

Opinion Mining and Sentiment analysis

Τμήμα Μηχανικών Η/Υ και Πληροφορικής
επιβλέπων καθηγητής: Μακρής Χρήστος



ΠΑΝΕΠΙΣΤΗΜΙΟ
ΠΑΤΡΩΝ
UNIVERSITY OF PATRAS

– Επισκόπηση και πειραματική αξιολόγηση τεχνικών
για opinion mining και sentiment analysis –

Παναγόπουλος Παναγιώτης

Sentiment Analysis ή Opinion Mining

- είναι ο ερευνητικός τομέας που αναφέρεται στην χρήση τεχνικών natural language processing, text analysis και computational linguistics ώστε να αναγνωριστούν και να εξαχθούν υποκειμενικές πληροφορίες από υλικό που έχει παραχθεί από ανθρώπινη δραστηριότητα.
- αναλύει τις ανθρώπινες απόψεις, τα συναισθήματα, τις αξιολογήσεις και τις στάσεις, προς οντότητες όπως προϊόντα, υπηρεσίες, οργανισμούς, πρόσωπα, εκδηλώσεις, θεματικές ενότητες, καθώς και προς τις ιδιότητες αυτών.
- αποσκοπεί στο να προσδιορίσει τη στάση ενός ομιλητή ή συγγραφέα σε σχέση με κάποιο θέμα ή την συνολική πολικότητα ενός κειμένου.

RoadMap

- Opinion Mining and Sentiment Analysis
 - εισαγωγή
 - εφαρμογές
 - ορισμός του προβλήματος
- Τεχνικές Opinion Mining and Sentiment Analysis
 - τα 3 επίπεδα ανάλυσης (κειμένου, πρότασης, χαρακτηριστικού)
 - sentiment λεξικό
 - opinion spamming
 - αναζήτηση και ανάκτηση απόψεων
 - σύνοψη απόψεων
- Opinion-based Κατάταξη Οντοτήτων
 - βελτιώνοντας την opinion-based κατάταξη οντοτήτων

Εισαγωγή στην εξόρυξη και ανάλυση απόψεων

- Με τον όρο Opinion Mining ή Sentiment Analysis συνήθως αναφερόμαστε μία διαδικασία άντλησης υποκειμενικών πληροφοριών υψηλής ποιότητας από κείμενικές πηγές.
- Για τις διαδικασίες του text mining συνήθως χρησιμοποιούνται τεχνικές ανάλυσης κειμένου όπως:
 - τεχνικές ανάκτησης πληροφορίας
 - λεξικής ανάλυσης
 - αναγνώρισης προτύπων
 - εξόρυξης δεδομένων
 - και τεχνικές natural language processing

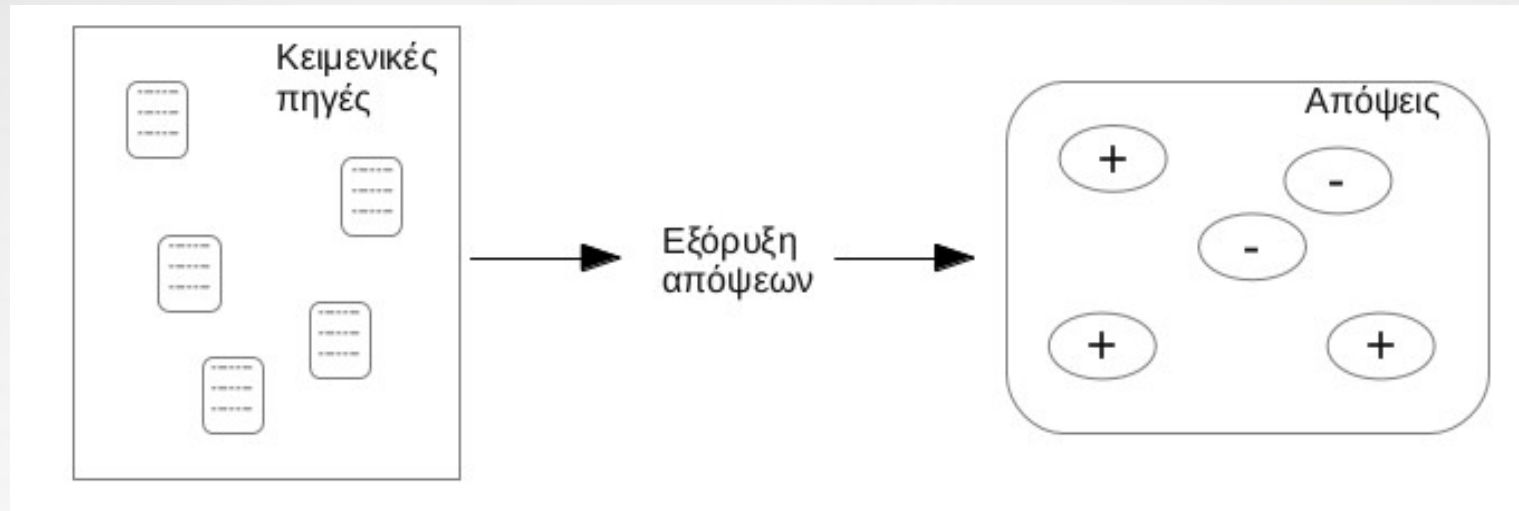
Κίνητρα

- Οι απόψεις αποτελούν σημαντικό κομμάτι της ανθρώπινης δραστηριότητας επειδή επηρεάζουν την συμπεριφορά μας.
- Κάθε φορά που χρειάζεται να πάρουμε μία απόφαση, θέλουμε να γνωρίσουμε τις απόψεις των άλλων.
 - οι εταιρείες και οργανισμοί θέλουν να έχουν εικόνα των απόψεων των καταναλωτών ή της κοινής γνώμης για τα προϊόντα ή τις υπηρεσίες τους
 - οι καταναλωτές θέλουν να γνωρίζουν τις απόψεις των άλλων χρηστών, πριν προχωρήσουν σε κάποια αγορά.
- Τα τελευταία χρόνια έχουμε δει πως η ανάρτηση υποκειμενικού περιεχομένου στα κοινωνικά, έχει οδηγήσει στην αναμόρφωση των μοντέλων των επιχειρήσεων και έχει επηρεάσει τα κοινωνικά και πολιτικά μας συστήματα.
- Από το 2000 ο τομέας του Opinion Mining έχει εξελιχθεί σε μία αρκετά ενεργή ερευνητική περιοχή
 - έχει ένα ευρύ φάσμα εφαρμογών
 - παρέχει πολλά ενδιαφέροντα ερευνητικά προβλήματα, τα οποία δεν έχουν μελετηθεί παλιότερα
 - για πρώτη φορά στην ανθρώπινη ιστορία, έχουμε διαθέσιμο ένα τεράστιο όγκο από υποκειμενικά δεδομένα στα κοινωνικά δίκτυα στο διαδίκτυο

Εφαρμογές της εξόρυξης και ανάλυσης απόψεων

- Κυρίως εφαρμογές κοινωνικού, πολιτικού και βιομηχανικού ενδιαφέροντος
 - πρόβλεψη της απόδοσης πωλήσεων
 - παραγωγή κατάταξης προϊόντων και προμηθευτών
 - συσχέτιση των κοινωνικών δικτύων με τις δημοσκοπήσεις απόψεων της κοινής γνώμης
 - πρόβλεψη των αποτελεσμάτων εκλογών
 - προβλεψη της εισπρακτικής επιτυχίας ταινιών
 - μελέτη στρατηγικών marketing
 - χαρακτηρισμός κοινωνικών σχέσεων

Ορισμός του προβλήματος



- οι εξαγώμενες υποκειμενικές πληροφορίες ονομάζονται απόψεις.
- μία άποψη αποτελείται από ένα στόχο και ένα συναίσθημα (g,s).
 - “ο καφές είναι καλός”
 - “η γεύση του καφέ είναι καλή”
- ο αναλυτικός ορισμός της γενικής άποψης είναι μία πεντάδα $(e_i, a_{ij}, s_{ijkl}, h_k, t_l)$
- υπάρχουν πολλές κατηγορίες απόψεων

Ορισμός του προβλήματος

- ο ορισμός της διαδικασίας της εξόρυξης απόψεων

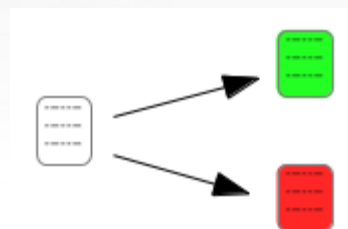
Δοθέντος ένος κειμένου d (ή μιας συλλογής κειμένων D), ανακάλυψε όλες τις πλειάδες απόψεων (e_i , a_{ij} , s_{ijkl} , h_k , tl) που υπάρχουν μέσα στο d . Η διαδικασία περιέχει 6 διεργασίες:

1. εξαγωγή όλων των οντοτήτων (entities) από το D και κατηγοριοποίηση των συνώνυμων σε ομάδες (entity clusters). Σε κάθε cluster δίνεται το αναγνωριστικό e_i
2. εξαγωγή όλων των χαρακτηριστικών (aspects) από το D και κατηγοριοποίηση των συνώνυμων σε ομάδες (aspects clusters). Σε κάθε cluster δίνεται το αναγνωριστικό a_{ij} του e_i
3. εξαγωγή των opinion holders και κατηγοριοποίηση σε ομάδες με αναγνωριστικό h_k
4. εξαγωγή του χρόνου και κανονικοποίηση σε ένα πρότυπο
5. προσδιορισμός της πολικότητας της άποψης (θετική/αρνητική/ουδέτερη), ή βαθμολόγηση (σε ποσοστό επί τις % ή 1-5 stars)
6. παραγωγή όλων των πλειάδων απόψεων (e_i , a_{ij} , s_{ijkl} , h_k , tl)

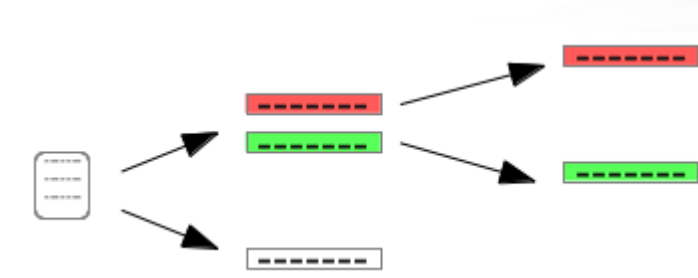
Τεχνικές εξόρυξης και ανάλυσης απόψεων

- Η εξόρυξη και ανάλυση απόψεων έχει ερευνηθεί κυρίως σε τρία επίπεδα

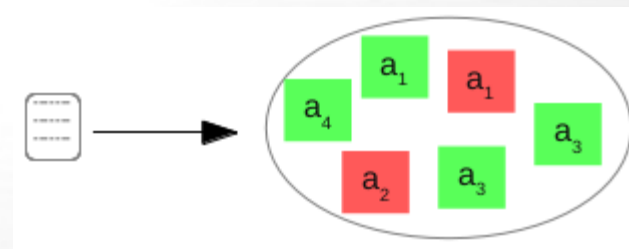
- το επίπεδο κειμένου



- το επίπεδο πρότασης



- το επίπεδο οντότητας ή χαρακτηριστικού



Παράγοντες που καθιστούν δύσκολη την εξόρυξη και ανάλυση απόψεων

- οι δυσκολίες απορρέουν από το γεγονός ότι απαιτείται από τους αλγορίθμους να είναι σε θέση να κατανοούν πλήρως την σημασιολογική πληροφορία των κειμένων
- οι εξαγωγή των οντοτήτων, των aspects, των opinion holders, και του χρόνου στον οποίο εκφράζονται οι απόψεις, μπορούμε να πούμε πως είναι κλασικά προβλήματα Named-entity recognition (NER). Γενικά δεν υπάρχει ικανοποιητική λύση.
- βασικό πρόβλημα είναι η χρήση συνώνυμων λέξεων, από τους χρήστες, οι οποίοι χρησιμοποιούν διαφορετικά ονόματα για να αναφερθούν στις ίδιες οντότητες και aspects.
- η ύπαρξη sentiment λέξεων σε μία πρόταση, συνήθως υποδουλώνει την ύπαρξη κάποιας άποψης. Όμως δεν συμβαίνει πάντα αυτό
- το πλήθος των διαφορετικών μορφών με τους οποίους μπορεί να εκφραστεί μία άποψη καθιστά την αναγνώριση της υποκειμενικότητας μία δύσκολη διαδικασία

Παράγοντες που καθιστούν δύσκολη την εξόρυξη και ανάλυση απόψεων

- ο προσδιορισμός της πολικότητας των απόψεων, πρέπει να πραγματοποιηθεί με προσοχή, διότι πρέπει να ληφθούν υπόψιν σχήματα άρνησης, σαρκασμού, σχήματα σύγκρισής και άλλα.
- γενικά η ποικιλομορφία της ανθρώπινης έκφρασης αποτελεί έναν από τους βασικότερους παράγοντες οι οποίοι καθιστούν δύσκολη μία διαδικασία ανάλυσης απόψεων
- οι πολικότητες των απόψεων σχετικά με συγκεκριμένα aspects εξαρτώνται από το domain στο οποίο εργαζόμαστε
- επίσης είναι απαραίτητος ο προσδιορισμός της ποιότητας των απόψεων
- επίσης παράγοντες που καθιστούν δύσκολη την διαδικασία της εξόρυξης και ανάλυσης απόψεων, και σχετίζονται με την ποιότητα των απόψεων, είναι η διαφήμιση, και τα φαινόμενα του opinion spamming και sock puppeting

Τεχνικές εξόρυξης και ανάλυσης απόψεων

- Στο επίπεδο ανάλυσης κειμένου το πρόβλημα της εξαγωγής και επεξεργασίας απόψεων παρουσιάζει δύο μορφές, ανάλογα με τη είδος του πεδίου τιμών της επιθυμητής εξόδου:
 - αν το πεδίο τιμών αποτελείται από διακριτές (κατηγορηματικές) τιμές, όπως θετικό ή αρνητικό, τότε είναι ένα πρόβλημα κατηγοριοποίησης
 - αν το πεδίο τιμών αποτελείται από αριθμητικές τιμές ή βαθμολογίες κατάταξης σε ένα προκαθορισμένο εύρος, όπως για παράδειγμα 1-5 stars, τότε είναι ένα regression πρόβλημα

Ανάλυση σε επίπεδο κειμένου

- Εποπτευόμενες τεχνικές sentiment κατηγοριοποίησης
 - μπορεί να χρησιμοποιηθεί οποιαδήποτε εποπτευόμενη μέθοδος, όπως η κατηγοριοποίηση naïve Bayes και η τεχνική support vector machines
 - το κλειδί για την κατηγοριοποίηση υποκειμενικής άποψης είναι η σύσταση ενός συνόλου αποτελεσματικών features
 - unigrams, n-grams
 - η Delta TFIDF ως μία νέα τεχνική απόδοσης βάρους στις λέξεις των κειμένων
 - η χρήση παρουσίας (presence) γλωσσολογικών features
 - “recency effect”, αρκεί να εξετάζονται μόνο οι τελευταίες προτάσεις αντί για όλο το μήκος του κειμένου
- Μη-εποπτευόμενες τεχνικές sentiment κατηγοριοποίησης
 - απλούστερες μέθοδοι οι οποίες εντοπίζουν στο κείμενο τις υποκειμενικές πληροφορίες και αποδίδουν σε αυτό, με κάποια τεχνική, μία τελική βαθμολογία, κατηγοριοποιώντας το ως θετικό ή αρνητικό
 - τεχνικές που βασίζονται στην χρήση συντακτικών patterns ή sentiment λεξικού

Ανάλυση σε επίπεδο κειμένου

- Πρόβλεψη sentiment βαθμολογιών
 - μπορεί να χρησιμοποιηθεί κάποια τεχνική regression, αν και υπάρχουν έρευνες οι οποίες ακολουθούν και άλλες προσεγγίσεις
 - PRank μοντέλο
 - n ανεξάρτητα/εξαρτημένα regression μοντέλα, ένα για κάθε aspect
 - αναπαράσταση των κειμένων ως bag-of-opinions, αντί για την αναπαράσταση bag-of-words, με βάση το γεγονός πως οι απόψεις εμφανίζονται συχνά ως n-grams patterns

Cross-domain sentiment κατηγοριοποίηση

- Η κατηγοριοποίηση υποκειμενικού περιεχομένου είναι αρκετά ευαίσθητη στην θεματική ενότητα που αυτή πραγματοποιείται
- τεχνικές για την αντιμετώπιση αυτού του φαινομένου βασίζονται κυρίως στην δημιουργία:
 - κατηγοριοποιητών χρησιμοποιώντας δεδομένα και από τα δύο domains
 - εκπαιδεύονται κατηγοριοποιητές με σύνολα κειμένων και από τα δύο domains (target και source)
 - επιλέγεται ένα feature space με χαρακτηριστικά από δύο domains
 - χρησιμοποιούνται semi-supervised τεχνικές
 - domain independent κατηγοριοποιητών
 - εκπαίδευση του κατηγοριοποιητή σε ένα σύνολο από reviews τα οποία προέρχονται από πολλά διαφορετικά domains
 - επιλέγεται ένα domain independent feature space με χαρακτηριστικά από πολλά domains

Ποιότητα των review κειμένων

- Ο προσδιορισμός της ποιότητας ενός review κειμένου είναι σημαντικός για τα συστήματα που διαχειρίζονται reviews χρηστών σχετικά με προϊόντα ή υπηρεσίες
 - όταν ο αριθμός των reviews σε ένα ιστότοπο είναι μεγάλος, είναι σημαντικό για τον αναγνώστη να εντοπίζει εύκολα και γρήγορα χρήσιμα reviews
 - “Was the review helpful to you?”
 - σε online συστήματα
 - είναι ευθύνη των χρηστών να παρέχουν feedback
- προσεγγίζεται συνήθως ως ένα regression πρόβλημα
 - το μοντέλο το οποίο σχηματίζεται, καταχωρεί μία βαθμολογία με βάση την ποιότητα σε κάθε review
 - τα δεδομένα εκπαίδευσης απαρτίζονται από τα το feedback των χρηστών για τα review
 - τα σύνολα χαρακτηριστικών μπορούν να είναι διάφορων ειδών:
 - χαρακτηριστικά δομής, unigrams και bigrams με TF_IDF βάρη, συντακτικά χαρακτηριστικά, σημασιολογικά χαρακτηριστικά, meta-data

Τεχνικές εξόρυξης και ανάλυσης απόψεων

- στο επίπεδο ανάλυσης πρότασης στόχος είναι η κατηγοριοποίηση των προτάσεων σε θετικής ή αρνητικής πολικότητας, ανάλογα με το αν εκφράζει θετικό ή αρνητικό συναίσθημα
 - κατηγοριοποίηση υποκειμενικότητας
 - κατηγοριοποίηση πολικότητας
- η sentiment κατηγοριοποίηση μιας πρότασης μπορεί να αντιμετωπιστεί
 - είτε ως ένα πρόβλημα κατηγοριοποίησης τριών κατηγοριών (θετική, αρνητική, ουδέτερη),
 - είτε ως δύο ξεχωριστά προβλήματα κατηγοριοποίησης
 - κατηγοριοποίηση υποκειμενικότητας
 - κατηγοριοποίηση πολικότητας

Ανάλυση σε επίπεδο πρότασης

- κατηγοριοποίηση υποκειμενικότητας
 - προσεγγίσεις βασισμένες στην εποπτευόμενη μάθηση
 - σύνολο χαρακτηριστικών
 - unigrams, bigrams, trigrams
 - μέρη του λόγου
 - η παρουσία sentiment λέξεων
 - η αρθροιστική πολικότητα ακολουθιών με sentiment λέξεις
 - μέρη του λόγου μαζί με την πολικότητά τους
- κατηγοριοποίηση πολικότητας
 - εποπτευόμενη μάθηση
 - lexicon-based μέθοδοι

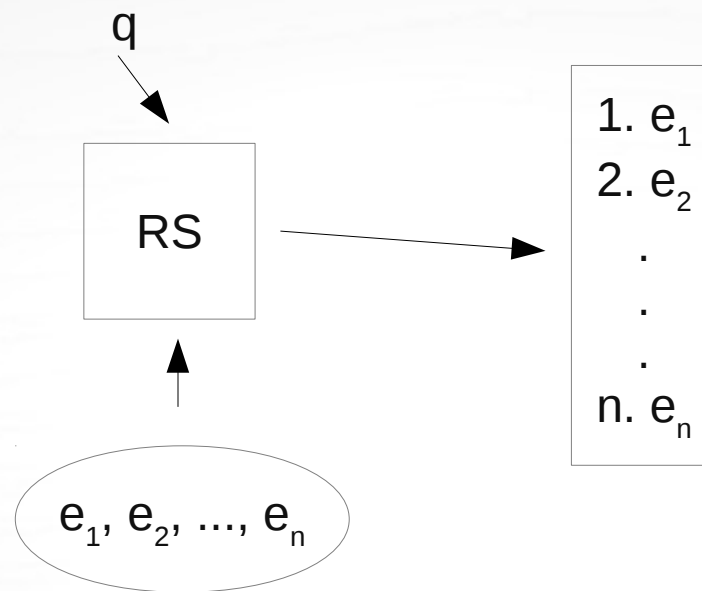
Ανάλυση σε επίπεδο χαρακτηριστικού

- εξαγωγή των οντοτήτων, των opinion holders, και του χρόνου
 - μεθοδοι βασισμένοι σε κανόνες
 - στατιστικές μέθοδοι
- εξαγωγή και ομαδοποίηση των aspects
 - ευκολότερη διαδικασία – μία άποψη έχει πάντα ένα στόχο
 - εξαγωγή χρησιμοποιώντας
 - την συχνότητα των ουσιαστικών και των ονοματικών εκφράσεων
 - την σχέση μεταξύ απόψης και του στόχου της άποψης
 - εποπτευόμενη μάθηση
 - θεματικά μοντέλα
- κατηγοριοποίηση υποκειμενικότητας
 - προσδιορισμός της πολικότητας των απόψεων για τα aspects
 - εποπτευόμενη μάθηση
 - βασισμένη σε λεξικό

Αναζήτηση και ανάκτηση απόψεων

- αναζήτηση
 - απόψεων σχετικά με μία συγκεκριμένη οντότητα ή κάποιο χαρακτηριστικό μιας οντότητας
 - απόψεων συγκεκριμένων ατόμων ή οργανισμών σχετικά με κάποια οντότητα ή κάποιο χαρακτηριστικό μιας οντότητας
- Ένα σύστημα ανάκτησης απόψεων αποτελείται από δύο μέρη:
 - retrieval component
 - opinion classification component

Opinion-based κατάταξη οντοτήτων



Το πρόβλημα της αξιολόγησης οντοτήτων ως πρόβλημα ανάκτησης πληροφορίας

- Η ιδέα για την διάταξη των οντοτήτων, είναι να αναπαρασταθεί κάθε οντότητα με το κείμενο όλων των διαθέσιμων review κειμένων τα οποία αναφέρονται για αυτήν, (Ganesan and ChengXiang, 2012)
- δοθέντος ενός ερωτήματος του χρήστη
 - το οποίο εκφράζει τα επιθυμητά aspects, τα οποία μία οντότητα πρέπει να διαθέτει
 - αξιολογούμε τις οντότητες με βάση το πόσο καλά τα review κείμενα ταιριάζουν τις προτιμήσεις του χρήστη
- το πρόβλημα της αξιολόγησης οντοτήτων μετατρέπεται σε πρόβλημα ανάκτησης πληροφορίας
 - μπορούμε να χρησιμοποιήσουμε οποιοδήποτε γνωστό μοντέλο ανάκτησης πληροφορίας, όπως το BM25

Βελτιώνοντας την opinion-based κατάταξη των οντοτήτων

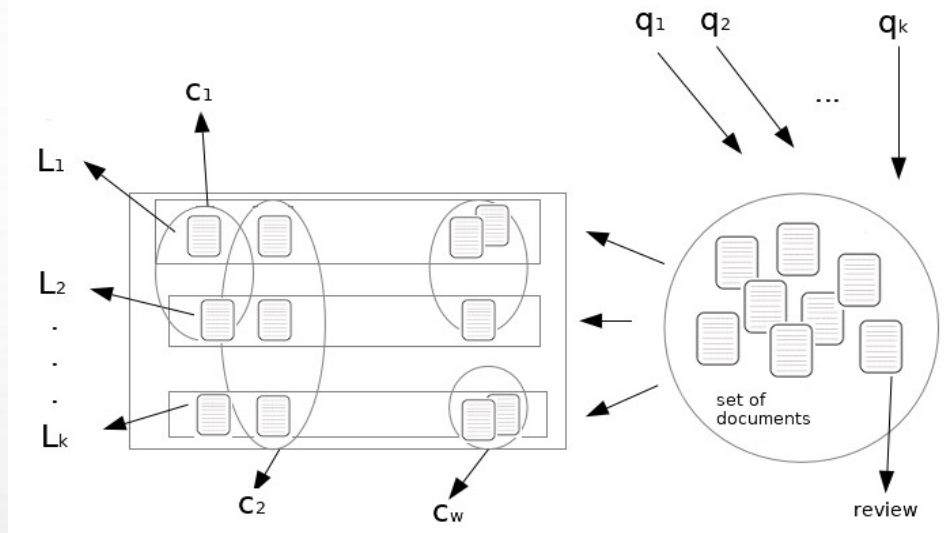
- Το setup που παρουσιάζεται στην έρευνα Opinion-based Entity Ranking (Ganesan and ChengXiang, 2012) δεν επιβάλλει κυρώσεις για τις αρνητικές απόψεις
- Επιχειρούμε να βελτιώσουμε ακόμα περισσότερο αυτή την προσέγγιση
 - αντί της χρήσης της σημαντικότητας των aspect keywords του ερωτήματος για ένα review κείμενο, δοκιμάζουμε να χρησιμοποιήσουμε την sentiment πληροφορία των απόψεων, οι οποίες εκφράζονται στο review κείμενο
 - επιχειρούμε να αναγνωρίσουμε παρόμοιες συμπεριφορές βαθμολόγησης ανάμεσα στους χρήστες, και να χρησιμοποιήσουμε αυτήν την πληροφορία για την βελτίωση της κατάταξης
 - προτείνουμε το naïve consumer μοντέλο ως ένα μη εποπτευόμενο σχήμα, το οποίο χρησιμοποιεί πληροφορίες από το διαδίκτυο ώστε να αποδόσει ένα βάρος σημαντικότητας σε κάθε ένα από τα aspects τα οποία χρησιμοποιούνται στην αξιολόγηση των οντοτήτων

Opinion-based βαθμολογίες χαρακτηριστικών

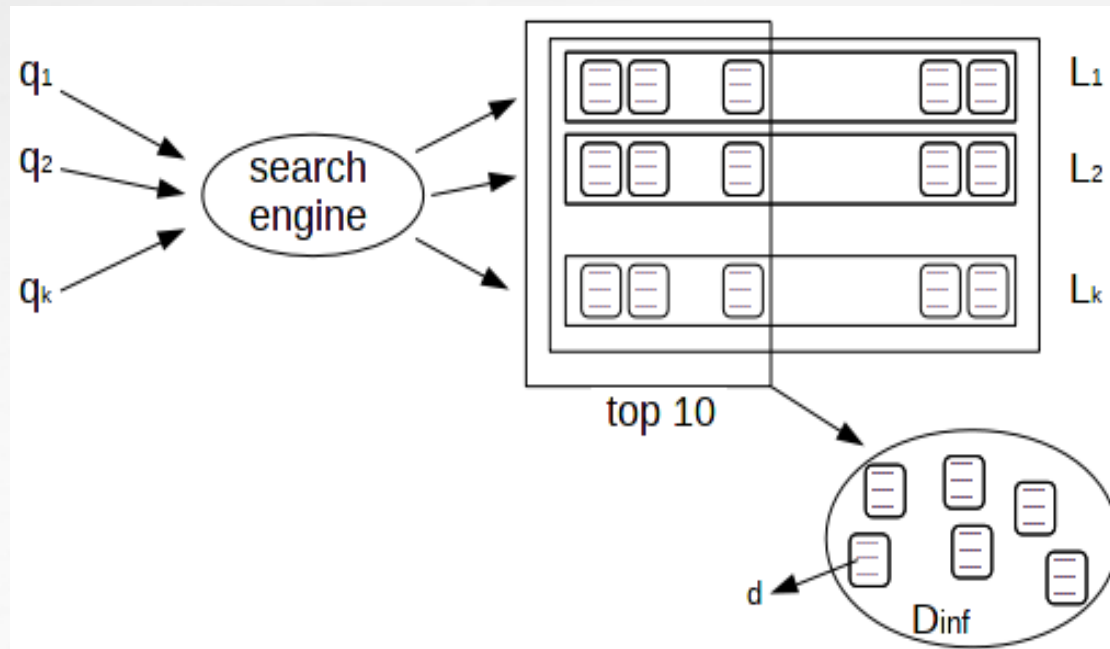
- Για να δημιουργήσουμε ένα μοντέλο το οποίο λαμβάνει υπόψιν του την sentiment πληροφορία των απόψεων οι οποίες εκφράζονται στα review κείμενα χρησιμοποιούμε δύο μη εποπτευόμενες τεχνικές sentiment ανάλυσης
 - lexicon-based
 - syntactic patterns-based
- Με βάση το γεγονός πως κάθε review περιέχει απόψεις χρηστών για μία συγκεκριμένη οντότητα, εφαρμόζουμε
 - aspect-based τεχνικές sentiment ανάλυσης για να εξάγουμε την sentiment πληροφορία η οποία εκφράζεται για τα aspects στις προτάσεις
 - και aspect-based summarization (ή feature-based summarization) για να υπολογίσουμε την συνολική βαθμολογία των aspects σε κάθε κείμενο

Εξομάλυνση της κατάταξης με Opinion-based Clusters

- προσπαθούμε να χρησιμοποιήσουμε clustering πληροφορία σχετικά με τα reviews ώστε να βελτιώσουμε την κατάταξη των οντοτήτων
- χρησιμοποιούμε τον αλγόριθμο ClustFuse του Kurland (Kurland, 2006), ο οποίος χρησιμοποιεί δύο συνιστώσες για να αποδόσει μία βαθμολογία σε ένα κείμενο
 - την πιθανότητα σχετικότητας του κειμένου με το ερώτημα
 - την υπόθεση πως οι συστάδες μπορούν να χρησιμοποιηθούν ως proxies για τα κείμενα



To Naive Consumer μοντέλο



$$scoreA(a_i) = \frac{\sum_{d \in D_{inf}} BM25(a_i)_d}{\sum_{a' \in A} \left(\sum_{d \in D_{inf}} BM25(a')_d \right)}$$

$$p(q, ri) = \sum_{t \in q} score(t)_{ri} * scoreA(t)$$

Πειράματα και Αποτελέσματα

method	1 st exp. set	2 nd exp. set
BM25std	0.87	0.936
BM25std+AvgScoreQAM+ opinExp	0.88	0.955
lexicon-based SA	0.865	0.91
syntactic patterns-based SA	0.869	0.94
BM25std+kurland	0.88	0.90
BM25std+kurland+opinion- based clusters	0.89	0.956
NCM1	0.891	0.938
NCM2	0.893	0.96
m-aspect PRank	-	0.95