παράτυπους χρήστες που χρησιμοποιούν τα σχόλια για διαφήμιση, προσωπική προώθηση ή τα σχόλιά τους είναι ανεπιθύμητα ή υβριστικά.

Στην [33] παρουσιάζονται τεχνικές με τις οποίες επιλέγονται για κάθε τοποθεσία σχόλια για προβολή στους χρήστες βάσει χαρακτηριστικών όπως γλώσσα, ποιότητα του σχολίου, προσωπικές προτιμήσεις του χρήστη και χρονική καταλληλότητα. Στην [34] οι συγγραφείς παρουσιάζουν και συγκρίνουν τεχνικές για την ανίχνευση συναισθήματος (θετικά ή αρνητικά ως προς κάποιο θέμα) σε σχόλια του Foursquare χρησιμοποιώντας τόσο μεθόδους μάθησης με επίβλεψη (κατηγοριοποιητές) όσο και χωρίς επίβλεψη.

Στις [37, 38] οι συγγραφείς εκμεταλλεύονται την δύναμη των μικρών κειμένων (micro-reviews) με σκοπό την επιλογή συνόλων αξιολογήσεων από άλλα συστήματα οι οποίες καλύπτουν αποτελεσματικά τα σημαντικότερα θέματα-χαρακτηριστικά των τοποθεσιών που αξιολογούνται. Οι ίδιοι συγγραφείς στην [39] αντιμετωπίζουν το πρόβλημα της δημιουργίας περιλήψεων από τα μικρά κείμενα έτσι ώστε να είναι αντιπροσωπευτικές, περιεκτικές και αναγνώσιμες.

Τέλος υπάρχουν αρκετές εργασίες που πραγματεύονται την εύρεση θεμάτων από διάφορα κείμενα κοινωνικών δικτύων, όπως η [36] η οποία παρουσιάζει μια μέθοδο που συνδυάζει τεχνικές επεξεργασίας φυσικής γλώσσας, τεχνικές που βασίζονται σε ετικέτες και σημασιολογικές τεχνικές.

2.3 Προσθήκη ετικετών σε πόρους

Η ανάθεση ετικετών-λέξεων κλειδιών (tags) σε κείμενα, εικόνες η βίντεο (γενικότερα σε πόρους), βελτιώνει την παρεχόμενη εμπειρία στον χρήστη, επιτρέποντας την ευκολότερη οργάνωση και ταξινόμηση και ανάκτηση των πόρων αυτών. Ειδικά στα κοινωνικά δίκτυα όπου η αύξηση του περιεχομένου είναι ραγδαία, η χρήση ετικετών γίνεται όλο και περισσότερο αναγκαία. Η διαδικασία αυτοματοποίησης της διαδικασίας προσθήκης ετικετών και ιδιαίτερα εξατομικευμένων ετικετών είναι ένα πολύ δημοφιλές ερευνητικό θέμα.

Στην [5] οι συγγραφείς δημιουργούν δύο σύνολα από ετικέτες για ένα URL, μία με δημοφιλή και μία με προτεινόμενα. Τα δημοφιλή είναι αυτά που χρησιμοποιούν πιο συχνά οι χρήστες για να αναθέσουν στο URL, ενώ τα προτεινόμενα είναι η τομή α) του λεξιλογίου που χρησιμοποιεί ο συγκεκριμένος χρήστης που μας ενδιαφέρει με β) όλα τις ετικέτες που έχουν ανατεθεί στο συγκεκριμένο URL. Μοντελοποιούν το πρόβλημα της προσθήκης εξατομικευμένων ετικετών σαν ένα "query and ranking" [52] πρόβλημα. Οι συγγραφείς επίσης κατασκευάζουν και χρησιμοποιούν ένα γράφημα συσχέτισης των κειμένων καθώς και ένα διμερές γράφημα μεταξύ κειμένων και χρηστών.

Στην [6] οι συγγραφείς υποστηρίζουν ότι μια τεχνική από το πεδίο της εξόρυξης δεδομένων

όπως το *clustering* μπορεί να χρησιμοποιηθεί για την ανίχνευση τάσεων και για ευκολότερη και πιο αποδοτική διαδικασία απόδοσης των tags. Υποστηρίζουν ότι η χρήση συστάδων από ετικέτες μπορεί να γεφυρώσει το χάσμα μεταξύ χρηστών και πόρων και παρουσιάζουν έναν αλγόριθμο που στηρίζεται σε μια ιεραρχία συστάδων από ετικέτες.

Στην [7] οι συγγραφείς συγκρίνουν δύο αλγορίθμους για προτάσεις ετικετών σε ένα πραγματικό σύνολο δεδομένων. Ο ένας βασίζεται σε συνεργατικό φιλτράρισμα χρηστών ενώ ο άλλος χρησιμοποιεί γραφήματα και είναι βασισμένος στον αλγόριθμο Folkrank [20]. Αποδεικνύουν ότι και οι δύο αλγόριθμοι έχουν καλύτερα αποτελέσματα από άλλες μη εξατομικευμένες μεθόδους δίνοντας προβάδισμα στον αλγόριθμο που βασίζεται σε γραφήματα.

2.4 Προτάσεις βασισμένες σε κείμενο

Τα συστήματα παροχής προτάσεων έχουν αναπτυχθεί για να βοηθούν τον χρήστη να ανακαλύψει "αντικείμενα" σύμφωνα με τις προσωπικές του προτιμήσεις. Αυτά τα αντικείμενα μπορεί να είναι ιστοσελίδες, βιβλία ταινίες προϊόντα, εστιατόρια κ.α. Αυτά τα συστήματα συνήθως χρησιμοποιούν προηγούμενη πληροφορία από τις προτιμήσεις του χρήστη προκειμένου να λειτουργήσουν αποδοτικά. Η μεγαλύτερη πρόκληση για τον δημιουργό ενός τέτοιου συστήματος είναι η συλλογή αρκετών αρχικών δεδομένων και η αποδοτική χρήση τους. Τα συστήματα προτάσεων που βασίζονται σε κείμενο έρχονται να αντιμετωπίσουν το πρόβλημα αυτό (γνωστό και ως πρόβλημα κρύας εκκίνησης ή cold-start) που είναι συχνό στα συστήματα που βασίζονται σε τεχνικές συνεργατικού φιλτραρίσματος και έχουν τη δυνατότητα να λειτουργούν και με νέα αντικείμενα ή με αντικείμενα χωρίς προηγούμενη πληροφορία.

Στην [8] οι συγγραφείς προτείνουν μια μέθοδο παροχής προτάσεων χρησιμοποιώντας μη δομημένα κείμενα. Η μέθοδός τους έχει δύο βήματα. Στο πρώτο βήμα, τα κείμενα που εκφράζουν υποκειμενική γνώμη μαρκάρονται με βάση τη γνώμη αυτή. Τα αποτελέσματα αυτά χρησιμοποιούνται στο δεύτερο βήμα για την παροχή προτάσεων με τη χρήση πλέον τεχνικής συνεργατικού φιλτραρίσματος. Αξιολογούν τη μέθοδό τους χρησιμοποιώντας αξιολογήσεις ταινιών. Τελικά αποδεικνύουν ότι είναι εφικτή η χρήση κειμένου από διαφορετικές πηγές για την τροφοδότηση ενός συστήματος προτάσεων, εφόσον χρησιμοποιηθεί η γνώμη που εκφράζει το κάθε κείμενο. Έτσι αντιμετωπίζεται το πρόβλημα κρύας εκκίνησης όταν δεν υπάρχουν δεδομένα από την κύρια πηγή.

Στην [9] οι συγγραφείς παρουσιάζουν ένα web-based σύστημα προτάσεων για ταινίες, το οποίο χρησιμοποιεί κατηγοριοποίηση κειμένου για εξαγωγή πληροφορίας από παρουσιάσεις ταινιών. Αποδεικνύουν ότι οι τεχνικές κατηγοριοποίησης κειμένου λειτουργούν αρκετά καλά

κάτω από προϋποθέσεις και προσπαθούν να ανακαλύψουν σε ποιες περιπτώσεις είναι κατάλληλο να χρησιμοποιηθούν.

2.5 Θεματικά μοντέλα και σύντομα κείμενα

Όσο η χρήση των διαδικτυακών κοινωνικών δικτύων αυξάνεται και αποτελεί σημαντικό εργαλείο επικοινωνίας μεταξύ των ανθρώπων, είναι επιτακτική η δυνατότητα αποδοτικής ανάλυσης αυτού του περιεχομένου. Τα θεματικά μοντέλα [1, 2, 3] είναι αποδεδειγμένα πολύ δυνατά εργαλεία στην ανάλυση κειμένων και στον εντοπισμό λανθανόντων μοτίβων μέσα σε αυτά. Χρησιμοποιούνται με ποικίλους τρόπους σε συνηθισμένα κείμενα και συνήθως με μεγάλη επιτυχία. Στη συνέχεια της εργασίας θα τα δούμε αναλυτικά και θα τα χρησιμοποιήσουμε.

Τα κείμενα που συναντούμε στα κοινωνικά δίκτυα είναι γενικά αρκετά διαφορετικά από τα παραδοσιακά κείμενα, κυρίως λόγω του μικρού μεγέθους τους αλλά και λόγω της ποικιλίας των διαφορετικών τρόπων που μπορεί να τα χειριστεί κάποιος. Υπάρχουν μερικές εργασίες που εστιάζουν στο περιεχόμενο του Twitter, το οποίο αποτελείται από μικρά κείμενα το πολύ 140 χαρακτήρων.

Στην [11] οι συγγραφείς εστιάζουν στο πώς θα εκπαιδεύσουν θεματικά μοντέλα αποδοτικά για σύντομα κείμενα από το Twitter. Έχουν δώσει ιδιαίτερη προσοχή στο πως θα ομαδοποιήσουν τα δεδομένα και πως κάθε διαφορετική στρατηγική ομαδοποίησης επιδρά στην ποιότητα των παραγόμενων μοντέλων. Επίσης μελετούν πως μπορούν να τροποποιηθούν τα κοινά εργαλεία θεματικών μοντέλων για να λειτουργούν με τέτοιους είδους κείμενα πιο αποδοτικά. Αποδεικνύουν ότι το μέγεθος του κειμένου παίζει μεγάλο ρόλο στην ποιότητα των θεμάτων που παράγονται, καθώς και ότι η ομαδοποίηση των κειμένων έχει ως αποτέλεσμα καλύτερα θέματα.

Στην [10] οι συγγραφείς χρησιμοποιούν θεματικά μοντέλα προκειμένου να συγκρίνουν το περιεχόμενο του Twitter με ένα παραδοσιακό μέσο ειδήσεων, την εφημερίδα New York Times. Αναπτύσσουν και αποδεικνύουν την αποτελεσματικότητα ενός θεματικού μοντέλου που το ονομάζουν Twitter-LDA το οποίο είναι βελτιστοποιημένο για σύντομα κείμενα.

Στην [12] γίνεται χρήση μιας ημι-επιβλεπόμενης εκδοχής των θεματικών μοντέλων προκειμένου να γίνει χαρακτηρισμός των χρηστών και των tweets με βάση διάφορα χαρακτηριστικά τους.

Τα θεματικά μοντέλα έχουν χρησιμοποιηθεί και για πολλούς ακόμα σκοπούς. Στην [13] οι συγγραφείς τα χρησιμοποιούν πάνω σε δεδομένα επιστημονικών συνεδρίων προκειμένου να αναλύσουν το πως μεταβάλλεται η θεματολογία στο πέρασμα των χρόνων. Στην [14] χρησιμο-

ποιούν τα θεματικά μοντέλα για να μοντελοποιήσουν και να ανακαλύψουν καθημερινές ασχολίες των ανθρώπων μετατρέποντας σε κείμενα δεδομένα από αισθητήρες τοποθετημένους στο ανθρώπινο σώμα.

Κεφάλαιο 3

Πρόβλεψη του Θέματος Μικρών Κειμένων

3.1 Επισκόπηση του προβλήματος

Όπως έχουμε ήδη αναφέρει και νωρίτερα, στα κοινωνικά δίκτυα βασισμένα σε τοποθεσία οι χρήστες έχουν τη δυνατότητα να αφήνουν σχόλια (tips) με τη μορφή σύντομου κειμένου στις τοποθεσίες (venues) που επισκέπτονται, αναφέροντας συνήθως τις εμπειρίες τους και συχνά αξιολογώντας την συγκεκριμένη τοποθεσία. Τα σχόλια που αφήνουν οι χρήστες συνήθως περιέχουν γνώμες για ένα η περισσότερα θέματα (topics) που αφορούν την τοποθεσία που αντιστοιχούν. Για παράδειγμα, ένα σχόλιο σε κάποιο εστιατόριο μπορεί να περιέχει μία πρόταση για ένα συγκεκριμένο πιάτο, ή ένα παράπονο για την εξυπηρέτηση ή τις τιμές.

Αν είχαμε τη δυνατότητα να βρίσκουμε με αυτοματοποιημένο τρόπο τη θεματολογία αυτών των σχολίων θα είχαμε μία πολύ χρήσιμη πληροφορία η οποία θα μπορούσε να χρησιμοποιηθεί σε διάφορες εφαρμογές. Ακόμα πιο ενδιαφέρον θα ήταν να προβλέπουμε τη θεματολογία των σχολίων που πρόκειται να αφήσει ένας χρήστης σε μια τοποθεσία, πράγμα που προαπαιτεί να έχουμε κάποια προηγούμενη γνώση και για τον χρήστη και για την τοποθεσία.

Μερικές από τις εφαρμογές που θα μπορούσε να έχει η παραπάνω γνώση είναι η παροχή εξατομικευμένων προτάσεων τοποθεσιών στους χρήστες, το φιλτράρισμα και η ομαδοποίηση των σχολίων προκειμένου να γίνεται η προβολή τους στον κάθε χρήστη με βάση τις προτιμήσεις του, αλλά και η παροχή πληροφορίας προς τις αντίστοιχες επιχειρήσεις με σκοπό τη βελτίωση των υπηρεσιών τους ή ακόμα και την χρήση τους για λόγους διαφήμισης.

Ωστόσο τα σχόλια αυτά έχουν ιδιαίτερα χαρακτηριστικά που τα κάνουν την ανάλυση τους δύσκολη. Συνήθως είναι μικρά και το περιεχόμενό τους συχνά δεν έχει σωστή δομή κειμένου, περιέχει αργκό και διάφορες καθημερινές εκφράσεις. Επιπλέον, είναι πολύ συχνή η παρουσία ανεπιθύμητων, διαφημιστικών ή παραπλανητικών σχολίων.

Στην εργασία αυτή έχουμε ως στόχο να προβλέψουμε το περιεχόμενο ενός σχολίου που θα αφήσει ένας δεδομένος χρήστης σε μια δεδομένη τοποθεσία. Η πρόβλεψη θα γίνει με τη μορφή του πιο κατάλληλου θέματος για το σχόλιο, του οποίου η μορφή διαφέρει ανάλογα με την προσέγγιση του προβλήματος. Στην περίπτωση της προσέγγισης με θεματικά μοντέλα, όπως θα δούμε και παρακάτω, το θέμα θα είναι μια κατανομή από λέξεις ενώ στην περίπτωση της προσέγγισης με κατηγοριοποιητές, θα είναι συγκεκριμένο θέμα και άμεσα κατανοητό. Στη διάθεσή μας έχουμε σύνολα δεδομένων από πραγματικά σύνολα σχολίων από χρήστες και τοποθεσίες που συλλέξαμε από το κοινωνικό δίκτυο του Foursquare, τα οποία θα δούμε αναλυτικά στο επόμενο κεφάλαιο.

3.2 Πιθανολογικά Θεματικά Μοντέλα

Η ποσότητα της πληροφορίας που είναι διαθέσιμη διαδικτυακά αυξάνεται με ραγδαίους ρυθμούς εδώ και αρκετά χρόνια και ένας από τους βασικούς λόγους είναι τα κοινωνικά δίκτυα. Αυτό έχει σαν αποτέλεσμα να είναι πολύ δύσκολη η ανάγνωση, η μελέτη και η ανάλυση πληροφορίας με τη μορφή κειμένων από τους ανθρώπους έτσι ώστε να λαμβάνουν αρκετή πληροφορία γρήγορα και εύκολα. Για αυτό το λόγο οι ερευνητές στο πεδίο της μηχανικής μάθησης ανέπτυξαν τα πιθανολογικά θεματικά μοντέλα, τα οποία είναι σύνολο αλγορίθμων με στόχο την εξερεύνηση μεγάλων συλλογών από κείμενα και την προσθήκη θεματικής πληροφορίας σε αυτά.

Οι *αλγόριθμοι θεματικών μοντέλων* [3] είναι στατιστικές μέθοδοι που αναλύουν τις λέξεις πρωτότυπων κειμένων με στόχο την αναγνώριση θεμάτων που περιέχουν, τον τρόπο που συσχετίζονται αυτά τα θέματα μεταξύ τους, καθώς και την εξέλιξη των θεμάτων αυτών στο χρόνο. Οι αλγόριθμοι αυτοί είναι "μη επιβλεπόμενοι", δηλαδή δεν χρειάζονται κάποια προηγούμενη πληροφορία ή ετικέτες για τα κείμενα, αλλά τα θέματα προκύπτουν από την ανάλυση των αρ-

χικών κειμένων. Ανάμεσα στα πλεονεκτήματα των αλγορίθμων θεματικών μοντέλων είναι ότι μπορούν να εφαρμοστούν σε τεράστιες συλλογές κειμένων, πράγμα που θα ήταν αδύνατο να γίνει με ανθρώπινο χέρι. Με τις νεότερες εξελίξεις μάλιστα μπορούν να εφαρμοστούν ακόμα και σε ζωντανές ροές κειμένων από το διαδίκτυο. Επίσης μπορούν να προσαρμοστούν σε πολλών ειδών δεδομένα, όπως σε γενετικά δεδομένα, σε εικόνες και σε κοινωνικά δίκτυα.

3.2.1 Λανθάνουσα Κατανομή του Dirichlet (LDA)

Σε αυτή την ενότητα θα δούμε μερικές πληροφορίες σε υψηλό επίπεδο για την *Λανθάνουσα Κατανομή του Dirichlet (Latent Dirichlet Allocation ή LDA)* [1], η οποία περιγράφει ίσως το πιο δημοφιλές αλλά και σχετικά απλό θεματικό μοντέλο σε σχέση με τα υπόλοιπα. Η βασική ιδέα της LDA είναι ότι τα κείμενα αναπαριστώνται ως τυχαίες μίξεις λέξεων, επιλεγμένων από θέματα. Κάθε θέμα χαρακτηρίζεται από μια κατανομή λέξεων. Στην εργασία αυτή θα τη χρησιμοποιήσουμε για την εξαγωγή των θεμάτων από τα δεδομένα μας.

Για να κατανοήσουμε την LDA πιο εύκολα, αρκεί να περιγράψουμε τη "φανταστική" απλοποιημένη διαδικασία με την οποία το μοντέλο υποθέτει ότι έχουν δημιουργηθεί τα κείμενα. Το μοντέλο θεωρεί ότι τα θέματα έχουν δημιουργηθεί πριν από τη δημιουργία των κειμένων και είναι ίδια για κάθε συλλογή. Αυτό που διαφέρει είναι η "αναλογία" του κάθε θέματος μέσα σε κάθε κείμενο. Το μοντέλο περιγράφει τη διαδικασία με την οποία κάθε κείμενο από το σύνολο δεδομένων μας έχει αποκτήσει τις λέξεις του.

Το σύνολο δεδομένων μας είναι μία συλλογή από κείμενα (documents), και κάθε κείμενο είναι μια συλλογή από λέξεις. Ορίζουμε:

- Το w αναπαριστά μία λέξη. Οι λέξεις αναπαρίστανται χρησιμοποιώντας διανύσματα μοναδιαίας βάσης τα οποία έχουν μία μοναδική συνιστώσα ίση με ένα και όλες τις άλλες συνιστώσες ίσες με μηδέν. Το w^u αναπαριστά την u-οστή λέξη του λεξικού μας και $w^u = 1$ ενώ $w^v = 0$ αν $u \neq v$
- Το **w** αναπαριστά ένα κείμενο, δηλαδή ένα διάνυσμα λέξεων, με $\mathbf{w} = (w_1, w_2, ..., w_N)$.
- Η a είναι παράμετρος της κατανομής Dirichlet [4].
- Το **z** αναπαριστά ένα διάνυσμα από θέματα, όπου αν το i-οστό στοιχείο του **z** είναι 1 τότε η επιλογή των λέξεων γίνεται από το i-οστό θέμα.
- Το β είναι ένας $k \times V$ πίνακας πιθανοτήτων λέξεων για κάθε θέμα (γραμμή) και κάθε λέξη (στήλη), όπου $\beta_{i,j} = p(w^j = 1 \mid z^i = 1)$

Με αυτά τα δεδομένα, η διαδικασία με την οποία θεωρεί το μοντέλο ότι προκύπτουν τα κείμενα είναι η εξής:

Για κάθε κείμενο:

(α) Επιλέγουμε μια κατανομή (μεγέθους *K*) για τα θέματα, έστω $\theta_d \sim Dir(a)$, όπου $Dir(a)$ μία ομοιόμορφη κατανομή Dirichlet με παράμετρο α.

(β) Για κάθε λέξη του κειμένου:

(i) Επιλέγουμε ένα από τα *K* θέματα από την κατανομή που επιλέξαμε στο βήμα (α), έστω $z_{d,n} \sim multi(\theta_d)$, όπου multi() είναι πολυωνυμική.

(ii) Επιλέγουμε μία από τις V λέξεις του παραπάνω θέματος, έστω $w_{d,n} \sim \beta_{z_{d,n}}$.

Από το μοντέλο παραγωγής κειμένου που περιγράφει η LDA, είναι ξεκάθαρο πως τα κείμενα αποτελούνται από πολλαπλά θέματα. Για παράδειγμα, ένα κείμενο σχετικό με την υγεία μπορεί να αποτελείται από λέξεις που προέρχονται από ένα θέμα σχετικό με τις εποχές , όπως η λέξη 'χειμώνας', και από λέξεις που προέρχονται από ένα θέμα σχετικό με ασθένειες, όπως η λέξη 'γρίπη'.

Το βήμα (α) αντικατοπτρίζει το γεγονός ότι κάθε κείμενο περιέχει θέματα σε διαφορετική αναλογία. Για παράδειγμα, ένα κείμενο από τη συλλογή του παραπάνω παραδείγματος μπορεί να περιέχει πολλές λέξεις από το θέμα με τις εποχές και καμία από το θέμα με τις ασθένειες, ενώ ένα άλλο κείμενο να περιέχει τον ίδιο αριθμό λέξεων από το κάθε θέμα.

Το βήμα (ii) αντικατοπτρίζει το γεγονός ότι κάθε λέξη του κειμένου είναι διαλεγμένη από ένα από τα *K* θέματα, σε αναλογία πάντα με την κατανομή των θεμάτων του κειμένου, όπως έχει επιλεχθεί από το βήμα (i). Η επιλογή της κάθε λέξης εξαρτάται από την κατανομή των λέξεων ολόκληρου του λεξικού, όπως έχει καθοριστεί από το θέμα που επιλέχθηκε.

Ένα σημείο που πρέπει να τονίσουμε είναι ότι το LDA μοντέλο δεν κάνει καμία θεώρηση τόσο για τη σειρά των λέξεων των κειμένων όσο και για τη σύνταξή τους. Η αναπαράσταση των κειμένων που ακολουθεί είναι η πολύ δημοφιλής bag-of-words [50]. Σύμφωνα με αυτήν, το κάθε κείμενο αναπαριστάται ως ένα διάνυσμα του οποίου κάθε θέση αναφέρεται σε μία λέξη του κειμένου και η τιμή της είναι το πλήθος εμφανίσεων της λέξης αυτής.

Ο βασικός στόχος των θεματικών μοντέλων είναι η αυτοματοποιημένη εξαγωγή των θεμάτων από μια συλλογή κειμένων, η ανακάλυψη δηλαδή της κρυφής δομής των κειμένων. Στην περιγραφή της διαδικασίας παραπάνω θεωρήσαμε ότι γνωρίζουμε από πριν τις κατανομές των θεμάτων και όλες τις παραμέτρους του μοντέλου, αυτό όμως προϋποθέτει ότι κάπως τα έχουμε προσδιορίσει. Η διαδικασία αυτή είναι ουσιαστικά η αντίστροφη της παραπάνω και ονομάζεται *statistical inference*. Ο στόχος της είναι να προσδιορίσει την *posterior* κατανομή των κρυφών μεταβλητών δοθέντων των κειμένων η οποία μπορεί να γραφεί με το συμβολισμό που έχουμε ορίσει ως:

$$p(\theta, \mathbf{z} \mid \mathbf{w}, a, \beta) = \frac{p(\theta, \mathbf{z}, \mathbf{w} \mid a, \beta)}{p(\mathbf{w} \mid a, \beta)} \tag{3.1}$$

Έχοντας την κατανομή των κρυφών αυτών μεταβλητών, δηλαδή το εκπαιδευμένο μοντέλο, μπορούμε να υπολογίσουμε για κάθε νέο κείμενο την κατανομή των θεμάτων που το απαρτίζουν. Θα αναφερόμαστε σε αυτήν την διαδικασία ως *topic inference*. Για το inference υπάρχουν διάφορες τεχνικές που εφαρμόζονται στο LDA, όμως η περιγραφή τους είναι αρκετά σύνθετη και ξεφεύγει από τους στόχους της παρούσας εργασίας. Αναλυτικές πληροφορίες υπάρχουν στο [1].

Στο Σχήμα 3.1 μπορούμε να δούμε οπτικά τις δύο διαδικασίες που αναφέρουμε παραπάνω. Αριστερά βλέπουμε την κατασκευαστική διαδικασία για τρία κείμενα. Το DOC1 διαλέγει τις λέξεις του με πιθανότητα 1.0 από το Topic 1, το DOC2 διαλέγει τις λέξεις του από το Topic 1 με πιθανότητα 0.5 και από το Topic 2 με πιθανότητα 0.5 ενώ το DOC3 διαλέγει τις λέξεις του από το Topic 2 με πιθανότητα 1.0. Δεξία βλέπουμε το πρόβλημα του inference, στο οποίο έχοντας τα κείμενα προσπαθούμε να προσδιορίσουμε τα θέματα και τις μεταβλητές του μοντέλου.

Συνοψίζοντας, παραθέτουμε την είσοδο και την έξοδο της κάθε διαδικασίας προκειμένου να γίνει πιο σαφές το πως χρησιμοποιούμε την LDA στην εργασία αυτή.

Για την εύρεση των θεμάτων (κατασκευή των μοντέλων):

- *Είσοδος:* Συλλογή κειμένων
- *Έξοδος:* Εκπαιδευμένο μοντέλο (θέματα, παράμετροι LDA)

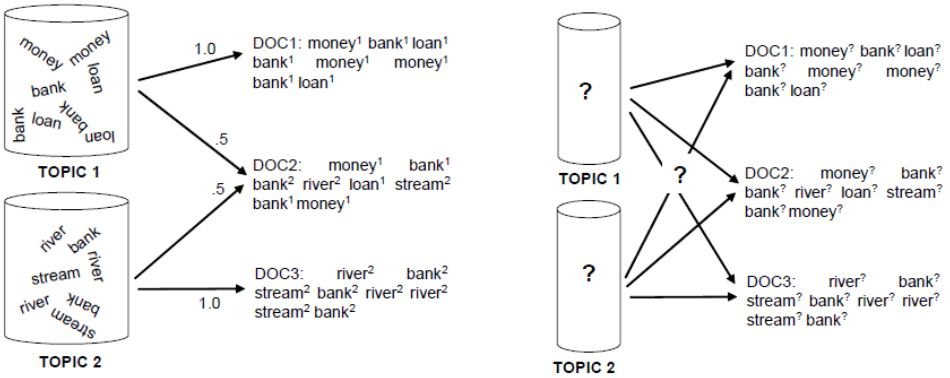
Για την εύρεση της κατανομής θεμάτων κάθε νέου κειμένου (topic inference):

- *Είσοδος:* Κείμενο και εκπαιδευμένο μοντέλο
- *Έξοδος:* Κατανομή θεμάτων

Ένα από τα πιο δύσκολα προβλήματα στη χρήση των θεματικών μοντέλων είναι η επιλογή του αριθμού των θεμάτων. Ο αριθμός αυτός είναι είσοδος για τον αλγόριθμο και η επιλογή του επηρεάζει την ερμηνεία των αποτελεσμάτων. Μια λύση με πολύ μικρό αριθμό γενικά έχει ως αποτέλεσμα πολύ ευρεία θέματα ενώ μια λύση με πολύ μεγάλο αριθμό έχει ως αποτέλεσμα θέματα που δεν έχουν ερμηνεία.

3.3 Πρόβλεψη Θέματος με χρήση Θεματικών Μοντέλων

Στην ενότητα αυτή θα παρουσιάσουμε μερικές προσεγγίσεις στο πρόβλημα της πρόβλεψης του θέματος ενός σχολίου δοθέντος ενός χρήστη και μιας τοποθεσίας, με χρήση θεματικών



Σχήμα 3.1: Διαδικασία παραγωγής κειμένου (αριστερά) και Inference (δεξιά). Πηγή [2]

.

μοντέλων κατασκευασμένων με LDA. Στο επόμενο κεφάλαιο θα παρουσιάσουμε πειραματικά αποτελέσματα με ποσοστά επιτυχίας για τις προσεγγίσεις αυτές και θα τις συγκρίνουμε μεταξύ τους.

3.3.1 Μοντέλα

Ένας από τους βασικότερους προβληματισμούς όταν ασχολείται κάποιος με θεματικά μοντέλα είναι ο τρόπος που θα χρησιμοποιήσει τα διαθέσιμα δεδομένα για την εκπαίδευση των μοντέλων. Η στρατηγική που θα ακολουθηθεί στην προετοιμασία των δεδομένων και στη συγχώνευση τους παίζει ρόλο, όπως έχουμε αναφέρει ξανά, στην ποιότητα των παραγόμενων μοντέλων.

Στην περίπτωσή μας έχουμε Χρήστες, Τοποθεσίες και σύνολα από Σχόλια για κάθε ένα από αυτά. Τα σχόλια είναι μικρά σε μέγεθος οπότε η χρήση ενός μόνο σχολίου ως κείμενο για την εκπαίδευση των μοντέλων δεν είναι η ενδεδειγμένη επιλογή. Η πληροφορία που περιέχει το καθένα είναι μικρή, δεδομένου ότι η λειτουργία των θεματικών μοντέλων στηρίζεται στην συν-εμφάνιση λέξεων. Επιλέξαμε λοιπόν να ομαδοποιήσουμε τα διαθέσιμα σχόλια ανά χρήστη και ανά τοποθεσία. Αυτές είναι εξάλλου και οι δύο οντότητες οι οποίες εμπλέκονται στην παραγωγή των μικρών κειμένων και αυτές που θα εξετάσουμε στη συνέχεια βάσει της κατανομής των θεμάτων τους. Τα κείμενα εκπαίδευσης των μοντέλων είναι λογικό να είναι ίδιου τύπου με τα κείμενα για τα οποία θα εξάγουμε κατανομές θεμάτων με την χρήση τους, για να έχει κάποιο νόημα η όλη διαδικασία.

Έτσι, δημιουργήσαμε δύο βασικές κατηγορίες μοντέλων:

Venue-data μοντέλο: Όλα τα σχόλια της τοποθεσίας συνενώθηκαν, ώστε να προκύψει ένα κείμενο για κάθε τοποθεσία. Τα κείμενα αυτά χρησιμοποιήθηκαν στην εκπαίδευση των

συγκεκριμένων μοντέλων. Τα θέματα που προκύπτουν από τα μοντέλα αυτά περιμένουμε να σχετίζονται με θέματα που αφορούν τις τοποθεσίες, όπως για παράδειγμα για το είδος τους (εστιατόρια, μπαρ, πάρκα κ.α.) ή για ιδιαιτερότητές τους (είδος κουζίνας, παροχές κ.α.)

User-data μοντέλο: Όλα τα σχόλια του κάθε χρήστη συνενώθηκαν, ώστε να προκύψει ένα κείμενο για κάθε χρήστη. Τα κείμενα αυτά χρησιμοποιήθηκαν στην εκπαίδευση των συγκεκριμένων μοντέλων. Τα θέματα που προκύπτουν από τα μοντέλα αυτά περιμένουμε να σχετίζονται με θέματα που αφορούν τους χρήστες, όπως για παράδειγμα με θέματα προτιμήσεων (κουζίνα, εξυπηρέτηση, τιμές) ή με γλωσσικές ιδιαιτερότητες κ.α.

Στο επόμενο κεφάλαιο θα δούμε αναλυτικές πληροφορίες για τα δεδομένα και τα μοντέλα που χρησιμοποιήσαμε στα πειράματα.

3.3.2 Προσεγγίσεις

Στις παρακάτω προσεγγίσεις προσπαθούμε να λύσουμε το πρόβλημα της πρόβλεψης του θέματος του σχολίου ενός δεδομένου χρήστη σε μια δεδομένη τοποθεσία με χρήση θεματικών μοντέλων. Θα αναφερόμαστε στο σχόλιο που δημιουργήθηκε από τον χρήστη *u* στην τοποθεσία *v* ως *tip(u,v)* και στο θέμα που θέλουμε να προβλέψουμε ως *topic(u,v)*.

Παρακάτω θεωρούμε ότι χρησιμοποιούμε κάποιο ή κάποια από τα ήδη εκπαιδευμένα LDA μοντέλα που έχουμε στη διάθεσή μας προκειμένου να εξάγουμε, με τη διαδικασία του topic inference, για κάθε document που μας ενδιαφέρει μία κατανομή από θέματα. Το κείμενο ενός χρήστη ή μιας τοποθεσίας είναι τα ομαδοποιημένα σχόλια για το κάθε ένα από αυτά, σε αντιστοιχία με τη διαδικασία κατασκευής των μοντέλων.

Για κάθε σχόλιο που μας ενδιαφέρει, εξάγουμε την κατανομή των θεμάτων του επίσης με τη διαδικασία του topic inference, χρησιμοποιώντας μόνο το ίδιο το σχόλιο ως κείμενο. Από την κατανομή αυτή επιλέγουμε το θέμα με τη μεγαλύτερη πιθανότητα και θεωρούμε ότι αυτό είναι το πραγματικό θέμα του συγκεκριμένου σχολίου. Αυτό το θέμα θα προσπαθήσουμε να προβλέψουμε στις παρακάτω προσεγγίσεις και με αυτό θα μετρήσουμε την επιτυχία στα πειράματά μας στη συνέχεια.

Προσέγγιση 1 - User top: Η πρώτη προσέγγιση είναι αρκετά απλή και βασίζεται στην υπόθεση ότι ένας συγκεκριμένος χρήστης είναι πολύ πιθανό στα σχόλια που αφήνει στις τοποθεσίες που επισκέπτεται, να αναφέρεται συχνότερα στα ίδια θέματα. Για παράδειγμα ένας χρήστης που επισκέπτεται εστιατόρια, να σχολιάζει πάντα την εξυπηρέτηση ή τις τιμές.

Η διαδικασία που ακολουθούμε είναι η εξής: Για κάθε χρήστη *u*, εξάγουμε με τη διαδικασία του topic inference την κατανομή των θεμάτων του, χρησιμοποιώντας ένα από τα διαθέσιμα μοντέλα. Ταξινομούμε την κατανομή αυτή και βρίσκουμε το πιο θέμα με την μεγαλύτερη πιθανότητα για τον χρήστη αυτόν, έστω *topic(u)*. Στη συνέχεια για να μετρήσουμε την επιτυχία της μεθόδου, ελέγχουμε για κάθε διαθέσιμο *tip(u,v)*, αν το πραγματικό *topic(u,v)* είναι το ίδιο με το *topic(u)* που βρήκαμε στο προηγούμενο βήμα.

Προσέγγιση 2 - Venue top: Η δεύτερη προσέγγιση βασίζεται στην υπόθεση ότι ένα σχόλιο σχετίζεται περισσότερο με την τοποθεσία στην οποία το έχει αφήσει ένας χρήστης και λιγότερο με τον ίδιο τον χρήστη. Αυτό σημαίνει ότι δοθείσας μιας τοποθεσίας, είναι πιο πιθανό ένας χρήστης να αναφερθεί στο σχόλιό του στα θέματα που αναφέρονται συνήθως οι χρήστες στην τοποθεσία αυτή.

Η διαδικασία που ακολουθούμε είναι η εξής: Για κάθε τοποθεσία *v*, εξάγουμε με τη διαδικασία του topic inference την κατανομή των θεμάτων της, χρησιμοποιώντας ένα από τα διαθέσιμα μοντέλα. Ταξινομούμε την κατανομή αυτή και βρίσκουμε το θέμα με την μεγαλύτερη πιθανότητα για την τοποθεσία αυτή, έστω *topic(v)*. Στη συνέχεια για να μετρήσουμε την επιτυχία της μεθόδου, ελέγχουμε για κάθε διαθέσιμο *tip(u,v)*, αν το πραγματικό *topic(u,v)* είναι το ίδιο με το *topic(v)* που βρήκαμε στο προηγούμενο βήμα.

Προσέγγιση 3 - Combined User-Venue: Στην επόμενη προσέγγιση σκεφτήκαμε ότι η αλήθεια μάλλον βρίσκεται κάπου ανάμεσα στις δύο προηγούμενες περιπτώσεις οπότε προσπαθήσαμε να τις συνδυάσουμε χρησιμοποιώντας δύο τρόπους:

Product: Για κάθε ζευγάρι *(u,v)*, υπολογίσαμε το γινόμενο των κατανομών των θεμάτων του χρήστη και της τοποθεσίας. Ταξινομούμε την κατανομή που προκύπτει και βρίσκουμε το θέμα με την μεγαλύτερη πιθανότητα για το ζευγάρι *(u,v)*. Στη συνέχεια για να μετρήσουμε την επιτυχία της μεθόδου, ελέγχουμε για κάθε διαθέσιμο *tip(u,v)*, αν το πραγματικό *topic(u,v)* είναι το ίδιο με το θέμα του ζευγαριού που βρήκαμε στο προηγούμενο βήμα.

Both: Αντί για το γινόμενο των δύο παραπάνω κατανομών, κρατάμε και τις δύο. Χρησιμοποιούμε τα θέματα με τις μεγαλύτερες πιθανότητες *topic(u)* και *topic(v)* και ελέγχουμε αν κάποιο από αυτά είναι ίδιο με το πραγματικό *topic(u,v)*.

Οι επόμενες προσεγγίσεις βασίζονται στην τεχνική του *Συνεργατικού Φιλτραρίσματος* (Collaborative Filtering) η οποία χρησιμοποιείται συνήθως σε συστήματα προτάσεων. Με τη γενική έννοια, είναι η διαδικασία φιλτραρίσματος πληροφορίας χρησιμοποιώντας τεχνικές που περιλαμβάνουν συνεργασία μεταξύ πολλαπλών πηγών δεδομένων, οντοτήτων, απόψεων κ.α.

Συνήθως η εφαρμογή της γίνεται σε πολύ μεγάλα σύνολα δεδομένων και μάλιστα μπορεί να εφαρμοστεί και σε διαφορετικούς τύπους δεδομένων.

Με τη στενότερη έννοια, το συνεργατικό φιλτράρισμα είναι μέθοδος αυτοματοποιημένης δημιουργίας προβλέψεων σχετικά με τα ενδιαφέροντα ενός χρήστη, συλλέγοντας πληροφορία και προτιμήσεις από πολλούς άλλους χρήστες (συνεργασία). Η υπόθεση που βασίζεται είναι ότι αν ένα άτομο Α έχει την ίδια γνώμη με ένα άτομο Β πάνω σε ένα συγκεκριμένο θέμα, τότε ο Α είναι πιο πιθανό να έχει την ίδια άποψη με τον Β και σε ένα διαφορετικό θέμα, από το να έχει την άποψη ενός άλλου τυχαίου ατόμου. Οι προβλέψεις που προκύπτουν είναι για συγκεκριμένο χρήστη κάθε φορά. Η τεχνική του συνεργατικού φιλτραρίσματος διαφέρει από την πιο απλή προσέγγιση του υπολογισμού μιας μέσης βαθμολογίας για κάθε θέμα που θέλουμε να προβλέψουμε, χρησιμοποιώντας μόνο αριθμό από ψήφους.

Στην πράξη η τεχνική μπορεί να χρησιμοποιηθεί με πολλές διαφορετικές μορφές και τύπους ανάλογα με την εφαρμογή. Αναλυτικότερες πληροφορίες υπάρχουν στα [51] και [19].

Για να υπολογίσουμε την ομοιότητα τόσο μεταξύ χρηστών όσο και μεταξύ τοποθεσιών χρησιμοποιούμε την ομοιότητα συνημιτόνου. Η *ομοιότητα συνημιτόνου* [49], είναι ένα μέτρο ομοιότητας μεταξύ δύο διανυσμάτων που προκύπτει από τον υπολογισμό του συνημιτόνου της γωνίας μεταξύ τους. Το συνημίτονο των 0° είναι 1 και για κάθε άλλη γωνία είναι μικρότερο του 1. Δύο διανύσματα με τον ίδιο προσανατολισμό έχουν ομοιότητα συνημιτόνου 1, δύο ορθογώνια διανύσματα έχουν ομοιότητα 0 ενώ δύο διανύσματα αντιδιαμετρικά έχουν ομοιότητα -1. Η ομοιότητα συνημιτόνου συνήθως χρησιμοποιείται με θετικές τιμές άρα και το αποτέλεσμα είναι στο [0,1]. Το αποτέλεσμα είναι στο διάστημα αυτό για όσες διαστάσεις και αν την χρησιμοποιήσουμε. Συνήθως η ομοιότητα συνημιτόνου χρησιμοποιείται σε πολλές διαστάσεις. Η ομοιότητα συνημιτόνου ορίζεται ως εξής:

$$similarity = \cos(A, B) = \frac{AB}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2}\sqrt{\sum_{i=1}^{n} (B_i)^2}} \tag{3.2}$$

Στον τομέα της Ανάκτησης Πληροφορίας και της Εξόρυξης Δεδομένων, όπου συχνά θέλουμε να κάνουμε σύγκριση κειμένων κάθε όρος είναι ουσιαστικά και μία διάσταση. Κάθε κείμενο χαρακτηρίζεται από ένα διάνυσμα, στο οποίο η τιμή της κάθε διάστασης αντιστοιχεί στο πλήθος των εμφανίσεων του συγκεκριμένου όρου μέσα στο κείμενο αυτό. Η ομοιότητα συνημιτόνου σε αυτή την περίπτωση δίνει ένα χρήσιμο μέτρο για το πόσο όμοια είναι δύο κείμενα μεταξύ τους με βάση το περιεχόμενό τους. Στην περίπτωση αυτή η τιμή της ομοιότητας θα είναι από 0 έως 1, αφού οι συχνότητες των όρων είναι θετικοί αριθμοί.

Η ομοιότητα συνημιτόνου είναι ιδιαίτερα δημοφιλής και ένας από τους λόγους είναι

πως ο υπολογισμός της είναι πολύ αποδοτικός, ειδικά αν έχουμε αραιά διανύσματα, αφού χρησιμοποιούνται μόνο οι μη-μηδενικές διαστάσεις τους.

Προσέγγιση 4 - Συνεργατική Χρηστών: Αυτή η προσέγγιση κάνει χρήση της τεχνικής του συνεργατικού φιλτραρίσματος υπολογίζοντας την ομοιότητα μεταξύ χρηστών. Η βασική ιδέα είναι ότι ένας χρήστης είναι πολύ πιθανό να αφήσει ένα σχόλιο σε μια τοποθεσία σχολιάζοντας τα ίδια θέματα με έναν άλλο χρήστη που έχει μεγάλη ομοιότητα με τον πρώτο και έχει αφήσει σχόλιο στην ίδια τοποθεσία.

Η διαδικασία που ακολουθούμε έχει ως εξής: Για κάθε σχόλιο που έχει αφήσει ένας χρήστης *u* σε μια τοποθεσία *v*, βρίσκουμε τον πιο όμοιο χρήστη *u'* στον χρήστη *u* που έχει αφήσει σχόλιο στην τοποθεσία *v*. Αν το πραγματικό θέμα του σχολίου που έχει αφήσει αυτός ο χρήστης είναι ίδιο με του σχολίου που έχει αφήσει ο χρήστης *u* που μας ενδιαφέρει, δηλαδή αν *topic(u',v)* είναι ίδιο με το *topic(u,v)*, τότε θεωρούμε ότι έχουμε επιτυχία.

Προσέγγιση 5 - Συνεργατική Τοποθεσιών: Άλλη μία προσέγγιση με χρήση της τεχνικής του συνεργατικού φιλτραρίσματος είναι να χρησιμοποιήσουμε την ομοιότητα μεταξύ των τοποθεσιών αντί για ομοιότητα μεταξύ των χρηστών. Με απλά λόγια η ιδέα της προσέγγισης αυτής είναι ότι αν ένας χρήστης αφήσει ένα σχόλιο σε μια τοποθεσία σχολιάζοντας κάποιο συγκεκριμένο θέμα, τότε είναι πολύ πιθανό να σχολιάσει το ίδιο θέμα και σε μια άλλη τοποθεσία που έχει μεγάλη ομοιότητα με την προηγούμενη.

Η διαδικασία που ακολουθούμε έχει ως εξής: Για κάθε σχόλιο που έχει αφήσει ένας χρήστης *u* σε μια τοποθεσία *v*, βρίσκουμε την πιο όμοια τοποθεσία *v'* στην τοποθεσία *v*, στην οποία ο χρήστης *u* έχει επίσης αφήσει σχόλιο. Αν το πραγματικό θέμα αυτού του σχολίου είναι ίδιο με το πραγματικό θέμα του σχολίου που έχει αφήσει ο χρήστης *u* στην τοποθεσία *v* που μας ενδιαφέρει, δηλαδή αν το *topic(u,v')* είναι ίδιο με το *topic(u,v)*, τότε θεωρούμε ότι έχουμε επιτυχία. Για να υπολογίσουμε την ομοιότητα μεταξύ των τοποθεσιών χρησιμοποιούμε και εδώ την ομοιότητα συνημιτόνου.

3.4 Πρόβλεψη Θέματος με χρήση Κατηγοριοποιητών

Στην ενότητα αυτή θα προσεγγίσουμε το πρόβλημα της πρόβλεψης του θέματος ενός σχολίου δοθέντος ενός χρήστη και μιας τοποθεσίας με χρήση κατηγοριοποιητών, αφού πρώτα κάνουμε μια εισαγωγή σε αυτούς και τη χρήση τους. Στο επόμενο κεφάλαιο θα παρουσιάσουμε

πειραματικά αποτελέσματα με ποσοστά επιτυχίας για τις προσεγγίσεις μας αυτές.

3.4.1 Κατηγοριοποίηση (Classification)

Η *κατηγοριοποίηση* [40] στον τομέα της μηχανικής μάθησης και της στατιστικής, είναι το πρόβλημα της εύρεσης σε ποια κατηγορία ανήκει ένα νέο στοιχείο, με βάση ένα σύνολο δεδομένων εκπαίδευσης το οποίο περιλαμβάνει στοιχεία για τα οποία η κατηγορία τους είναι γνωστή. Το σύνολο των κατηγοριών είναι προκαθορισμένο από πριν. Ο στόχος της διαδικασίας είναι η δημιουργία ενός καλού μοντέλου, το οποίο θα κατηγοριοποιεί σωστά καινούρια δεδομένα. Για παράδειγμα μπορεί τα δεδομένα να είναι καινούρια e-mail τα οποία πρέπει να ταξινομηθούν στις κατηγορίες 'spam' ή 'non-spam'.

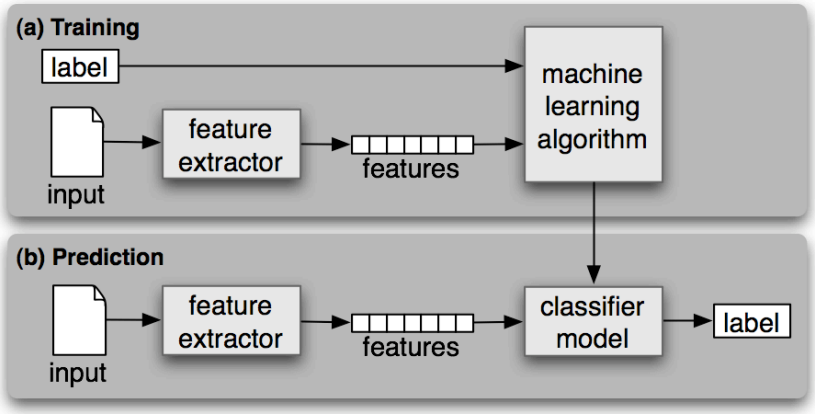
Στον τομέα της μηχανικής μάθησης η κατηγοριοποίηση ανήκει στην κατηγορία της μάθησης με επίβλεψη (supervised learning), επειδή προϋπόθεση για την λειτουργία του είναι η ύπαρξη ενός συνόλου δεδομένων εκπαίδευσης (training set), τα στοιχεία του οποίου είναι σωστά κατηγοριοποιημένα από την αρχή. Η αντίστοιχη διαδικασία χωρίς επίβλεψη ονομάζεται συσταδοποίηση (clustering).

Θα μπορούσαμε να περιγράψουμε τη διαδικασία της κατηγοριοποίησης με δύο βήματα, τα οποία φαίνονται και παραστατικά στο Σχήμα 3.2:

1. *Εκμάθηση:* Στο βήμα αυτό δημιουργείται το μοντέλο με χρήση ενός αλγορίθμου κατηγοριοποίησης και του συνόλου δεδομένων εκπαίδευσης (training set) που είναι ήδη κατηγοριοποιημένο. Το μοντέλο που δημιουργείται ονομάζεται κατηγοριοποιητής (classifier) και αναπαρίσταται με τη μορφή κανόνων κατηγοριοποίησης(classification rules), δέντρων απόφασης(decision trees) ή μαθηματικών τύπων.

2. *Κατηγοριοποίηση:* Αφού δημιουργηθεί ο κατηγοριοποιητής, το επόμενο βήμα είναι ο έλεγχος της αποτελεσματικότητάς του. Για το σκοπό αυτό χρησιμοποιούμε ένα δοκιμαστικό σύνολο δεδομένων (test set) επίσης κατηγοριοποιημένο, διαφορετικό όμως από το training set. Κατηγοριοποιούμε τα δοκιμαστικά δεδομένα και συγκρίνουμε τα αποτελέσματα που μας έδωσε ο κατηγοριοποιητής με τα σωστά που είχαμε ήδη από πριν. Έτσι υπολογίζουμε το ποσοστό των στοιχείων που κατηγοριοποιήθηκαν σωστά και αν είναι καλό σημαίνει ότι ο κατηγοριοποιητής που εκπαιδεύσαμε είναι καλός και μπορεί να χρησιμοποιηθεί και για άγνωστα δεδομένα.

Οι περισσότεροι αλγόριθμοι κατηγοριοποίησης περιγράφουν κάθε στοιχείο προς κατηγοριοποίηση ως ένα διάνυσμα από ξεχωριστά και μετρήσιμα χαρακτηριστικά (features) του στοι-



Σχήμα 3.2: Οπτικοποίηση διαδικασίας κατηγοριοποίησης. Πηγή: NLTK Book [16]

.

χείου. Τα χαρακτηριστικά αυτά μπορεί να είναι διαφόρων ειδών ανάλογα με τα στοιχεία που έχουμε, όπως για παράδειγμα δυαδικά (όπως 'true' ή 'false'), ακέραιοι (όπως εμφανίσεις λέξεων σε κείμενα), πραγματικές τιμές (όπως κάποια μέτρηση), κατηγορίες (όπως 'A','B','C') και άλλα. Αν τα στοιχεία είναι εικόνες, μπορεί να τα χαρακτηριστικά να είναι η ένταση των pixels. Επίσης κάποιοι αλγόριθμοι έχουν περιορισμούς στα χαρακτηριστικά, όπως να δουλεύουν μόνο με διακριτές τιμές.

Μερικές από τεχνικές κατηγοριοποίησης που υπάρχουν:

- Δέντρα απόφασης
- Αλγόριθμοι κοντινότερου γείτονα
- Νευρωνικά Δίκτυα
- Support Vector Machines
- Naive Bayes
- Σύνολα κανόνων

Για τους σκοπούς της εργασίας αυτής επιλέξαμε να χρησιμοποιήσουμε τον Naive Bayes κατηγοριοποιητή, ο οποίος είναι δημοφιλής για την κατηγοριοποίηση κειμένου.

3.4.2 Ο Naive Bayes κατηγοριοποιητής

Ο Naive Bayes κατηγοριοποιητής είναι ένας πιθανοτικός κατηγοριοποιητής που βασίζεται στο θεώρημα του Bayes. Ο στόχος του είναι επίσης η ανάθεση κατηγορίας-κλάσης σε ένα νέο στοιχείο από ένα σύνολο προκαθορισμένων κατηγοριών και αυτό γίνεται υπολογίζοντας την πιθανότητα που έχει το στοιχείο να ανήκει στην κάθε κατηγορία. Λέγεται Naive επειδή η λει-

τουργία του στηρίζεται σε μια παραδοχή ανεξαρτησίας. Πιο συγκεκριμένα, η παραδοχή είναι ότι η παρουσία ή απουσία ενός χαρακτηριστικού (feature) σε μία κατηγορία είναι ανεξάρτητη από την παρουσία ή απουσία ενός άλλου χαρακτηριστικού στην ίδια κατηγορία. Στην πραγματικότητα μπορεί να μη συμβαίνει πάντα αυτό, ωστόσο με την παραδοχή αυτή αποφεύγονται πολύπλοκοι υπολογισμοί.

Η όλη φιλοσοφία του Naive Bayes είναι να επιλέξει μια την πιο πιθανή κατηγορία για μια είσοδο (posterior probability), όπου η εκ των πρότερων πιθανότητα (prior probability) για κάθε ετικέτα είναι γνωστή. Από εκεί και πέρα, ανάλογα με τη συνεισφορά του κάθε χαρακτηριστικού ξεχωριστά, υπολογίζεται το πόσο πιθανό είναι να ανήκει σε μια κατηγορία.

Ουσιαστικά ο Naive Bayes υπολογίζει την παρακάτω πιθανότητα:

$$p(C_k \mid F_1, F_2, F_3...F_n) = \frac{p(C_k) \times p(F_1, F_2, F_3...F_n \mid C_k)}{p(F_1, F_2, F_3...F_n)} \tag{3.3}$$

ή με απλά λόγια:

$$prosterior = \frac{prior \times likelihood}{evidence} \tag{3.4}$$

Η C_k είναι η αντίστοιχη κατηγορία και $F_1, F_2, F_3...F_n$ τα διαφορετικά χαρακτηριστικά που έχουμε. Η (3.3) υπολογίζει την πιθανότητα ένα στοιχείο με χαρακτηριστικά $F_1, F_2, F_3...F_n$ να ανήκει στην κατηγορία C_k. Ο παρονομαστής εξαρτάται μόνο από τα χαρακτηριστικά του συνόλου εκπαίδευσης οπότε οι βασικός υπολογισμός που μας ενδιαφέρει είναι στον αριθμητή. Το δεξί μέρος του αριθμητή λόγω της Naive παραδοχής που αναφέραμε μπορεί να γραφεί απλοποιημένα ως:

$$p(F_1, F_2...F_n \mid C_k) = p(F_1 \mid C_k) \times p(F_2 \mid C_k)... \times p(F_n \mid C_k) \tag{3.5}$$

Η Naive Bayes κατηγοριοποίηση έχει αποδειχθεί ότι δουλεύει ικανοποιητικά, ακόμα και για σχετικά μικρά σύνολα εκπαίδευσης [41].

3.4.3 Επιλογή Χαρακτηριστικών

Η επιλογή των χαρακτηριστικών (features) και ο τρόπος κωδικοποίησης τους προκειμένου να χρησιμοποιηθούν σε μια μέθοδο εκπαίδευσης συνήθως παίζει σημαντικό ρόλο στην ικανότητα παραγωγής ενός καλού μοντέλου. Από τις πρώτες δυσκολίες στην κατασκευή ενός κατηγοριοποιητή είναι η επιλογή των χαρακτηριστικών που είναι χρήσιμα και ο τρόπος αναπαράστασής τους. Σε πολλές περιπτώσεις είναι δυνατό να έχουμε αξιοπρεπή απόδοση χρησιμοποιώντας απλά και προφανή σύνολα χαρακτηριστικών, ωστόσο μια πιο προσεκτική επιλογή ίσως μας δώσει ακόμα καλύτερα αποτελέσματα.

Τυπικά, η εξαγωγή των χαρακτηριστικών είναι μια διαδικασία δοκιμής-απόρριψης, αποφασίζοντας διαισθητικά ποιες πληροφορίες είναι σχετικές με το πρόβλημα και ποιες όχι. Συνήθως

ξεκινάμε με πολλά χαρακτηριστικά και αφαιρούμε. Παρ' όλα αυτά κάθε αλγόριθμος έχει και τους δικούς του περιορισμούς στον αριθμό των χαρακτηριστικών που μπορούν να χρησιμοποιηθούν. Χρησιμοποιώντας πολλά χαρακτηριστικά είναι πιθανό ο αλγόριθμος να εξαρτάται από λεκτικές ιδιομορφίες των δεδομένων εκπαίδευσης και να έχει κακή γενίκευση σε καινούρια δεδομένα. Ουσιαστικά είναι σαν να απομνημονεύει τα δεδομένα εκπαίδευσης. Το πρόβλημα αυτό είναι γνωστό σαν *overfitting* και είναι εντονότερο σε μικρά σύνολα δεδομένων εκπαίδευσης.

Μια συνήθης επιλογή που γίνεται στην περίπτωση της κατηγοριοποίησης κειμένου είναι τα χαρακτηριστικά των κατηγοριοποιητών να υποδεικνύουν παρουσία λέξεων στο κάθε στοιχείο. Εμείς επιλέγουμε να χρησιμοποιήσουμε τις πιο συχνές λέξεις από τα σύνολα δεδομένων μας. Στο επόμενο κεφάλαιο θα δούμε αναλυτικότερες πληροφορίες για αυτά και κυρίως για την επιλογή του πλήθους των λέξεων.

3.4.4 Προσεγγίσεις

Η κατηγοριοποίηση είναι μέθοδος μάθησης με επίβλεψη όπως έχουμε αναφέρει. Σε αντίθεση με τα θεματικά μοντέλα που η εξαγωγή των θεμάτων γίνεται από τον ίδιο τον αλγόριθμο μέσα από τα αρχικά κείμενα, εδώ θα χρησιμοποιήσουμε σύνολα δεδομένων εκπαίδευσης και ελέγχου από σχόλια στα οποία έχουμε αναθέσει εμείς κατηγορίες από πριν. Στο επόμενο κεφάλαιο θα παραθέσουμε πληροφορίες για τα σύνολα δεδομένων, τις κατηγορίες, την επιλογή των χαρακτηριστικών και των κατηγοριοποιητών.

Χρήση ενιαίου κατηγοριοποιητή για όλες τις κατηγορίες

Η πρώτη προσέγγιση προϋποθέτει την εκπαίδευση ενός Naive Bayes κατηγοριοποιητή ο οποίος θα κατηγοριοποιεί τα σχόλιο σε μία από τις προκαθορισμένες κατηγορίες που έχουμε επιλέξει. Τα χαρακτηριστικά των δειγμάτων θα είναι ένα σύνολο από λέξεις που περιλαμβάνονται στα σχόλια και πιο συγκεκριμένα η παρουσία ή η απουσία της κάθε λέξης στο κάθε στοιχείο.

Ο στόχος μας εδώ είναι να προβλέψουμε την κατηγορία που ανήκει ένα σχόλιο, δοθέντος του χρήστη που το άφησε και την τοποθεσία στην οποία το άφησε. Κατηγοριοποιούμε κάθε σχόλιο που μας ενδιαφέρει, έστω *tip(u,v)* με τον ενιαίο κατηγοριοποιητή και κρατάμε την κατηγορία του, έστω *cat(u,v)*.

Σε αντιστοιχία με τις προσεγγίσεις που κάναμε νωρίτερα στο LDA, έχουμε:

- *User-top:* Για κάθε χρήστη *u*, κατηγοριοποιούμε ένα ένα τα σχόλιά του με τον ενιαίο κατηγοριοποιητή μας και εξάγουμε ένα σύνολο από κατηγορίες. Θεωρούμε την κατηγορία που εμφανίζεται πιο συχνά ως κατηγορία του συγκεκριμένου χρήστη, έστω *cat(u)*. Στη

συνέχεια για να μετρήσουμε την επιτυχία της μεθόδου, ελέγχουμε για κάθε διαθέσιμο *tip(u,v)*, αν η πραγματική κατηγορία του σχολίου, *cat(u,v)*, είναι η ίδια με την *cat(u)* που βρήκαμε στο προηγούμενο βήμα.

- *Venue-top:* Για κάθε τοποθεσία *v*, κατηγοριοποιούμε ένα ένα τα σχόλιά του με τον ενιαίο κατηγοριοποιητή μας και εξάγουμε ένα σύνολο από κατηγορίες. Θεωρούμε την κατηγορία που εμφανίζεται πιο συχνά ως κατηγορία της συγκεκριμένης τοποθεσίας, έστω *cat(v)*. Στη συνέχεια για να μετρήσουμε την επιτυχία της μεθόδου, ελέγχουμε για κάθε διαθέσιμο *tip(u,v)*, αν η πραγματική κατηγορία του σχολίου, *cat(u,v)*, είναι η ίδια με την *cat(v)* που βρήκαμε στο προηγούμενο βήμα.

- *Both:* Για κάθε ζευγάρι *(u,v)* χρησιμοποιούμε τις πιο συχνές κατηγορίες *cat(u)* και *cat(v)* που βρήκαμε στις προηγούμενες προσεγγίσεις και ελέγχουμε για κάθε διαθέσιμο *tip(u,v)*, αν η πραγματική κατηγορία του σχολίου, *cat(u,v)*, είναι η ίδια με κάποια από αυτές.

Χρήση ενός κατηγοριοποιητή για κάθε κατηγορία

Η δεύτερη προσέγγιση που κάναμε προϋποθέτει ξεχωριστό κατηγοριοποιητή για κάθε προκαθορισμένη κατηγορία που έχουμε επιλέξει να αναθέσουμε στα σχόλια. Τα χαρακτηριστικά των δειγμάτων θα είναι όπως και στην προηγούμενη περίπτωση η παρουσία ή απουσία λέξεων από το κάθε σχόλιο. Εδώ όμως το σύνολο εκπαίδευσης και ελέγχου του κάθε κατηγοριοποιητή θα είναι ήδη κατηγοριοποιημένο σε δύο κατηγορίες, 'True' ή 'False', ανάλογα με το αν το στοιχείο περιλαμβάνεται στην συγκεκριμένη κατηγορία κάθε φορά. Για παράδειγμα ένα σχόλιο που ανήκει στην κατηγορία Α θα έχει ετικέτα True στο σύνολο δεδομένων του κατηγοριοποιητή της κατηγορίας Α και False στα υπόλοιπα σύνολα.

Το αποτέλεσμα κατηγοριοποίησης για κάθε σχόλιο θα είναι το σύνολο των απαντήσεων όλων των κατηγοριοποιητών. Αυτό σημαίνει ότι είναι πολύ πιθανό να υπάρχουν tips που ανήκουν σε παραπάνω από μία κατηγορίες, εμείς όμως θα κρατήσουμε μόνο μία από αυτές. Για να βρούμε την πιο ισχυρή 'πραγματική' κατηγορία για τα σχόλια στα οποία θα μετρήσουμε την επιτυχία των προσεγγίσεων μας αργότερα θα χρησιμοποιήσουμε τις πιθανότητες που μας δίνει ο κάθε κατηγοριοποιητής για την περίπτωση 'True'. Άρα για κάθε *tip(u,v)* που θέλουμε να προβλέψουμε, θεωρούμε *cat(u,v)* την κατηγορία με τη μεγαλύτερη πιθανότητα από τις παραπάνω.

Στη συνέχεια σε αντιστοιχία με τις προηγούμενες προσεγγίσεις έχουμε:

- *User-top:* Για κάθε χρήστη *u* με πλήθος σχολίων *K*, κατηγοριοποιούμε ένα ένα τα σχόλια του με όλους τους κατηγοριοποιητές μας και εξάγουμε ένα σύνολο από *K* διανύσματα

με τόσες απαντήσεις για το καθένα, όσες είναι και οι κατηγορίες. Θεωρούμε την πιο ισχυρή κατηγορία για τον χρήστη, αυτήν που εμφανίζει τα πιο πολλά 'True' στο σύνολο των κατηγοριοποιημένων tip, έστω *cat(u)*. Στη συνέχεια για να μετρήσουμε την επιτυχία της μεθόδου, ελέγχουμε για κάθε διαθέσιμο *tip(u,v)*, αν η πραγματική κατηγορία του σχολίου που αναφέραμε παραπάνω, *cat(u,v)*, είναι η ίδια με την *cat(u)* που βρήκαμε στο προηγούμενο βήμα.

- *Venue-top:* Για κάθε τοποθεσία *v* με πλήθος σχολίων *K*, κατηγοριοποιούμε ένα ένα τα σχόλια του με όλους τους κατηγοριοποιητές μας και εξάγουμε ένα σύνολο από *K* διανύσματα με τόσες απαντήσεις για το καθένα, όσες είναι και οι κατηγορίες. Θεωρούμε την πιο ισχυρή κατηγορία για την τοποθεσία, αυτήν που εμφανίζει τα πιο πολλά 'True' στο σύνολο των κατηγοριοποιημένων σχολίων, έστω *cat(v)*. Στη συνέχεια για να μετρήσουμε την επιτυχία της μεθόδου, ελέγχουμε για κάθε διαθέσιμο *tip(u,v)*, αν η πραγματική κατηγορία του σχολίου που αναφέραμε παραπάνω, *cat(u,v)*, είναι η ίδια με την *cat(v)* που βρήκαμε στο προηγούμενο βήμα.

- *Both:* Για κάθε ζευγάρι *(u,v)* χρησιμοποιούμε τις πιο συχνές κατηγορίες *cat(u)* και *cat(v)* που βρήκαμε στις προηγούμενες προσεγγίσεις και ελέγχουμε για κάθε διαθέσιμο *tip(u,v)*, αν η πραγματική κατηγορία του σχολίου, *cat(u,v)*, είναι η ίδια με κάποια από αυτές.

Κεφάλαιο 4

Πειραματικά Αποτελέσματα

Στο κεφάλαιο αυτό θα δούμε Πειραματικά Αποτελέσματα από τις προσεγγίσεις στα προβλήματα που αναφέραμε στο προηγούμενο. Αρχικά θα δώσουμε πληροφορίες και στατιστικά για τα δεδομένα που χρησιμοποιούμε καθώς και τον τρόπο συλλογής τους και στη συνέχεια θα δούμε ποσοστά επιτυχίας των μεθόδων και συγκρίσεις μεταξύ τους.

4.1 Συλλογή Δεδομένων

4.1.1 Το API του Foursquare

Το πρώτο βήμα για να πραγματοποιήσουμε την ανάλυση και τα πειράματά μας ήταν η συλλογή ικανού αριθμού κατάλληλων δεδομένων. Για το σκοπό χρησιμοποιήσαμε δεδομένα από το δίκτυο του Foursquare, το οποίο όπως έχουμε αναφέρει είναι ίσως ο πιο χαρακτηριστικός

εκπρόσωπος των κοινωνικών δικτύων βασισμένων σε τοποθεσία με τουλάχιστον 55 εκατομμύρια χρήστες και πάνω από 6 δισεκατομμύρια check-ins παγκοσμίως τη στιγμή που γράφουμε αυτήν την εργασία.

Το Foursquare, όπως και η πλειοψηφία των κοινωνικών δικτύων παρέχει μια προγραμματιστική διεπαφή (Application Programming Interface - API) [42]. Το API αυτό παρέχει τη δυνατότητα στους προγραμματιστές να συνδέσουν σχετικά εύκολα τις εφαρμογές τους με την πλατφόρμα του Foursquare και να χρησιμοποιήσουν τα δεδομένα του. Έτσι, τρίτες εφαρμογές έχουν την δυνατότητα να αποκτήσουν παρόμοια λειτουργικότητα με την επίσημη εφαρμογή του Foursquare για κινητές συσκευές αλλά και με το web περιβάλλον του και πολλές φορές να την επεκτείνουν.

Οργάνωση δεδομένων του API

Το API του Foursquare παρέχει πρόσβαση σε μεγάλη ποικιλία δεδομένων, οργανωμένων σε οντότητες και από λειτουργίες πάνω σε αυτές που τις ονομάζει endpoints. Κάθε οντότητα είναι και ένα διαφορετικό είδος πόρου (resource) όπως για παράδειγμα venues, users, tips και πολλά άλλα. Κάθε πόρος έχει ένα σύνολο από endpoints που μπορεί να ανήκουν σε μία από τις παρακάτω τρεις κατηγορίες:

- *General:* Γενικές λειτουργίες όπως αναζήτηση για τοποθεσίες ή χρήστες.
- *Aspects:* Λειτουργίες που αναφέρονται σε συγκεκριμένο αντικείμενο μιας οντότητας, όπως ανάκτηση των σχολίων ενός συγκεκριμένου χρήστη ή μιας τοποθεσίας.
- *Actions:* Λειτουργίες που επίσης αναφέρονται σε συγκεκριμένο αντικείμενο και μπορούμε να πούμε πως είναι πιο 'διαχειριστικές', όπως αλλαγή των χαρακτηριστικών μιας τοποθεσίας ή διαγραφή ενός φίλου από τους χρήστες.

Στον Πίνακα 4.1 φαίνονται τα endpoints για τρεις βασικές κατηγορίες πόρων του Foursquare API. Πλήρης λίστα υπάρχει διαθέσιμη στο [42].

Μορφή ερωτήσεων και απαντήσεων

Η πρόσβαση στα δεδομένα και τις λειτουργίες του Foursquare API γίνεται με την αποστολή κατάλληλα διαμορφωμένων ερωτήσεων και στη συνέχεια λήψη κατάλληλα δομημένων απαντήσεων. Οι ερωτήσεις γίνονται με τη μορφή URL διευθύνσεων στις οποίες έχουν προστεθεί τα κατάλληλα ορίσματα κάθε φορά. Ας δούμε μερικά παραδείγματα:

- Έστω ότι θέλουμε να βρούμε ποια venues είναι κοντά σε μία περιοχή, για παράδειγμα στην πόλη των Ιωαννίνων. Θα χρησιμοποιήσουμε το venues/search endpoint ορίζοντας

Resource	General	Aspects	Actions
users	requests,search	checkins,friends,lists,mayorships photos,tastes,tips,venuelikes venuelikes	approve,setpings,unfriend deny,unfriend,update
venues	add,categories,search managed,suggestcompletion explore,timeseries,trending	events,herenow,hours likes,links,listed menu,nextvenues,photos similar,stats,tips	claim,dislike,edit flag,like,proposeedit setrole,setsinglelocation
tips	add	likes,listed,saves	flag,like,unmark

Πίνακας 4.1: Οργάνωση οντοτήτων στο Foursquare API

την παράμετρο near, ως εξής:
`https://api.foursquare.com/v2/venues/search?near=Ioannina`

- Έστω ότι θέλουμε να βρούμε τα tips ενός venue, για το οποίο ξέρουμε ήδη το venue id του. Θα χρησιμοποιήσουμε το venues/tips endpoint ώς εξής:
`https://api.foursquare.com/v2/venues/VENUE-ID/tips`

Τα παραδείγματα αυτά είναι αρκετά απλοποιημένα. Έχουμε παραλείψει και άλλες απαραίτητες παραμέτρους όπως authentication tokens που θα αναφερθούμε αργότερα.

Οι απαντήσεις που επιστρέφουν από το API είναι καλά δομημένες σε ένα πρότυπο που ονομάζεται *JavaScript Object Notation (JSON)* [44]. Το πρότυπο αυτό χρησιμοποιεί κείμενο σε αναγνώσιμη μορφή από τον άνθρωπο και η οργάνωσή τους είναι σε αντικείμενα με τη μορφή κλειδιού-τιμής. Η JSON μπορεί αρχικά να προήλθε από την γλώσσα Javascript, ωστόσο είναι τελείως ανεξάρτητη από γλώσσα. Είναι συμβατή με γενικές δομές δεδομένων που περιλαμβάνουν όλες οι μοντέρνες γλώσσες και αυτό την κάνει πολύ ευέλικτη. Επίσης υπάρχουν βιβλιοθήκες για εύκολη ανάγνωση και γραφή σε JSON για σχεδόν όλες τις ευρέως χρησιμοποιούμενες γλώσσες προγραμματισμού.

Διαδικασία Σύνδεσης

Προκειμένου να χρησιμοποιήσουμε το Foursquare API, αρχικά εγγράψαμε μία καινούρια web εφαρμογή στο Foursquare. Αυτό είναι απαραίτητο γιατί από αυτή την εφαρμογή και τους χρήστες της θα φαίνεται ότι γίνονται οι ερωτήσεις προς το API.

Στη συνέχεια δημιουργήσαμε κώδικα για την ανάκτηση, την αποθήκευση και την επεξεργασία των εισερχόμενων δεδομένων. Για το σκοπό αυτό χρησιμοποιήσαμε την γλώσσα προγραμματισμού Python, επειδή προσφέρει μεγάλη ευελιξία με πληθώρα βιβλιοθηκών και δομών για αποθήκευση και επανάκτηση τέτοιου είδους δεδομένων.

Το Foursquare API όπως και όλα σχεδόν τα μοντέρνα API των κοινωνικών δικτύων χρησιμοποιούν πιστοποίηση των εφαρμογών που συνδέονται με το πρότυπο *OAuth2* το οποίο προϋποθέτει την ανταλλαγή μιας σειράς από tokens τόσο στην αρχή της χρήσης μιας εφαρμογής όσο και σε κάθε ερώτημα που αποστέλλεται. Για την πιστοποίηση, την εύκολη αποστολή σωστά δομημένων ερωτημάτων καθώς και την λεκτική ανάλυση των απαντήσεων χρησιμοποιήσαμε τις βιβλιοθήκες pyfoursquare[45] και foursquare[46] για Python.

Περιορισμοί του API

Το Foursquare έχει θέσει για λόγους ασφαλείας αλλά και αξιοπιστίας της πλατφόρμας του κάποιους περιορισμούς στην πρόσβαση και συλλογή δεδομένων από το API του. Ο βασικότερος περιορισμός ως προς τον όγκο των δεδομένων του είναι το όριο των 500 ερωτημάτων ανά ώρα και ανά χρήστη της κάθε εγγεγραμμένης εφαρμογής. Πολλά είδη ερωτημάτων έχουν όρια στις απαντήσεις, όπως για παράδειγμα ο μέγιστος αριθμός από τοποθεσίες που μπορούμε να ζητήσουμε από μια περιοχή είναι τα 50 ανά ερώτημα. Ωστόσο στο συγκεκριμένο ερώτημα το API επιτρέπει μέχρι και 5000 ερωτήματα ανά ώρα με την προϋπόθεση ότι δε θα γίνουν εκ μέρους συγκεκριμένου χρήστη της εφαρμογής (userless query) και συνεπώς οι απαντήσεις δε θα είναι προσωποποιημένες. Περιορισμοί υπάρχουν και ως προς το είδος των δεδομένων που έχει πρόσβαση μια τρίτη εφαρμογή. Ο σημαντικότερος είναι η αδυναμία ανάκτησης των check-ins των χρηστών παρά μόνο του ενεργού χρήστη. Ο περιορισμός αυτός μας στερεί από μια πολύτιμη πληροφορία στα πλαίσια της ανάλυσης και ανάπτυξης μιας υπηρεσίας και δίνει το πλεονέκτημα στο ίδιο το Foursquare. Από την άλλη εξασφαλίζει την ιδιωτικότητα των χρηστών του και αποτρέπει πιθανή κακόβουλη χρήση τους.

4.1.2 Διαδικασία Συλλογής Δεδομένων

Για την ανάλυση μας επιλέξαμε να συλλέξουμε δεδομένα από δύο μεγάλες μεγάλες πόλεις των Ηνωμένων Πολιτειών της Αμερικής όπου η χρήση του Foursquare είναι πολύ διαδεδομένη. Η μία είναι η Νέα Υόρκη και η άλλη το Σαν Φρανσίσκο.

Η διαδικασία της συλλογής και για τις δύο πόλεις είχε ως εξής:

- Αρχικά χρησιμοποιήσαμε το venues/search endpoint για να βρούμε τις διαθέσιμες τοποθεσίες στις κεντρικές περιοχές των δύο πόλεων και να συλλέξουμε τα ID τους. Η σάρωση έγινε δίνοντας συντεταγμένες σημείων με τη μορφή ενός πλέγματος που κάλυπτε την κάθε περιοχή.
- Στη συνέχεια χρησιμοποιώντας τα venue id συλλέξαμε το πλήρες προφίλ για κάθε τοποθεσία καθώς και το σύνολο των σχολίων τους, χρησιμοποιώντας τα endpoints

Πόλη	Venues	Venue Tips	Users	User Tips	Τριάδες UVT
Νέα Υόρκη	41.178	194.764	18.260	170.642	48.215
Σαν Φρανσίσκο	42.724	154.320	14.478	175.181	55.095

Πίνακας 4.2: Στατιστικά των δεδομένων που συλλέχθηκαν

venues και venues/tips.

- Για χρήστες, επιλέξαμε να συλλέξουμε πληροφορίες για τους δημάρχους (mayors) των τοποθεσιών που είχαμε ήδη στα χέρια μας, με το σκεπτικό ότι είναι πιο ενεργοί χρήστες και κατ' επέκταση αφήνουν και μεγάλο αριθμό από σχόλια. Αφού βρήκαμε τα user id των δημάρχων, συλλέξαμε τα προφίλ τους και το σύνολο των σχολίων τους χρησιμοποιώντας τα endpoints users και users/tips.

Η συλλογή των δεδομένων πραγματοποιήθηκε τον Μάιο του 2013 για την πόλη της Νέας Υόρκης και τον Δεκέμβριο του 2013 για το Σαν Φρανσίσκο. Στον Πίνακα 4.2 μπορούμε να δούμε μερικά στατιστικά για τα σύνολα δεδομένων που συλλέχθηκαν. Τα νούμερα στην τελευταία στήλη είναι ο αριθμός των σχολίων τα οποία βρέθηκαν τόσο στα σχόλια χρηστών όσο και στα σχόλια τοποθεσιών στο αντίστοιχο σύνολο δεδομένων, πρόκειται δηλαδή για τα σχόλια για τα οποία γνωρίζουμε και τους χρήστες που τα δημιούργησαν και τις τοποθεσίες στις οποίες τα άφησαν.

4.1.3 Προ-επεξεργασία Κειμένου

Στη συνέχεια της εργασίας αυτής θα ασχοληθούμε κυρίως με το κείμενο των σχολίων. Για το λόγο αυτό, πριν τη χρήση του κειμένου είτε για την κατασκευή θεματικών μοντέλων είτε για την εκπαίδευση κατηγοριοποιητών πρώτα εκτελέσαμε μια προ-επεξεργασία στο περιεχόμενο των σχολίων. Πιο αναλυτικά:

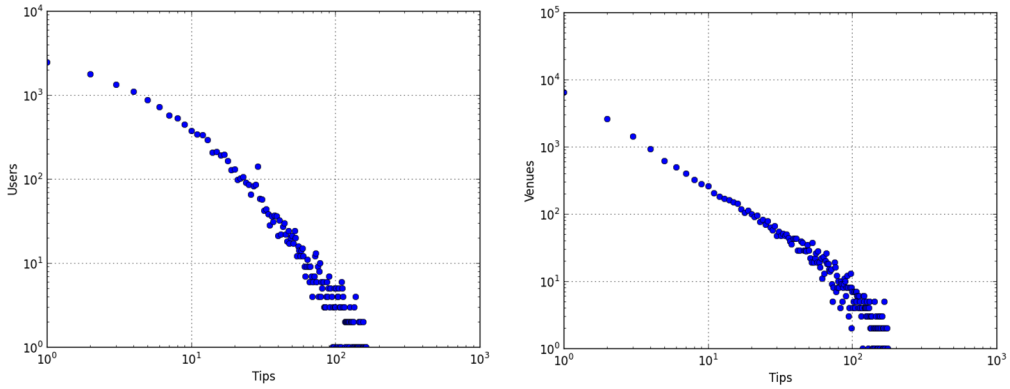
- Μετατρέψαμε όλους τους χαρακτήρες σε μικρούς λατινικούς.
- Αφαιρέσαμε σημεία στίξης και σύμβολα, url διευθύνσεις και αριθμούς.
- Αφαιρέσαμε τα πιο κοινά stopwords.
- Μετατρέψαμε τα ουσιαστικά στον ενικό αριθμό με χρήση της βιβλιοθήκης inflect της Python [47].
- Τέλος, αφαιρέσαμε τις λέξεις που εμφανίζονται μόνο μία φορά στο κάθε σύνολο δεδομένων καθώς και τις λέξεις που έχουν λιγότερους από τρεις χαρακτήρες.

4.2 Διάφορα Στατιστικά

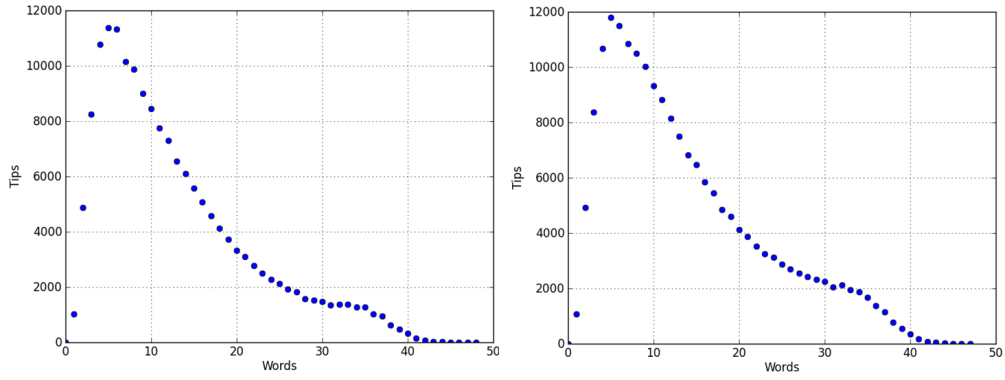
Σε αυτή την ενότητα θα δούμε μερικά στατιστικά που προέκυψαν από την ανάλυση των δεδομένων που συλλέξαμε.

4.2.1 Κατανομές σχολίων στις τοποθεσίες και τους χρήστες

Στο Σχήμα 4.1 μπορούμε να παρατηρήσουμε πως κατανέμονται τα σχόλια στους χρήστες και στις τοποθεσίες σε λογαριθμική κλίμακα, για το σύνολο δεδομένων της Νέας Υόρκης. Όπως φαίνεται, η πλειοψηφία των χρηστών και των τοποθεσιών έχουν λιγότερα από 10 σχόλια. Στο Σχήμα 4.2 φαίνεται αντίστοιχα η κατανομή των λέξεων στα σχόλια, με την πλειοψηφία των σχολίων να αποτελείται από λιγότερες από 10 λέξεις.



Σχήμα 4.1: Κατανομή των σχολίων στους χρήστες (αριστερά) και στις τοποθεσίες (δεξιά) σε λογαριθμική κλίμακα. (Νέα Υόρκη)



Σχήμα 4.2: Κατανομή λέξεων στα σχόλια των χρηστών (αριστερά) και στων τοποθεσιών (δεξιά). (Νέα Υόρκη)

4.2.2 Στατιστικά ανά κατηγορία τοποθεσίας

Το Foursquare επιτρέπει στους δημιουργούς των τοποθεσιών να τους αναθέτουν μία ή περισσότερες κατηγορίες κατά τη δημιουργία τους. Η ιεραρχία των κατηγοριών του Foursquare αριθμεί εκατοντάδες κατηγορίες, οργανωμένες σε δενδρική δομή.

Εμείς ταξινομήσαμε τις κατηγορίες που μαζέψαμε σε επτά γενικές κατηγορίες, λαμβάνοντας υπόψη την πρώτη κατηγορία της κάθε τοποθεσίας, εφόσον βέβαια υπήρχε και αυτή. Στους Πίνακες 4.3 και 4.4 και αντίστοιχα στα Σχήματα 4.3 και 4.4 υπάρχουν διαθέσιμα στατιστικά για τις τοποθεσίες της Νέας Υόρκης και του Σαν Φρανσίσκο ανά κατηγορία.

Κατηγορίες	Venues		Checkins		Tips		Words	
Building	10.649	**25,9%**	2.889.958	11,1%	10.175	5,2%	132.853	5,0%
Education	849	2,0%	638.596	2,5%	1.428	0,7%	18.463	0,7%
Nightlife	2.897	7,0%	3.843.434	14,8%	31.472	16,2%	421.266	15,7%
Food	7.293	17,7%	8.672.397	**33,5%**	108.844	**56,2%**	1.493.946	**55,8%**
Travel	1.827	4,4%	2.787.081	10,8%	2.944	1,5%	42.898	1,6%
Parks&Outdoors	2.364	5,7%	1.753.723	6,8%	4.592	2,3%	63.328	2,4%
Shops	8.525	20,7%	3.744.627	14,4%	25.971	13,4%	373.894	14,0%
Arts&Entertainment	2.504	6,0%	1.393.044	5,4%	7.758	4,0%	112.369	4,2%
No Category	4.270	10,3%	202.624	0,8%	1.580	0,8%	16.210	0,6%
Totals	**41.178**		**25.925.484**		**194.764**		**2.675.227**	

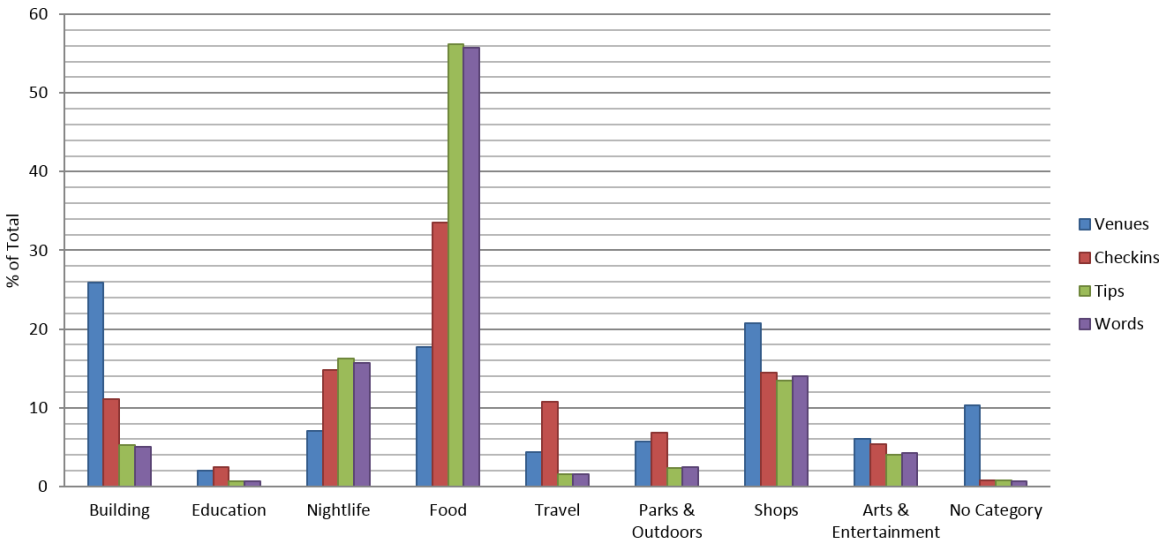
Πίνακας 4.3: Στατιστικά ανά κατηγορία τοποθεσίας (Νέα Υόρκη)

Κατηγορίες	Venues		Checkins		Tips		Words	
Building	10.390	**24,3%**	2.698.161	12,9%	8.717	5,6%	110.064	5,5%
Education	980	2,3%	352.195	1,7%	905	0,6%	12.099	0.6%
Nightlife	2.568	6,0%	2.312.799	11,0%	16.789	10,9%	211.253	10,6%
Food	6.185	14,5%	7.850.117	**37.4%**	88.725	**57,5%**	1.132.013	**56.6%**
Travel	3.100	7,3%	1.351.055	6,4%	7.148	4,6%	111.233	5,6%
Parks&Outdoors	2.815	6,6%	1.732.396	8,3%	5.937	3,8%	79.798	4,0%
Shops	9.127	21,4%	3.308.499	15,8%	19.141	12,4%	252.586	12,6%
Arts&Entertainment	2.564	6,0%	1.268.311	6,1%	6.221	4,0%	82.318	4,1%
No Category	4.995	11,4%	89.812	0,4%	737	0,5%	10.079	0,5%
Totals	**42.724**		**20.963.345**		**154.320**		**2.001.443**	

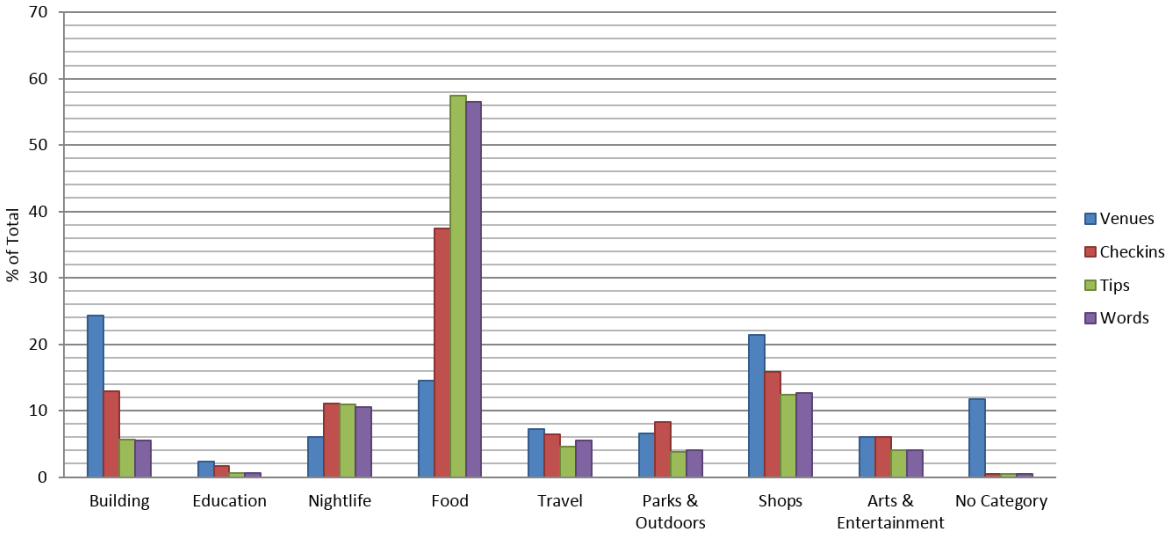
Πίνακας 4.4: Στατιστικά ανά κατηγορία τοποθεσίας (Σαν Φρανσίσκο)

4.2.3 Εύρεση ετικετών για Χρήστες και Τοποθεσίες

Όπως έχουμε αναφέρει ξανά, η διαδικασία προσθήκης ετικετών σε πόρους είναι πολύ διαδεδομένη τα τελευταία χρόνια, και ιδιαίτερα στα κοινωνικά δίκτυα και σε άλλους διαδικτυακούς πόρους. Μερικά από τα πλεονεκτήματα των ετικετών είναι η βελτιωμένη εμπειρία χρήσης και η αποδοτικότερη κατασκευή ευρετηρίων για τους πόρους.



Σχήμα 4.3: Στατιστικά ανά κατηγορία τοποθεσίας (Νέα Υόρκη)



Σχήμα 4.4: Στατιστικά ανά κατηγορία τοποθεσίας (Σαν Φρανσίσκο)

Εδώ χρησιμοποιήσαμε την TF-IDF μετρική προκειμένου να προσθέσουμε σκορ σε κάθε λέξη από κάθε διαθέσιμο χρήστη και τοποθεσία του συνόλου δεδομένων μας. Η *Term Frequency - Inverse Document Frequency (TF-IDF)* είναι μια μετρική που μας αποκαλύπτει πόσο σημαντική είναι μία λέξη σε ένα κείμενο, όταν έχουμε μια συλλογή κειμένων. Χρησιμοποιείται συχνά στον τομέα της Ανάκτησης Πληροφορίας καθώς της Εξόρυξης Δεδομένων. Η τιμή της είναι ανάλογη με το πλήθος των εμφανίσεων μιας λέξης μέσα σε ένα κείμενο (tf), αλλά είναι αντιστρόφως ανάλογη με το πλήθος των εμφανίσεων της λέξης μέσα σε όλη τη συλλογή (df). Κάποιες παραλλαγές της tf-idf χρησιμοποιούνται και από τις μηχανές αναζήτησης για την εύρεση της συνάφειας ενός κειμένου με κάποιο ερώτημα, αλλά και για διάφορους άλλους λόγους όπως για παράδειγμα φιλτράρισμα λέξεων από κείμενα. Η πιο απλή είναι μέθοδος υπολογισμού της tf-idf είναι:

$$\textit{tf-idf} = tf \times idf, \text{ όπου } \textit{idf} = \log \frac{N}{df} \text{ και } N\text{=συνολικός αριθμός κειμένων}$$

Στην περίπτωσή μας για κάθε λέξη χρησιμοποιήσαμε:

TF : Το πλήθος εμφανίσεων της λέξης στα σχόλια χρηστών ή τοποθεσιών.
DF : Το πλήθος των τοποθεσιών-χρηστών στα σχόλια των οποίων περιλαμβάνεται η λέξη.

Αφού υπολογίσαμε μία βαθμολογία για κάθε λέξη του κάθε χρήστη ή τοποθεσίας, ταξινομήσαμε τα αποτελέσματα και θεωρήσαμε ετικέτες τις Κ λέξεις με τις μεγαλύτερες βαθμολογίες για κάθε χρήστη και τοποθεσία ξεχωριστά. Στον Πίνακα 4.5 είναι διαθέσιμο ένα παράδειγμα από ετικέτες για τις τοποθεσίες της Νέας Υόρκης.

Venue Name	Tags
Jacques Torres Chocolate	chocolate torre wicked jacque cookie icecream chip mrchocolate download iphone
Almondine Bakery	croissant almondine almond macaron quiche muffin chocolate pretzel pari french
Manhattan Bridge	bridge view pedestrian train biker jump roadway brooklyn manhattan busta
Beekman Beer Garden	beekman beach view sand garden beer ches river pong seaport
Nitehawk Cinema	movie nitehawk theater popcorn queso film theatre drafthouse alamo flick
St. Anselm	butcher steak mashed potato sardine gratin grilled saddle hanger mustache
Radegast Hall & Biergarten	beer pretzel german liter radegast sausage bratwurst brat oktoberfest grill
Da Gennaro	italy ravioli umberto oil dipping cramped spaghetti gallo trap attentive
Los Feliz	margarita downstair taco tequila hibiscu feliz climb dancing cilantro jalapeno
Apple Store	store apple geniu ipad iphone soho workshop ipod computer appointment

Πίνακας 4.5: Δείγμα ετικετών για τοποθεσίες της Νέας Υόρκης με χρήση tf-idf (K=10)

4.3 Εκτίμηση πρόβλεψης με Θεματικά Μοντέλα

4.3.1 Επιλογή Δεδομένων και Κατασκευή Μοντέλων

Πριν την δημιουργία των κατάλληλων θεματικών μοντέλων για τα πειράματά μας, ήταν απαραίτητο να αποφασίσουμε ποια δεδομένα θα χρησιμοποιήσουμε καθώς και να κάνουμε μια προ-επεξεργασία αυτών.

Επιλέξαμε δύο υποσύνολα από τα αρχικά σύνολα δεδομένων κρατώντας μόνο τοποθεσίες και χρήστες που είχαν ικανό αριθμό από σχόλια. Θέσαμε αυτό το όριο στα 50 σχόλια για το σύνολο δεδομένων της Νέας Υόρκης και στα 40 σχόλια για το σύνολο δεδομένων του Σαν Φρανσίσκο. Στον Πίνακα 4.6 παραθέτουμε στατιστικά για τα δύο υποσύνολα δεδομένων που έμειναν, ενώ στον Πίνακα 4.7 μπορούμε να δούμε πως κατανέμονται σε κατηγορίες οι τοποθεσίες στα σύνολα δεδομένων που έμειναν.

Πόλη	Venues	Venue Tips	Users	User Tips	Τριάδες UVT
Νέα Υόρκη	953	87.899	561	52.607	6.023
Σαν Φρανσίσκο	878	66.213	947	77.143	8.624

Πίνακας 4.6: Υποσύνολο δεδομένων για δημιουργία μοντέλων

Κατηγορία	Food	Nightlife	Shops	Arts&Entert.	Building	Parks&Outdoor	Travel	Education
Τοποθεσίες NY	628	186	48	43	22	14	8	4
Ποσοστό	65,8%	19,6%	5,0%	4,5%	2,3%	1,5%	0,9%	0,4%
Τοποθεσίες SF	599	119	44	35	18	25	35	3
Ποσοστό	68,2%	13,6%	5,0%	4,0%	2,1%	2,8%	4,0%	0,3%

Πίνακας 4.7: Κατανομή τοποθεσιών σε κατηγορίες για τα επιλεγμένα σύνολα δεδομένων

Όπως αναφέραμε στο προηγούμενο κεφάλαιο, επιλέξαμε να ομαδοποιήσουμε τα σχόλια σε ένα κείμενο για κάθε χρήστη και για κάθε τοποθεσία. Με τα δεδομένα αυτά έχουμε δύο βασικές κατηγορίες μοντέλων: Venue-data μοντέλα και User-data μοντέλα.

Για κάθε κατηγορία από αυτές κατασκευάσαμε δύο εκδόσεις μοντέλων:

α) Χρησιμοποιώντας ολόκληρο το υποσύνολο των σχολίων χρηστών και τοποθεσιών αντίστοιχα.

β) Αφαιρώντας τα κοινά σχόλια μεταξύ χρηστών και τοποθεσιών από τα δεδομένα πριν την συγχώνευσή τους. Τα κοινά αυτά σχόλια είναι αυτά που συμμετέχουν στις τριάδες χρήστη-τοποθεσίας-σχολίου που χρησιμοποιούμε στη συνέχεια για να μετρήσουμε τα ποσοστά επιτυχίας των προσεγγίσεων μας.

Συνολικά λοιπόν κατασκευάσαμε 8 μοντέλα, 4 για τα δεδομένα της Νέας Υόρκης και 4 για τα δεδομένα του Σαν Φρανσίσκο. Στατιστικά για το πλήθος των σχολίων που χρησιμοποιήθηκαν στην κατασκευή των μοντέλων παρουσιάζουμε στον Πίνακα 4.8

Σύνολο δεδομένων	Συνολικά	Μετά την αφαίρεση	Αφαιρέθηκαν	
NY Χρήστες	52.607	46.584	6.023	11,5%
NY Τοποθεσίες	87.899	81.876	6.023	6,8%
SF Χρήστες	77.143	68.519	8.624	11,2%
SF Τοποθεσίες	66.213	57.589	8.624	13,0%

Πίνακας 4.8: Πλήθος σχολίων που χρησιμοποιήθηκαν στην εκπαίδευση

Για την κατασκευή των μοντέλων καθώς και για το inference των θεμάτων στη συνέχεια χρησιμοποιήσαμε το *Gensim* [48]. Το Gensim είναι μια δωρεάν βιβλιοθήκη για Python η οποία προσφέρει πληθώρα εργαλείων για επεξεργασία φυσικής γλώσσας και αλγορίθμους πιθανολογικών μοντέλων. Ανάμεσα σε αυτούς είναι φυσικά και το LDA. Οι υλοποιήσεις που προσφέρει το Gensim είναι αρκετά αποδοτικές και εύκολες στη χρήση.

Οι παράμετροι για την κατασκευή των μοντέλων που θα χρησιμοποιήσουμε στη συνέχεια είναι πλήθος θεμάτων ίσο με 100 και οι επαναλήψεις του LDA αλγορίθμου (περάσματα των κειμένων) ίσες επίσης με 100. Για όλα τα μοντέλα δημιουργήσαμε ένα κοινό λεξικό που περιλαμβάνει ολόκληρο το λεξιλόγιο των δεδομένων μας.

Παραδείγματα Θεματικών Μοντέλων

Για να καταλάβουμε πως μοιάζει ένα θεματικό μοντέλο σε υψηλό επίπεδο (λέξεων), στους Πίνακες 4.9 και 4.10 παραθέτουμε μερικά παραδείγματα από θέματα που έχουν δημιουργηθεί από τον LDA αλγόριθμο χρησιμοποιώντας τα σχόλια τοποθεσιών και χρηστών αντίστοιχα.

Επειδή τα θέματα περιλαμβάνουν μεγάλο αριθμό από λέξεις, στα παραδείγματα τυπώνουμε μόνο μερικές από αυτές που έχουν τα μεγαλύτερα βάρη σε κάθε θέμα. Με μια πρώτη ματιά μπορούμε να παρατηρήσουμε ότι τα θέματα που δημιουργήθηκαν με τα σχόλια τοποθεσιών είναι περισσότερο προσανατολισμένα σε χαρακτηριστικά των τοποθεσιών και βγάζουν μεγαλύτερο νόημα από τα θέματα που δημιουργήθηκαν με τα σχόλια των χρηστών.

4.3.2 Αποτελέσματα πρόβλεψης με θεματικά μοντέλα

Στην ενότητα αυτή παρουσιάζουμε ποσοστά επιτυχίας από τις προσεγγίσεις που παρουσιάσαμε στο προηγούμενο κεφάλαιο. Για κάθε προσέγγιση παραθέτουμε ποσοστά επιτυχίας τόσο για το σύνολο δεδομένων της Νέας Υόρκης όσο και για του Σαν Φρανσίσκο.

Topic	Words
0	chicken ramen fried best pork noodles spicy good salad food curry soup rice beef tea
1	beer great bar place good best happy awesome hour drinks night drink beers music selection
2	german boot schnitzel liter bratwurst suppenkuche suckling kiss pig bier das geisha's buddha venison
3	pizza best deep dish good crust great gym zoo slice time game star class encountered
4	games dog machines belly donuts dogs sliders arcade fortune fun bottomless play mimosas laughing quarters
5	offer san bring francisco front month desk phone expires sansome pie months commitment
6	burrito tacos best good super food taco salsa mexican great carnitas fish tea amazing order
7	great good food best burger amazing cheese delicious chicken fries service place order breakfast salad
8	cocktails cocktail drinks drink bar great place best order punch bartenders good bartender bourbon style
9	apple giants sausage store drinks best hour tea game day iphone plow happy boba buy
10	wine great north pizza beach place italian best taxi northbeachsf nbsf excludes nice anytime
11	pizza best amazing great food menu wine delicious good pedi wait bar order restaurant
12	chicken ceviche sandwich best sausage rotisserie roast sandwiches amazing sangria peruvian jerk good pisco
13	food good great best chicken service amazing place delicious order menu spicy thai awesome restaurant
14	coffee great best good sandwich place wifi iced delicious free latte breakfast love nice awesome
15	sangria sea donut lions crab pier apple planetarium donuts place wharf best cioppino maple great
16	cream ice chocolate caramel best good salted delicious movie cookies hot amazing banana flavors
17	garlic crab beach ocean wasabi poke tofu calamari strawberry mochi noodles salmon roasted mahi sand
18	great place view free hotel room nice san check best francisco city day wifi views
19	sushi roll best dim sum dumplings breakfast secret good sashimi sake tuna fish rolls

Πίνακας 4.9: Παραδείγματα θεμάτων από τα σχόλια των τοποθεσιών του Σαν Φρανσίσκο

Στους πίνακες υπάρχει διαχωρισμός ανάλογα με τα δεδομένα που χρησιμοποιήθηκαν για να δημιουργηθεί το θεματικό μοντέλο που χρησιμοποιείται (χρηστών ή τοποθεσιών). Επίσης υπάρχει διαχωρισμός ανάλογα με το αν τα σύνολα δεδομένων χρησιμοποιήθηκαν ολόκληρα ή αν αφαιρέθηκαν πρώτα τα κοινά σχόλια μεταξύ χρηστών και τοποθεσιών.

Για τις περισσότερες προσεγγίσεις επίσης δίνουμε ποσοστά επιτυχίας αν χρησιμοποιήσουμε τα πρώτα δύο θέματα με τις μεγαλύτερες πιθανότητες από τις κατανομές των θεμάτων των χρηστών και τοποθεσιών αντίστοιχα.

Η τυχαία προσέγγιση

Αρχικά θα ήταν χρήσιμο να έχουμε κάποιο μέτρο σύγκρισης, προκειμένου να αξιολογήσουμε καλύτερα τα ποσοστά επιτυχίας που θα δούμε στη συνέχεια. Αν εκτελούσαμε το πείραμα μας με τυχαίο τρόπο, τότε η διαδικασία θα ήταν η εξής:

- Για κάθε διαθέσιμη τριάδα user-venue-tip:
  - Επιλέγουμε έναν αριθμό μεταξύ 1-N (πλήθος των θεμάτων)
  - Επιλέγουμε ακόμα έναν (top-1 μέθοδοι) ή δύο (top-2 μέθοδοι) αριθμούς μεταξύ 1-N
  - Ελέγχουμε αν οποιοσδήποτε από τους αριθμούς στο βήμα 2 είναι ίδιος με τον αριθμό στο βήμα 1 και αν είναι το καταγράφουμε σαν επιτυχία.

Topic	Words
0	und der die ist mit sehr das man den hier ein für auf nicht nur
1	nom prices refresher lime swim tastebuds drink drinks holiday pumpkin going bean start july
2	title today great forward estate alamo real meeting hospitality contract van time love
3	great good best food place amazing service free awesome delicious order chicken bar nice coffee
4	los muy las con son está para pero del por están servicio hay buena mejor
5	vip stay slow helsinki ghetto sue japanese fashionable swa bay alto omg don't tea paid
6	great place good food service nice best free time coffee people staff love day parking
7	good service les food best pour check top wifi très nice pas great early place
8	established continuously operated family owned bar built original best oldest sauce famous century restaurant
9	wifi очень place для тут есть место super можно good что free все wi-fi только
10	free wifi check hour open happy power outlets work gluten places breakfast order list car
11	chicken delicious good best pork fried great order chocolate cream coffee beef spicy amazing
12	pants wear route esta best san blood francisco list este magazine bowls marrow bone
13	jewish read story family jews true learn find children lives young holocaust years survive life
14	fondu petals burger burrito adobo cha mama sandwich fastrak waffle croque-madame loco california sisig
15	great today nice service gas best good food friendly place awesome clean super fresh waste
16	wi-fi password network 4sqwifi.com open day hours website brass jon swim check band p.m
17	course golf jet leavin plane disc best great seat beers open holes burrito play place
18	een voor met ook het scattando music school lekker muzieklessen van bij maar hier eens
19	não para atendimento muito uma bom mas tem mais melhor bem sempre são pra excelente

Πίνακας 4.10: Παραδείγματα θεμάτων από τα σχόλια χρηστών του Σαν Φρανσίσκο

- Υπολογίζουμε το συνολικό ποσοστό επιτυχίας ως: $\frac{επιτυχιες}{τριαδες} \times 100$

Για $N = 100$ θέματα, η τυχαία προσέγγιση έχει περίπου **1%** επιτυχία για τις top-1 μεθόδους και **2%** επιτυχία για τις top-2 μεθόδους, ανεξάρτητα από τον αριθμό των εκτελέσεων.

Η User-top προσέγγιση

Στον Πίνακα 4.11 παραθέτουμε ποσοστά επιτυχίας για την User-top προσέγγιση. Όπως παρατηρούμε, τα αποτελέσματα είναι σημαντικά καλύτερα όταν χρησιμοποιούμε μοντέλα κατασκευασμένα με τα δεδομένα των χρηστών. Η παρουσία όμως των κοινών σχολίων στα δεδομένα εκπαίδευσης έχει μεγάλη επίδραση στα ποσοστά επιτυχίας και ιδιαίτερα στα δεδομένα της Νέας Υόρκης. Αυτό αποτελεί βασικό πρόβλημα για την προσέγγιση αυτή. Οι top-2 μέθοδοι δίνουν καλύτερα ποσοστά όπως είναι αναμενόμενο.

(α) Νέα Υόρκη

Μοντέλο για τα top tip topics	Όλα τα tips		Χωρίς κοινά	
	Top-1	Top-2	Top-1	Top-2
user model	23,0%	**30,8%**	6,4%	9,7%
venue model	6,0%	10,9%	4,5%	8,1%

(β) Σαν Φρανσίσκο

Μοντέλο για τα top tip topics	Όλα τα tips		Χωρίς κοινά	
	Top-1	Top-2	Top-1	Top-2
user model	26,0%	**39,0%**	14,8%	25,2%
venue model	7,0%	13,7%	8,6%	15,1%

Πίνακας 4.11: Ποσοστά επιτυχίας πρόβλεψης θέματος για την User-top προσέγγιση

Η Venue-top προσέγγιση

Στον Πίνακα 4.12 παραθέτουμε ποσοστά επιτυχίας για την Venue-top προσέγγιση. Όπως παρατηρούμε, εδώ έχουμε καλύτερα αποτελέσματα όταν χρησιμοποιούμε τα μοντέλα που εκπαιδεύτηκαν με τα δεδομένα των τοποθεσιών, σε αντιστοιχία με τα αποτελέσματα της προηγούμενης προσέγγισης. Εδώ αξίζει να παρατηρήσουμε ότι τα ποσοστά είναι σχετικά καλά ακόμα και αν χρησιμοποιήσουμε τα μοντέλα που δεν περιλαμβάνουν τα κοινά σχόλια μεταξύ χρηστών και τοποθεσιών.

(a) Νέα Υόρκη

Μοντέλο για τα top tip topics	Όλα τα tips		Χωρίς κοινά	
	Top-1	Top-2	Top-1	Top-2
user model	8,5%	24,7%	6,3%	12,2%
venue model	29,0%	**36,3%**	20,0%	29,1%

(b) Σαν Φρανσίσκο

Μοντέλο για τα top tip topics	Όλα τα tips		Χωρίς κοινά	
	Top-1	Top-2	Top-1	Top-2
user model	14,0%	29,9%	18,7%	26,5%
venue model	38,0%	**51,7%**	24,1%	34,7%

Πίνακας 4.12: Ποσοστά επιτυχίας πρόβλεψης θέματος για την Venue-top προσέγγιση

Οι Συνδυασμένες προσέγγισεις

Στους Πίνακες 4.13 και 4.14 παραθέτουμε ποσοστά επιτυχίας για τις συνδυασμένες προσεγγίσεις. Όπως μπορούμε να παρατηρήσουμε και στις δύο περιπτώσεις η χρήση των μοντέλων που εκπαιδεύτηκαν με τα δεδομένα των τοποθεσιών γενικά δίνουν καλύτερα ποσοστά επιτυχίας, κυρίως για το σύνολο δεδομένων του Σαν Φρανσίσκο. Για ακόμη μια φορά η παρουσία

των κοινών σχολίων έχει επίδραση στα αποτελέσματα. Γενικά παρατηρούμε ότι η απλούστερη λύση συνδυασμού 'Both' δίνει καλά αποτελέσματα σε όλες σχεδόν τις περιπτώσεις.

(a) Νέα Υόρκη

Μοντέλο για τα top tip topics	Όλα τα tips		Χωρίς κοινά	
	Top-1	Top-2	Top-1	Top-2
user model	15,5%	25,5%	5,5%	6,6%
venue model	23,5%	**26,1%**	11,5%	16,8%

(b) Σαν Φρανσίσκο

Μοντέλο για τα top tip topics	Όλα τα tips		Χωρίς κοινά	
	Top-1	Top-2	Top-1	Top-2
user model	19,0%	27,3%	17,5%	27,9%
venue model	31,0%	**36,3%**	18,7%	30,4%

Πίνακας 4.13: Ποσοστά επιτυχίας πρόβλεψης θέματος για την συνδυασμένη προσέγγιση (γινόμενο)

(a) Νέα Υόρκη

Μοντέλο για τα top tip topics	Όλα τα tips (Top-2)	Χωρίς κοινά (Top-2)
user model	30,6%	10,4%
venue model	30,1%	22,7%

(b) Σαν Φρανσίσκο

Μοντέλο για τα top tip topics	Όλα τα tips (Top-2)	Χωρίς κοινά (Top-2)
user model	31,3%	17,4%
venue model	42,3%	28,9%

Πίνακας 4.14: Ποσοστά επιτυχίας πρόβλεψης θέματος για την συνδυασμένη προσέγγιση (Both)

Οι Συνεργατικές προσεγγίσεις (Συνεργατικό Φιλτράρισμα)

Στους Πίνακες 4.15 και 4.16 παραθέτουμε ποσοστά επιτυχίας για τις συνεργατικές προσεγγίσεις χρηστών και τοποθεσιών. Εδώ έχουμε συμπεριλάβει περισσότερους συνδυασμούς χρήσης των διαθέσιμων θεματικών μοντέλων και αυτό γιατί σε αυτές τις προσεγγίσεις είναι δυνατή η χρήση διαφορετικού μοντέλου για την εξαγωγή της κατανομής των θεμάτων για τα σχόλια και διαφορετικού μοντέλου για την εξαγωγή των κατανομών των θεμάτων που χρησιμοποιούνται για τον υπολογισμό της ομοιότητας μεταξύ χρηστών και μεταξύ τοποθεσιών.

Τα ποσοστά επιτυχίας σε αυτές τις προσεγγίσεις δεν είναι τόσο καλά όσο στις προηγούμενες, έχουν όμως ένα πλεονέκτημα. Επηρεάζονται σχετικά λιγότερο από την ύπαρξη των κοινών σχολίων μεταξύ χρηστών και τοποθεσιών στα δεδομένα εκπαίδευσης του μοντέλου.

Στην περίπτωση της συνεργατικής προσέγγισης χρηστών έχουμε γενικά καλύτερα αποτελέσματα όταν χρησιμοποιούμε τα μοντέλα τοποθεσιών για τις κατανομές θεμάτων των σχολίων ενώ στην περίπτωση της συνεργατικής προσέγγισης τοποθεσιών έχουμε σχετικά καλύτερα αποτελέσματα όταν χρησιμοποιούμε τα μοντέλα χρηστών για τις κατανομές θεμάτων των σχολίων.

Μια παρατήρηση που μπορούμε να κάνουμε εδώ είναι ότι η επιλογή του μοντέλου για τον υπολογισμό της ομοιότητας μεταξύ χρηστών ή μεταξύ τοποθεσιών δε φαίνεται να παίζει ιδιαίτερο ρόλο στα ποσοστά επιτυχίας των δύο προσεγγίσεων. Μικρό προβάδισμα ωστόσο έχουμε όταν χρησιμοποιούμε τα μοντέλα με τα σύνολα δεδομένων των χρηστών για τον υπολογισμό της ομοιότητας μεταξύ χρηστών και αντίστοιχα τα μοντέλα με τα σύνολα δεδομένων των τοποθεσιών για τον υπολογισμό της ομοιότητας μεταξύ των τοποθεσιών.

(a) Νέα Υόρκη

Μοντέλο για τα top tip topics	Μοντέλο για την ομοιότητα *u-u*			
	User model		Venue model	
	Όλα τα tips	Χωρίς κοινά	Όλα τα tips	Χωρίς κοινά
user model	5,5%	**6,0%**	4,0%	**5,5%**
venue model	16,5%	**11,1%**	17,0%	**11,3%**

(b) Σαν Φρανσίσκο

Μοντέλο για τα top tip topics	Μοντέλο για την ομοιότητα *u-u*			
	user model		venue model	
	Όλα τα tips	Χωρίς κοινά	Όλα τα tips	Χωρίς κοινά
user model	12,0%	**9,7%**	10,0%	**9,5%**
venue model	24,5%	**13,7%**	24,5%	**14,0%**

Πίνακας 4.15: Ποσοστά επιτυχίας πρόβλεψης θέματος για την συνεργατική προσέγγιση χρηστών

(a) Νέα Υόρκη

Μοντέλο για τα top tip topics	Μοντέλο για την ομοιότητα *ν-ν*			
	user model		venue model	
	Όλα τα tips	Χωρίς κοινά	Όλα τα tips	Χωρίς κοινά
user model	14,0%	**6,0%**	14,5%	**6,5%**
venue model	5,0%	**6,1%**	8,5%	**7,5%**

(b) Σαν Φρανσίσκο

Μοντέλο για τα top tip topics	Μοντέλο για την ομοιότητα *ν-ν*			
	user model		venue model	
	Όλα τα tips	Χωρίς κοινά	Όλα τα tips	Χωρίς κοινά
user model	20,0%	**12,5%**	19,0%	**12,0%**
venue model	6,0%	**5,5%**	12,0%	**9,0%**

Πίνακας 4.16: Ποσοστά επιτυχίας πρόβλεψης θέματος για την συνεργατική προσέγγιση τοποθεσιών

Σύγκριση και Συμπέρασμα

Στον Πίνακα 4.17 παραθέτουμε συγκεντρωμένα αποτελέσματα από όλες τις προσεγγίσεις προκειμένου να κάνουμε μια σύγκριση μεταξύ τους. Επιλέξαμε από κάθε προσέγγιση τα αποτελέσματα για τα μοντέλα χωρίς τα κοινά σχόλια και για τις top-1 μεθόδους, για να είναι πιο αντικειμενική η σύγκριση. Για τις συνεργατικές μεθόδους επιλέξαμε τα καλύτερα αποτελέσματα ανεξάρτητα από το το μοντέλο που χρησιμοποιήθηκε για τον υπολογισμό ομοιότητας μεταξύ χρηστών ή μεταξύ τοποθεσιών. Επιπλέον, προσθέσαμε στην τελευταία στήλη το ποσοστό επιτυχίας που θα είχαμε αν προβλέπαμε πάντα το πιο πιθανό θέμα για το αντίστοιχο σύνολο δεδομένων και μοντέλο. Στα Σχήματα Π.3 και Π.4 του Παραρτήματος παρουσιάζουμε ενδεικτικά κάποιες κατανομές των πιο πιθανών θεμάτων.

Όπως παρατηρούμε, τα μεγαλύτερα ποσοστά επιτυχίας για κάθε σύνολο δεδομένων έχει η Venue-top μέθοδος με χρήση των μοντέλων με τα δεδομένα των τοποθεσιών, ενώ τα μικρότερα η User-top με χρήση επίσης των ίδιων μοντέλων. Το ίδιο μοτίβο ακολουθούν και οι top-2 μέθοδοι. Γενικά μπορούμε να θεωρήσουμε ποσοστά της τάξης του 20% και παραπάνω ως αρκετά ικανοποιητικά, αν συνυπολογίσουμε ότι το τυχαίο πείραμα θα μας έδινε ποσοστά της τάξης του 1-2%. Καταλήγουμε λοιπόν στο αποτέλεσμα ότι τα δεδομένα των τοποθεσιών μας δίνουν πιο ποιοτικά θέματα και κατ' επέκταση καλύτερα ποσοστά πρόβλεψης. Οι συνεργατικές μέθοδοι δίνουν σχετικά μικρά ποσοστά επιτυχίας και ένας από τους λόγους φαίνεται να είναι η αδυναμία καλού υπολογισμού της ομοιότητας μεταξύ των χρηστών και μεταξύ των τοποθεσιών. Αυτό συμβαίνει και λόγω της φύσης των μοντέλων αλλά και λόγω του σχετικά μικρού αριθμού χρηστών και τοποθεσιών που δεν μας επιτρέπει πάντα να βρούμε σχόλια που

να πληρούν τις προϋποθέσεις που θέσαμε στις προσεγγίσεις αυτές.

Πόλη	Μοντέλο	User-top	Venue-top	Γινόμενο	Collab-U	Collab-V	Global-top
Νέα Υόρκη	Χρήστες	6,4%	6,3%	5,5%	6,0%	6,5%	5,4%
”	Τοποθεσίες	4,5%	20,0%	11,5	11,3%	7,5%	5,0%
Σαν Φρανσίσκο	Χρήστες	14,8%	18,7%	17,5%	9,7%	12,5%	7,9%
”	Τοποθεσίες	8,6%	24,1%	18,7%	14,0%	9,0%	3,9%

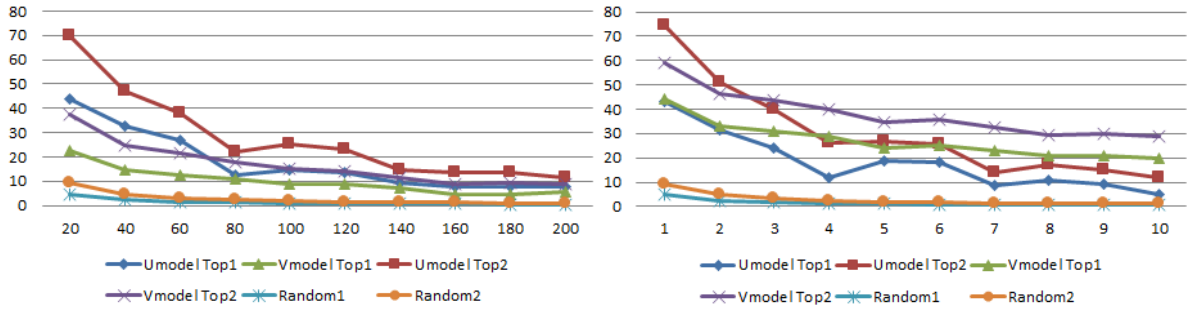
Πίνακας 4.17: Σύγκριση top-1 προσεγγίσεων

4.3.3 Η επίδραση του αριθμού των Θεμάτων

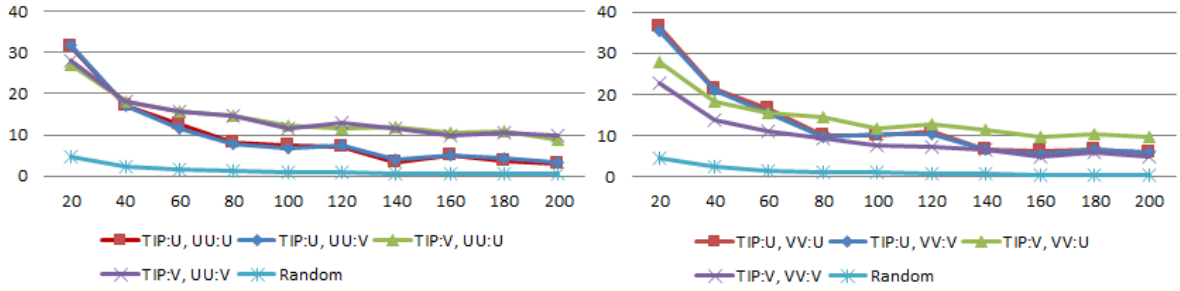
Όπως έχουμε αναφέρει στο προηγούμενο κεφάλαιο, ένα από τα βασικά προβλήματα των θεματικών μοντέλων είναι να αποφασίσει κάποιος τον ακριβή αριθμό των θεμάτων εκ των προτέρων. Στα προηγούμενα πειράματα χρησιμοποιήσαμε μοντέλα με σταθερό αριθμό από 100 θέματα κατά την κατασκευή τους. Προκειμένου να εξετάσουμε πώς ο αριθμός των θεμάτων επηρεάζει τα ποσοστά επιτυχίας, εκπαιδεύσαμε 10 διαφορετικά μοντέλα χρησιμοποιώντας το σύνολο δεδομένων από το Σαν Φρανσίσκο χωρίς τα κοινά σχόλια μέσα.

Στα Σχήματα 4.5 και 4.6 μπορούμε να δούμε πώς ο αριθμός των θεμάτων (από 20 μέχρι 200) επηρέασε τα ποσοστά επιτυχίας από τις User-top, Venue-top και Συνεργατικές προσεγγίσεις, αφήνοντας όλες τις άλλες παραμέτρους ίδιες. Όπως μπορούμε να παρατηρήσουμε, για μικρό αριθμό θεμάτων έχουμε σχετικά μεγάλα ποσοστά επιτυχίας, πράγμα που είναι αναμενόμενο. Στις User-top και Venue-top προσεγγίσεις υπάρχει μια γρήγορη μείωση στα ποσοστά επιτυχίας από τα 20 μέχρι τα 80 θέματα και στη συνέχεια υπάρχει μια σχετική σταθεροποίηση με μια πολύ μικρή αύξηση γύρω στα 100-120 θέματα και αυτή είναι κοινή τάση για όλα τα μοντέλα. Στις συνεργατικές προσεγγίσεις τα ποσοστά ακολουθούν την ίδια τάση με σχετική σταθεροποίηση μετά τα 80 θέματα και χωρίς την μικρή αύξηση που παρατηρήσαμε προηγουμένως. Στα Σχήματα επίσης περιλαμβάνονται και τα ποσοστά επιτυχίας των τυχαίων προσεγγίσεων που είναι πολύ χαμηλά.

Επιλέξαμε τα 100 θέματα για τα πειράματά μας, προκειμένου να έχουμε πιο αντικειμενικά αποτελέσματα και μικρά ποσοστά επιτυχίας ακόμα και στην τυχαία προσέγγιση. Ωστόσο αν θέλαμε τα θέματα να είναι πιο γενικά, θα μπορούσαμε να χρησιμοποιήσουμε μικρότερο αριθμό.



Σχήμα 4.5: User-top (Αριστερά) και Venue-top (Δεξιά) Προσεγγίσεις (Σαν Φρανσίσκο, Χωρίς κοινά σχόλια)



Σχήμα 4.6: Συνεργατικές Προσεγγίσεις Χρηστών (Αριστερά) και Τοποθεσιών (Δεξιά) (Σαν Φρανσίσκο, Χωρίς κοινά σχόλια)

4.4 Εκτίμηση πρόβλεψης με Κατηγοριοποιητές

4.4.1 Προετοιμασία Συνόλων Δεδομένων

Για την εκτέλεση των πειραμάτων της ενότητας αυτής χρειαζόμαστε α) τα δεδομένα των τριάδων user-venue-tip πάνω στα οποία θα εφαρμόσουμε την πρόβλεψη και β) τα σύνολα δεδομένων εκπαίδευσης και ελέγχου των κατηγοριοποιητών μας.

α) Τα δεδομένα πάνω στα οποία εφαρμόζουμε την πρόβλεψη είναι τριάδες user-venue-tip από το σύνολο δεδομένων του Σαν Φρανσίσκο, το οποίο είχαμε επιλέξει τόσο για τη δημιουργία των LDA μοντέλων όσο και για την πρόβλεψη θέματος στα αντίστοιχα πειράματα. Πρόκειται για τοποθεσίες και χρήστες με περισσότερα από 40 σχόλια έκαστος. Τα στατιστικά τους βρίσκονται στον Πίνακα 4.6. Επίσης στο κείμενο έγινε και η προεπεξεργασία που εξηγούμε στην Ενότητα 4.1.3.

β) Για τα δεδομένα εκπαίδευσης αρχικά επιλέξαμε ένα σύνολο από τυχαία σχόλια από το

αρχικό σύνολο δεδομένων του Σαν Φρανσίσκο, τα οποία όμως δεν περιλαμβάνονται στο σύνολο (α). Στη συνέχεια σε περίπου 1000 από αυτά αναθέσαμε κατηγορίες από τις ακόλουθες: {*food, drink, place, service, price*}. Στον Πίνακα 4.18 παραθέτουμε μερικά παραδείγματα σχολίων ανά κατηγορία. Η ανάθεση έγινε σύμφωνα με τη θεματολογία του κάθε σχολίου, προσέχοντας να είναι ξεκάθαρος ο προσανατολισμός του χρήστη που το γράφει. Επίσης δώσαμε προσοχή στην τήρηση της αναλογίας των κατηγοριών. Το σύνολο αυτό το χωρίσαμε με τυχαίο τρόπο σε σύνολο εκπαίδευσης και σύνολο ελέγχου με αναλογία περίπου 75% και 25% αντίστοιχα.

Για τα δεδομένα εκπαίδευσης των κατηγοριοποιητών μίας κατηγορίας δημιουργήσαμε από το παραπάνω σύνολο πέντε μικρότερα. Για κάθε σύνολο επιλέξαμε τα 200 σχόλια που ανήκουν στην κάθε κατηγορία και προσθέσαμε σε αυτά άλλα 200 τυχαία από τα υπόλοιπα που ανήκουν σε άλλες κατηγορίες. Το κάθε σύνολο το χωρίσαμε με τυχαίο τρόπο σε σύνολο εκπαίδευσης και σύνολο ελέγχου με αναλογία περίπου 75% και 25% αντίστοιχα, προσέχοντας και πάλι την αναλογία θετικών και αρνητικών δειγμάτων στο κάθε σύνολο.

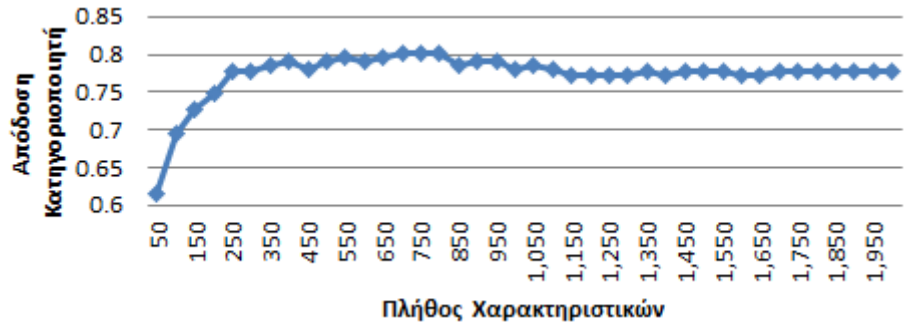
food	The tuna tower (with mango and avocado) is splendid indeed Jicama salad with lots of hot sauce is a refreshing, healthy lunch
drink	Get the French Champagne Cocktail Coffee is strong and excellent here
place	The view from the pureprofile office in San Francisco - looks like stormy weather for the night Fun place with outdoor seating and at this exact moment, an awesome Giants game
service	Super knowledgeable staff, so take advantage of them if you're looking to start a new series Expect to wait 2 hours on a normal weekend
price	Too expensive for my yuppie self Show your SF Bike Coalition membership card, and get 10% off

Πίνακας 4.18: Παραδείγματα σχολίων ανά κατηγορία

4.4.2 Εκπαίδευση Κατηγοριοποιητών

Επιλογή Χαρακτηριστικών

Η επιλογή των χαρακτηριστικών που θα χρησιμοποιηθούν στην εκπαίδευση ενός κατηγοριοποιητή είναι σημαντική στην απόδοση του. Μία συνήθης τακτική όταν πρόκειται για κατηγοριοποίηση κειμένου, την οποία και εφαρμόσαμε, είναι το κάθε χαρακτηριστικό να πρόκειται και για μία λέξη. Πιο συγκεκριμένα το κάθε χαρακτηριστικό είναι η παρουσία ή η απουσία (True ή False) της συγκεκριμένης λέξης μέσα στο στοιχείο προς εκπαίδευση ή προς κατηγοριοποίηση.



Σχήμα 4.7: Απόδοση κατηγοριοποιητή ανάλογα με το πλήθος των χαρακτηριστικών

Για να επιλέξουμε τις κατάλληλες λέξεις:

- Βρήκαμε όλες τις λέξεις που εμφανίζονται μέσα στο σύνολο δεδομένων μας, υπολογίσαμε τη συχνότητα εμφάνισής της κάθε μίας και τις ταξινομήσαμε με βάση αυτή.
- Εκπαιδεύσαμε κατηγοριοποιητές χρησιμοποιώντας τις top-n λέξεις κάθε φορά για n από 50-2000, διατηρώντας σταθερό το σύνολο εκπαίδευσης και ελέγχου που αναφέραμε παραπάνω. Από τα αποτελέσματα που φαίνονται και στο Σχήμα 4.7 βλέπουμε ότι για την περίπτωση του ενιαίου κατηγοριοποιητή, έχουμε την καλύτερη απόδοση χρησιμοποιώντας τις 700-800 πιο συχνές λέξεις ως χαρακτηριστικά. Αντίστοιχα για τους μικρότερους ξεχωριστούς κατηγοριοποιητές της δεύτερης προσέγγισης έχουμε καλύτερη απόδοση χρησιμοποιώντας λιγότερα χαρακτηριστικά. Αποφασίσαμε λοιπόν να χρησιμοποιήσουμε τις πρώτες 750 λέξεις ως χαρακτηριστικά για τον ενιαίο κατηγοριοποιητή και 500 λέξεις για τους υπόλοιπους. Αυτά τα νούμερα βέβαια έχουν να κάνουν με την απόδοση στο συγκεκριμένο σύνολο δεδομένων.

Ακρίβεια Κατηγοριοποιητών

Για τα πειράματα αυτής της ενότητας χρησιμοποιούμε την υλοποίηση του Naive Bayes κατηγοριοποιητή που βρίσκεται στο πακέτο εργαλείων επεξεργασίας φυσικής γλώσσας NLTK της Python [43].

Η ακρίβεια (ποσοστό των στοιχείων από το σύνολο δεδομένων ελέγχου που ταξινομήθηκαν σωστά) του ενιαίου κατηγοριοποιητή που εκπαιδεύσαμε και χρησιμοποιήσαμε είναι περίπου **80%**. Στον Πίνακα 4.19 παραθέτουμε την ακρίβεια των κατηγοριοποιητών που χρησιμοποιήσαμε για την περίπτωση του ενός κατηγοριοποιητή ανά κατηγορία. Επίσης στον Πίνακα 4.20 παραθέτουμε τα χαρακτηριστικά με τη μεγαλύτερη 'βαρύτητα' στην κατηγοριοποίηση για τον ενιαίο κατηγοριοποιητή.

Κατηγορία	food	drink	place	service	price
Ακρίβεια	0.89	0.85	0.73	0.80	0.87

Πίνακας 4.19: Ακρίβεια Κατηγοριοποιητών

	Χαρακτηριστικό	Κατηγορία		Χαρακτηριστικό	Κατηγορία		Χαρακτηριστικό	Κατηγορία
1	coffee	drink	**11**	hour	price	**21**	free	price
2	place	place	**12**	juice	drink	**22**	super	service
3	chicken	food	**13**	iced	drink	**23**	bring	price
4	staff	service	**14**	views	place	**24**	expensive	price
5	slow	service	**15**	spicy	food	**25**	espresso	drink
6	tea	drink	**16**	amazing	food	**26**	cheese	food
7	service	service	**17**	salad	food	**27**	beef	food
8	cocktail	drink	**18**	fried	food	**28**	music	place
9	cheap	price	**19**	drink	drink	**29**	bar	drink
10	nice	place	**20**	ice	drink	**30**	orange	drink

Πίνακας 4.20: Χαρακτηριστικά ενιαίου κατηγοριοποιητή με τη μεγαλύτερη βαρύτητα

4.4.3 Αποτελέσματα πρόβλεψης με χρήση ενιαίου κατηγοριοποιητή

Για να έχουμε και εδώ κάποιο μέτρο σύγκρισης, προκειμένου να αξιολογήσουμε τα ποσοστά επιτυχίας θα υπολογίσουμε το ποσοστό επιτυχίας που θα είχαμε εκτελούσαμε το πείραμα μας με τυχαίο τρόπο. Στην περίπτωση του ενιαίου κατηγοριοποιητή, η διαδικασία είναι η ίδια με αυτήν που περιγράψαμε για τα θεματικά μοντέλα στην Ενότητα 4.3.2.

Για πέντε κατηγορίες και πολλές εκτελέσεις του πειράματος (όσες είναι οι διαθέσιμες τριάδες) το ποσοστό επιτυχίας θα ήταν γύρω στο **20%**, το οποίο βέβαια είναι από μόνο του αρκετά μεγάλο ποσοστό.

Στον Πίνακα 4.21 παραθέτουμε ποσοστά επιτυχίας για τις προσεγγίσεις με χρήση ενός ενιαίου κατηγοριοποιητή. Όπως παρατηρούμε, τα αποτελέσματα είναι αρκετά καλύτερα από αυτά του τυχαίου πειράματος. Επίσης επιβεβαιώνεται και εδώ η διαπίστωση ότι τα σχόλια έχουν μεγαλύτερη συσχέτιση με την ίδια την τοποθεσία και λιγότερη με τον χρήστη που το δημιούργησε.

	Τυχαίο Πείραμα	User-top	Venue-top	Both
Επιτυχία	20%	40,5%	48,5%	62,8%

Πίνακας 4.21: Αποτελέσματα πρόβλεψης με χρήση ενιαίου κατηγοριοποιητή

4.4.4 Αποτελέσματα πρόβλεψης με χρήση ενός κατηγοριοποιητή ανά κατηγορία

Στον Πίνακα 4.22 παραθέτουμε ποσοστά επιτυχίας για τις προσεγγίσεις με χρήση ενός κατηγοριοποιητή ανά κατηγορία. Όπως μπορούμε να παρατηρήσουμε, υπάρχει μια μικρή βελτίωση στα ποσοστά επιτυχίας της πρόβλεψης κατά 2% περίπου για κάθε μέθοδο αντίστοιχα, σε σχέση με την προσέγγιση του ενιαίου κατηγοριοποιητή. Δεδομένου ότι χρησιμοποιούμε ακριβώς τα ίδια δεδομένα, η βελτίωση αυτή δε μπορεί να θεωρηθεί αμελητέα.

	Τυχαίο Πείραμα	User-top	Venue-top	Both
Επιτυχία	20%	42,6%	50,7%	62,9%

Πίνακας 4.22: Αποτελέσματα πρόβλεψης με χρήση πολλών κατηγοριοποιητών

4.5 Θέματα προς συζήτηση

Η χρήση των πιθανολογικών θεματικών μοντέλων για την εύρεση θεμάτων από μικρά κείμενα όπως αυτά που είδαμε στην εργασία αυτή είναι αναμφισβήτητα μια ελκυστική λύση. Στην πραγματικότητα βέβαια υπάρχουν πολλά ζητήματα, όπως η ευαισθησία των παραγόμενων μοντέλων στα δεδομένα εισόδου, καθώς μικρές αλλαγές ακόμα και στον τρόπο της προετοιμασίας των κειμένων μπορεί να προκαλέσει αλλαγές στα παραγόμενα θέματα. Το ίδιο συμβαίνει και με την προσθήκη ή αφαίρεση κειμένων από το σύνολο εκπαίδευσης, ειδικά αν αυτό είναι σχετικά μικρό. Άλλο σημαντικό ζήτημα είναι ότι τα θέματα δεν είναι άμεσα κατανοητά αν δεν υπάρξει κάποιου είδους ανθρώπινη παρέμβαση όπως η προσθήκη ετικετών σε αυτά. Από την άλλη όμως είναι πολύ χρήσιμα εργαλεία όταν έχουμε μεγάλες συλλογές κειμένων που έχουν κάποιες ιδιαιτερότητες και υπάρχει η ανάγκη για εξερεύνηση της κρυφής δομής και των θεμάτων τους.

Από την άλλη πλευρά, η χρήση επιβλεπόμενων μεθόδων μάθησης όπως η κατηγοριοποίηση ίσως είναι πιο ενδεδειγμένη λύση αν γνωρίζουμε την δομή των δεδομένων και τα θέματα που περιλαμβάνουν. Μας οδηγεί σε θέματα κατανοητά και άμεσα αναγνωρίσιμα αλλά είναι συχνά μια αρκετά επίπονη και χρονοβόρα διαδικασία.

Μια πιθανή επέκταση είναι ο συνδυασμός επιβλεπόμενων και μη επιβλεπόμενων μεθόδων μάθησης. Τα αποτελέσματα των θεματικών μοντέλων θα μπορούσαν να χρησιμοποιηθούν σαν επιπλέον χαρακτηριστικά σε μεθόδους κατηγοριοποίησης με πιθανή αύξηση της απόδοσης τους.

Κεφάλαιο 5

Συμπεράσματα και Μελλοντικές Επεκτάσεις

Στην Εργασία αυτή αρχικά κάναμε μια εισαγωγή στα Διαδικτυακά Κοινωνικά Δίκτυα Βασισμένα σε Τοποθεσία (LBSN). Αναφέραμε τα ιδιαίτερα χαρακτηριστικά τους, παρουσιάσαμε τους τρόπους που ενσωματώνεται η πληροφορία σε αυτά και τα κατατάξαμε σε κάποιες κατηγορίες με βάση αυτούς, αναφέροντας και σχετικά παραδείγματα.

Στη συνέχεια αναφερθήκαμε σε μια πληθώρα ερευνητικών θεμάτων που έχουν προκύψει από τέτοιου είδους δίκτυα και σε αντίστοιχες εργασίες που έχουν ασχοληθεί με αυτά. Επίσης αναφέραμε σχετικές εργασίες που έχουν ασχοληθεί με κείμενο, με θέματα όπως προσθήκη ετικετών σε πόρους, προτάσεις βασισμένες σε κείμενο, και χρήση θεματικών μοντέλων σε μικρά κείμενα.

Στο κύριο μέρος της εργασίας ασχοληθήκαμε με το θέμα της πρόβλεψης του θέματος σε μικρά κείμενα (tips) που αφήνουν οι χρήστες στις τοποθεσίες (venues) που επισκέπτονται, μέσω των LBSN. Αρχικά προσεγγίσαμε το θέμα με χρήση θεματικών μοντέλων. Αφού κάναμε μια εισαγωγή σε αυτά τα μοντέλα και ιδιαίτερα στο LDA παρουσιάσαμε κάποιες προσεγγίσεις στο θέμα, που έχουν ως στόχο να διαπιστώσουμε αν είναι δυνατόν να προβλέψουμε το θέμα των κειμένων αυτών με τη βοήθεια μιας σειράς από θέματα που παράγονται αυτοματοποιημένα από τον αλγόριθμο. Οι προσεγγίσεις μας βασίζονται σε απλοϊκές θεωρήσεις όπως η επιμονή των χρηστών να γράφουν είτε για συγκεκριμένα θέματα, είτε σχετικά με ιδιαιτερότητες των τοποθεσιών. Σε άλλες προσεγγίσεις λαμβάνουμε υπόψη τη συμπεριφορά όμοιων χρηστών ή όμοιων τοποθεσιών στα ζητούμενα.

Στη συνέχεια προσεγγίσαμε το θέμα με χρήση κατηγοριοποιητών, αφού κάναμε μία εισαγωγή σε αυτούς, σε σημαντικά θέματα που αφορούν την χρήση τους και ιδιαίτερα στον Naive Bayes κατηγοριοποιητή τον οποίο και χρησιμοποιούμε. Προσεγγίσαμε το θέμα με χρήση ενός

αλλά και πολλαπλών κατηγοριοποιητών, αυτή τη φορά με χρήση αντίστοιχων δεδομένων που έχουν έχουν κατηγοριοποιηθεί από πριν σε ”θέματα”, σε αντίθεση με τα θεματικά μοντέλα που η παραγωγή τους είναι αυτοματοποιημένη.

Στο πειραματικό κομμάτι της εργασίας τώρα, αρχικά ασχολούμαστε με τη συλλογή των δεδομένων που θα χρησιμοποιήσουμε. Μελετάμε την προγραμματιστική διεπαφή (API) του δικτύου Foursquare, τον τρόπο που είναι οργανωμένα τα δεδομένα, τη διαδικασία με την οποία έχουμε πρόσβαση σε αυτά καθώς και τους περιορισμούς που τίθενται από το ίδιο το δίκτυο. Στη συνέχεια παρουσιάζουμε πληροφορίες για τα δεδομένα που συλλέξαμε και διάφορα στατιστικά σχετικά με αυτά.

Συνεχίσαμε με την κατασκευή θεματικών μοντέλων με χρήση του LDA αλγορίθμου, αρχικά αναφέροντας τις επιλογές μας σχετικά με τα δεδομένα και παρουσιάζοντας παραδείγματα και στη συνέχεια υλοποιώντας τις προσεγγίσεις μας πάνω στα πραγματικά δεδομένα με χρήση αυτών. Παρουσιάζουμε τα αποτελέσματα και δείχνουμε ότι όντως είναι δυνατή η πρόβλεψη με κάποιες μεθόδους που βασίζονται στην συνάφεια των κειμένων με τις τοποθεσίες να υπερέχουν. Δείχνουμε επίσης την ευαισθησία των προβλέψεών μας ανάλογα με τα δεδομένα που χρησιμοποιούνται.

Τέλος, προετοιμάζουμε τα σύνολα δεδομένων και εκπαιδεύουμε ένα σύνολο κατηγοριοποιητών τους οποίους αξιολογούμε και στη συνέχεια χρησιμοποιούμε για να εφαρμόσουμε τις προσεγγίσεις μας. Παρουσιάζουμε τα αποτελέσματα και επιβεβαιώνουμε τη λειτουργία των μεθόδων μας, με μικρή υπεροχή της προσέγγισης με πολλαπλούς κατηγοριοποιητές. Καταλήγουμε και εδώ στο συμπέρασμα ότι τα σχόλια σχετίζονται περισσότερο με τις τοποθεσίες και κάπως λιγότερο με τους χρήστες. Αυτό εκτός από την προφανή ερμηνεία ότι οι χρήστες γράφουν για συγκεκριμένα δυνατά ή αδύναμα σημεία της κάθε τοποθεσίας, μπορεί να έχει και διάφορες άλλες ερμηνείες για τη συμπεριφορά τους, όπως για παράδειγμα ότι ένας χρήστης μπορεί να επηρεάζεται από τα σχόλια των άλλων όταν γράφει το δικό του.

Για την παρούσα εργασία υπάρχουν διάφορες μελλοντικές επεκτάσεις που μπορούν να υλοποιηθούν και ιδιαίτερα στο κομμάτι της πρόβλεψης με χρήση κατηγοριοποιητών. Μερικά παραδείγματα είναι η χρήση συνδυασμένων κατηγοριών στα σχόλια, η αυτοματοποιημένη κατηγοριοποίηση με λέξεις κλειδιά, μια πιο έξυπνη μέθοδος επιλογής χαρακτηριστικών, καθώς και πιο έξυπνους τρόπους συνδυασμού των αποτελεσμάτων. Αρκετή δουλειά μπορεί να γίνει επίσης για να αντιμετωπιστούν οι ιδιομορφίες των κειμένων αυτών, καθώς συνήθως δεν έχουν ολοκληρωμένη δομή προτάσεων και χρησιμοποιούν συντομογραφίες και καθημερινή γλώσσα.

Βιβλιογραφία

[1] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. ”Latent dirichlet allocation.” the Journal of machine Learning research 3 (2003): 993-1022.

[2] Steyvers, Mark, and Tom Griffiths. ”Probabilistic topic models.” Handbook of latent semantic analysis 427.7 (2007): 424-440.

[3] David M. Blei. 2012. Probabilistic topic models. Commun. ACM 55, 4 (April 2012), 77-84.

[4] Frigyik, Bela A., Amol Kapila, and Maya R. Gupta. ”Introduction to the Dirichlet distribution and related processes.” Department of Electrical Engineering, University of Washignton, UWEETR-2010-0006 (2010).

[5] Ziyu Guan, Jiajun Bu, Qiaozhu Mei, Chun Chen, and Can Wang. 2009. Personalized tag recommendation using graph-based ranking on multi-type interrelated objects. In Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval (SIGIR ’09)

[6] Shepitsen, Andriy, et al. ”Personalized recommendation in social tagging systems using hierarchical clustering.” Proceedings of the 2008 ACM conference on Recommender systems. ACM, 2008.

[7] Jäschke, Robert, et al. ”Tag recommendations in folksonomies.” Knowledge Discovery in Databases: PKDD 2007. Springer Berlin Heidelberg, 2007. 506-514.

[8] Poirier, Damien, et al. ”Towards text-based recommendations.” Adaptivity, Personalization and Fusion of Heterogeneous Information. LE CENTRE DE HAUTES ETUDES INTERNATIONALES D’INFORMATIQUE DOCUMENTAIRE, 2010.

[9] Mak, Harry, Irena Koprinska, and Josiah Poon. ”Intimate: A web-based movie recommender using text categorization.” Web Intelligence, 2003. WI 2003. Proceedings. IEEE/WIC International Conference on. IEEE, 2003.

[10] Zhao, Wayne Xin, et al. ”Comparing twitter and traditional media using topic models.” Advances in Information Retrieval. Springer Berlin Heidelberg, 2011. 338-349.

[11] Hong, Liangjie, and Brian D. Davison. ”Empirical study of topic modeling in twitter.” Proceedings of the First Workshop on Social Media Analytics. ACM, 2010.

[12] Ramage, Daniel, Susan T. Dumais, and Daniel J. Liebling. ”Characterizing Microblogs with Topic Models.” ICWSM 5.4 (2010): 130-137.

[13] Hall, David, Daniel Jurafsky, and Christopher D. Manning. ”Studying the history of ideas using topic models.” Proceedings of the conference on empirical methods in natural language processing. Association for Computational Linguistics, 2008.

[14] Huynh, Tâm, Mario Fritz, and Bernt Schiele. ”Discovery of activity patterns using topic models.” Proceedings of the 10th international conference on Ubiquitous computing. ACM, 2008.

[15] Hoffman, Matthew, Francis R. Bach, and David M. Blei. ”Online learning for latent dirichlet allocation.” advances in neural information processing systems. 2010.

[16] Steven Bird, Ewan Klein, and Edward Loper. Natural Language Processing with Python. O’Reilly Media, 2009

[17] Zheng, Yu. ”Tutorial on location-based social networks.” Proceedings of the 21st international conference on World wide web, WWW. Vol. 12. 2012.

[18] Zheng, Yu, and Xiaofang Zhou. Computing with spatial trajectories. Springer Science & Business Media, 2011.

[19] Su, Xiaoyuan, and Taghi M. Khoshgoftaar. ”A survey of collaborative filtering techniques.” Advances in artificial intelligence 2009 (2009): 4.

[20] Hotho, Andreas, et al. ”Folkrank: A ranking algorithm for folksonomies.” LWA. Ed. Klaus-Dieter Althoff. Vol. 1. 2006.

[21] Li, Quannan, et al. ”Mining user similarity based on location history.” Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems. ACM, 2008.

[22] Cranshaw, Justin, et al. ”Bridging the gap between physical location and online social networks.” Proceedings of the 12th ACM international conference on Ubiquitous computing. ACM, 2010.

[23] Zheng, Yu, et al. ”Recommending friends and locations based on individual location history.” ACM Transactions on the Web (TWEB) 5.1 (2011): 5.

[24] Xiao, Xiangye, et al. ”Finding similar users using category-based location history.” Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems. ACM, 2010.

[25] Hung, Chih-Chieh, Chih-Wen Chang, and Wen-Chih Peng. ”Mining trajectory profiles for discovering user communities.” Proceedings of the 2009 International Workshop on Location Based Social Networks. ACM, 2009.

[26] Zheng, Yu, et al. ”Mining interesting locations and travel sequences from GPS trajectories.” Proceedings of the 18th international conference on World wide web. ACM, 2009.

[27] Yoon, Hyoseok, et al. ”Smart itinerary recommendation based on user-generated GPS trajectories.” Ubiquitous Intelligence and Computing. Springer Berlin Heidelberg, 2010. 19-34.

[28] Zheng, Vincent W., et al. ”Collaborative location and activity recommendations with gps history data.” Proceedings of the 19th international conference on World wide web. ACM, 2010.

[29] Zheng, Yu, and Xing Xie. ”Learning location correlation from gps trajectories.” Mobile Data Management (MDM), 2010 Eleventh International Conference on. IEEE, 2010.

[30] Lee, Ryong, and Kazutoshi Sumiya. ”Measuring geographical regularities of crowd behaviors for Twitter-based geo-social event detection.” Proceedings of the 2nd ACM SIGSPATIAL international workshop on location based social networks. ACM, 2010.

[31] Vasconcelos, Marisa Affonso, et al. ”Tips, dones and todos: uncovering user profiles in foursquare.” Proceedings of the fifth ACM international conference on Web search and data mining. ACM, 2012.

[32] Aggarwal, Anupama, Jussara Almeida, and Ponnurangam Kumaraguru. ”Detection of spam tipping behaviour on foursquare.” Proceedings of the 22nd international conference on World Wide Web companion. International World Wide Web Conferences Steering Committee, 2013.

[33] Sklar, Max, and Kristian J. Concepcion. ”Timely Tip Selection for Foursquare Recommendations.” RecSys 2014 Poster Proceedings

[34] Moraes, Felipe, et al. ”Polarity Detection of Foursquare Tips.” Social Informatics. Springer International Publishing, 2013. 153-162.

[35] Clifton, Chris, and Robert Cooley. ”TopCat: Data mining for topic identification in a text corpus.” Principles of Data Mining and Knowledge Discovery. Springer Berlin Heidelberg, 1999. 174-183.

[36] Muñoz-García, O., et al. ”Identifying topics in social media posts using dbpedia.” (2011).

[37] Nguyen, T., Hady Lauw, and Panayiotis Tsaparas. ”Review Selection Using Micro-Reviews.” (2014).

[38] Nguyen, Thanh-Son, Hady W. Lauw, and Panayiotis Tsaparas. ”Using micro-reviews to select an efficient set of reviews.” Proceedings of the 22nd ACM international conference on Conference on information & knowledge management. ACM, 2013.

[39] Nguyen, Thanh-Son, Hady W. Lauw, and Panayiotis Tsaparas. ”Review Synthesis for Micro-Review Summarization.”

[40] Μιχάλης Βαζιργιάννης, Μαρία Χαλκίδη, Εξόρυξη Γνώσης από Βάσεις Δεδομένων και τον Παγκόσμιο Ιστό, Εκδ. Gutenberg

[41] Forman, George, and Ira Cohen. ”Learning from little: Comparison of classifiers given little training.” Knowledge Discovery in Databases: PKDD 2004. Springer Berlin Heidelberg, 2004. 161-172.

[42] Foursquare API. https://developer.foursquare.com

[43] Natural Language ToolKit. http://www.nltk.org/

[44] JavaScript Object Notation. http://www.json.org/

[45] PyFoursquare. https://github.com/marcelcaraciolo/foursquare

[46] Foursquare library. https://pypi.python.org/pypi/foursquare

[47] Inflect python library. https://pypi.python.org/pypi/inflect

[48] Gensim topic modeling python library. http://radimrehurek.com/gensim/

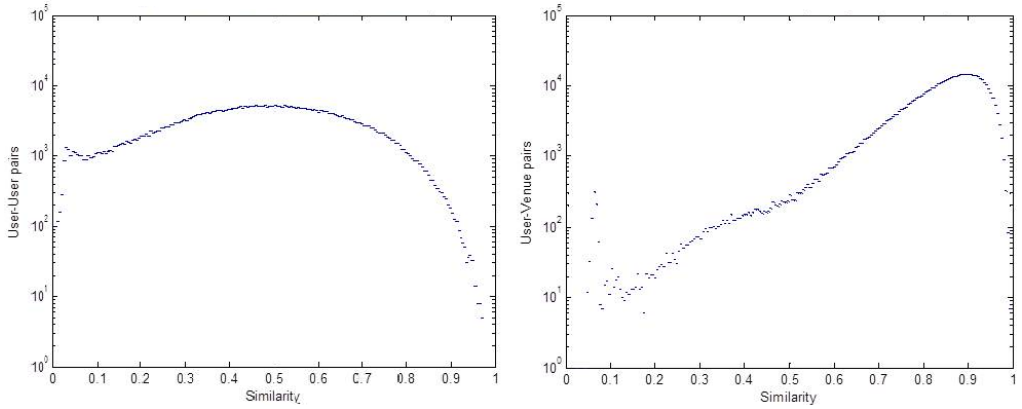
[49] Cosine Similarity http://en.wikipedia.org/wiki/Cosine_similarity

[50] Bag-of-words Model. http://en.wikipedia.org/wiki/Bag-of-words_model

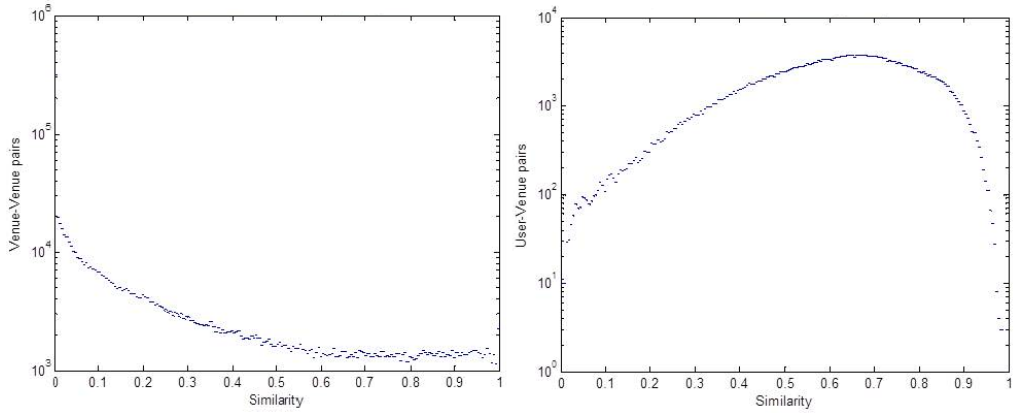
[51] Collaborative Filtering http://en.wikipedia.org/wiki/Collaborative_filtering

[52] Query and Ranking. http://en.wikipedia.org/wiki/Ranking_(information_retrieval)

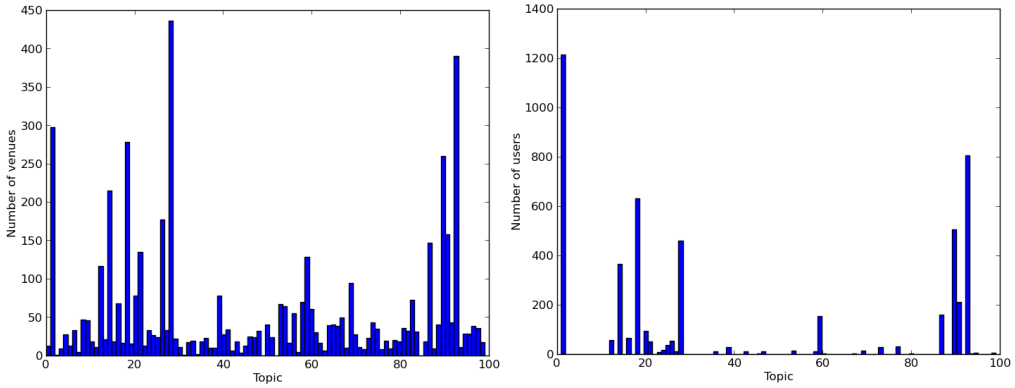
Παράρτημα



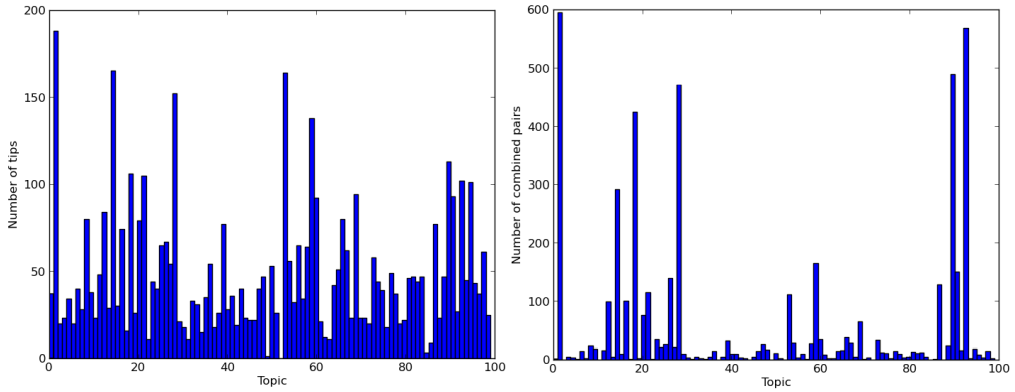
Σχήμα Π.1: Ιστόγραμμα ομοιότητας μεταξύ χρηστών (αριστερά) και μεταξύ χρήστη-τοποθεσίας (δεξιά) με χρήση μοντέλου με δεδομένα χρηστών (Νέα Υόρκη, Όλα τα σχόλια)



Σχήμα Π.2: Ιστόγραμμα ομοιότητας μεταξύ τοποθεσιών (αριστερά) και μεταξύ χρήστη-τοποθεσίας (δεξιά) με χρήση μοντέλου με δεδομένα τοποθεσιών (Νέα Υόρκη, Όλα τα σχόλια)



Σχήμα Π.3: Ιστόγραμμα με τα πιο πιθανά θέματα για τις τοποθεσίες (αριστερά) και τους χρήστες (δεξιά) με χρήση μοντέλου με δεδομένα τοποθεσιών (Νέα Υόρκη, Όλα τα σχόλια)



Σχήμα Π.4: Ιστόγραμμα με τα πιο πιθανά θέματα για τα σχόλια (αριστερά) και συνδυασμού γινομένου χρήστη-τοποθεσίας (δεξιά) με χρήση μοντέλου με δεδομένα τοποθεσιών (Νέα Υόρκη, Όλα τα σχόλια)

Σύντομο Βιογραφικό

Ο Ιωάννης Κοτρώτσιος γεννήθηκε το 1984 στη Θεσσαλονίκη, όπου και μεγάλωσε. Το 2011 απέκτησε Πτυχίο από το Τμήμα Πληροφορικής της Σχολής Θετικών Επιστημών του Πανεπιστημίου Ιωαννίνων. Από το 2012 παρακολουθεί το μεταπτυχιακό πρόγραμμα του τμήματος Μηχανικών Η/Υ και Πληροφορικής του Πανεπιστημίου Ιωαννίνων με κατεύθυνση το Λογισμικό. Επιπρόσθετα, από το 2013 είναι μέλος του Ερευνητικού Εργαστηρίου Κατανεμημένης Διαχείρισης και Επεξεργασίας Δεδομένων (D.M.O.D) του παραπάνω τμήματος.