OPINION MINING USING TEXT DATA

Εργασία στην Ανάκτηση Πληροφορίας
Χειμερινό Εξάμηνο 2016-2017

1. Περιγραφή

Στην εργασία αυτή, θα μελετήσετε τεχνικές εξόρυξης γνώσης από δεδομένα κειμένου (text mining). Πιο συγκεκριμένα, θα αντιμετωπίσετε το πρόβλημα που είναι γνωστό ως **Opinion Mining** ή **Sentiment Analysis**. Σκοπός των τεχνικών αυτών είναι ο προσδιορισμός των συναισθημάτων ή απόψεων που κρύβονται σε κείμενα.

Το πρόβλημα που θα μελετήσετε είναι αρκετά απλό και σχετίζεται με την ανάλυση κριτικών ταινιών. Αναλυτικότερα, δίνεται ένα σύνολο από κριτικές οι οποίες χαρακτηρίζονται είτε ως **θετικές** είτε ως **αρνητικές**. Για παράδειγμα, η επόμενη κριτική χαρακτηρίζεται ως θετική, διότι αποτυπώνει μία θετική εικόνα για τη συγκεκριμένη ταινία:

"This movie is really not all that bad. But then again, this movie genre is right down my alley. Sure, the sets are cheap, but they really did decent with what they had. If you like cheap, futuristic, post-apocalyptic B movies, then you'll love this one!! I sure did!"

Ο στόχος είναι να προσδιορίσουμε αν μία “νέα” κριτική είναι θετική ή αρνητική, βασισμένοι σε ένα σύνολο κριτικών για τις οποίες γνωρίζουμε αν είναι θετικές ή αρνητικές. Το πρόβλημα αυτό μπορεί να αντιμετωπιστεί τόσο με supervised τεχνικές όσο και με unsupervised αλγορίθμους. Οι supervised τεχνικές συνήθως αντιμετωπίζουν το συγκεκριμένο πρόβλημα ως **πρόβλημα classification** και με βάση τα δεδομένα εκπαίδευσης προσπαθούν να προσδιορίσουν ένα μοντέλο κατάλληλο να δίνει σωστά αποτελέσματα. Από την άλλη πλευρά, οι unsupervised τεχνικές αντιμετωπίζουν διαφορετικά το πρόβλημα προσπαθώντας να προσδιορίσουν αποδοτικούς αλγορίθμους **χωρίς εκπαίδευση** που να δίνουν λύση στο πρόβλημα. Χαρακτηριστικό παράδειγμα είναι η χρήση **k-nearest-neighbor** τεχνικών για τον προσδιορισμό της κλάσης ενός αντικειμένου. Έχετε τη δυνατότητα να χρησιμοποιήσετε και supervised και unsupervised μεθόδους. Υπάρχει βοηθητικό υλικό που καλύπτει ένα μεγάλο φάσμα των τεχνικών που χρησιμοποιούνται στην πράξη.

Το πρόβλημα που θα μελετήσετε είναι στενά συνδεδεμένο με τεχνικές ανάκτησης πληροφορίας. Θα πρέπει να προσδιορίσετε τα κατάλληλα χαρακτηριστικά (features) από τα κείμενα ώστε να μπορέσετε να συσχετίσετε το περιεχόμενο των κειμένων με την κατηγορία στην οποία ανήκουν (positive/negative). Στο μάθημα έχουμε μελετήσει τέτοιου είδους τεχνικές, οι οποίες στηρίζονται σε διανύσματα βαρών που εκφράζουν τη σημαντικότητα ενός όρου σε ένα έγγραφο κειμένου (π.χ., με χρήση tf-idf).

Επίσης, θα πρέπει να εφαρμόσετε κατάλληλες τεχνικές προεπεξεργασίας των κειμένων, μετασχηματισμού των όρων αν απαιτείται και ίσως και απαλοιφή των συχνά εμφανιζόμενων λέξεων (stopwords). Ωστόσο, απαιτείται προσοχή διότι πολλές φορές **τα επίθετα και τα ρήματα** δίνουν πολύ σημαντική πληροφορία και επομένως η απαλοιφή τους μπορεί να μειώσει την ακρίβεια του αποτελέσματος. Επιπλέον, υπάρχουν τεχνικές όπως η χρήση *n*-grams, όπου λαμβάνονται οι *n* συνεχόμενες λέξεις ενός κειμένου και χρησιμοποιούνται μαζί ως χαρακτηριστικά. Επίσης, η χρήση stemming πολλές φορές βοηθάει.

Θα πρέπει να στοχεύσετε στα εξής:

1. η λύση που θα δώσετε θα πρέπει να έχει μεγάλη ακρίβεια (όσο μεγαλύτερη μπορείτε να πετύχετε).
2. ο χρόνος εκτέλεσης θα πρέπει να είναι όσο μικρότερος γίνεται.

Είναι προφανές ότι τα δύο αυτά χαρακτηριστικά είναι πολλές φορές αντικρουόμενα και επομένως μπορεί να μην υπάρχει μία τεχνική που να πετυχαίνει και τα δύο στο βέλτιστο βαθμό. Επίσης, τονίζεται ότι σε περίπτωση που μελετήσετε κάποια supervised τεχνική, θα πρέπει να μετρηθεί ξεχωριστά ο χρόνος training από το χρόνο testing.

2. Παραδοτέα - Υποβολή

Στόχος της εργασίας είναι **να ασχοληθείτε όσο γίνεται περισσότερο με το πρόβλημα και να εξετάσετε διαφορετικές τεχνικές**. Τονίζεται επομένως ότι θα εκτιμηθεί περισσότερο μία προσπάθεια μελέτης διαφορετικών τεχνικών ακόμη και αν δε δίνει τα αναμενόμενα αποτελέσματα ακρίβειας, από μία εργασία που εξετάζει μία μόνο τεχνική που έχει πολύ καλή ακρίβεια.

Η εργασία λαμβάνει το **40%** του συνολικού βαθμού. Για την εκπόνηση της εργασίας θα πρέπει να εργαστείτε σε **ομάδες 2 ή 3 ατόμων**. **Εργασίες που θα εκπονηθούν ατομικά δε θα γίνουν δεκτές**. Φροντίστε ώστε η συνεισφορά του κάθε μέλους της ομάδας να είναι ουσιαστική. Η προθεσμία παράδοσης ορίζεται ως η ημερομηνία εξέτασης του μαθήματος στην εξεταστική του Ιανουαρίου-Φεβρουαρίου 2016. Θα πρέπει να υποβληθούν τα ακόλουθα:

- πηγαίος κώδικας (JAVA ή C++)
- τεχνική αναφορά στην οποία να περιγράφονται οι τεχνικές που χρησιμοποιήσατε καθώς και συγκριτικά αποτελέσματα μεταξύ διαφορετικών προσεγγίσεων
- ένα αρχείο με όνομα **predictions.txt** το οποίο περιέχει τις προβλέψεις σας για τα **test data** και πρόκειται για αρχείο **txt** που αποτελείται από δύο στήλες: η πρώτη στήλη περιέχει το review identifier (το όνομα του αρχείου) και η δεύτερη στήλη είναι είτε 0 (negative) είτε 1 (positive). Παράδειγμα:

```
00000 0
00001 0
00002 1
00003 0
00004 1
00005 0
00006 1
00007 0
00008 1
00009 0
```

Παρακαλώ, χρησιμοποιήστε τα δεδομένα που περιέχονται στο data.zip που υπάρχει στο elarning. Το αρχείο αυτό περιέχει δύο καταλόγους train και test. Ο κατάλογος train περιέχει δύο υποκαταλόγους, pos και neg, που περιέχουν positive/negative reviews. Το όνομα των αρχείων είναι της μορφής reviewID_rating.txt. Ο κατάλογος test περιέχει μόνο τα αρχεία των reviews και προφανώς δεν υπάρχει η πληροφορία αν είναι θετικές ή αρνητικές.

Οι εργασίες θα υποβληθούν μέσω του **elearning** χρησιμοποιώντας κατάλληλο σύνδεσμο ο οποίος θα ενεργοποιηθεί.

Για απορίες και παρατηρήσεις papadopo@csd.auth.gr