Τεχνικές Εξόρυξης Δεδομένων: Χειμερινό Εξάμηνο 2011

Πρώτη Άσκηση (ομάδες έως 2 ατόμων)

Ημερομηνία Παράδοσης: 8/1/2011

Σε αυτή την άσκηση **θα μελετήσετε τρόπους προεπεξεργασίας των δεδομένων, αλγορίθμους κατηγοριοποίησης δεδομένων και θα δοκιμάσετε να υλοποιήσετε τον δικό σας απλό αλγόριθμο ανάλυσης συναισθήματος**.

1) Πιο συγκεκριμένα στην άσκηση αυτή θα χρησιμοποιήσετε τους αλγορίθμους Naïve Bayes και Δέντρο Αποφάσεων που είναι υλοποιημένοι στο weka. Οι αλγόριθμοι αυτοί θα εφαρμοστούν στο σύνολο δεδομένων που είναι αναρτημένο στη σελίδα: http://cgi.di.uoa.gr/~ys11/
   Το συγκεκριμένο σύνολο δεδομένων περιλαμβάνει 2000 reviews ταινιών. Τα reviews αυτά ανήκουν σε δύο κατηγορίες: τα 1000 από αυτά είναι χαρακτηρισμένα ως «θετικά» και τα άλλα 1000 είναι χαρακτηρισμένα ως «αρνητικά». Κάθε review είναι ένα σύνολο από λέξεις, επομένως χρησιμοποιώντας το μοντέλο αναπαράστασης δεδομένων bag-of-words κάθε review μπορεί να αναπαρασταθεί ως ένα vector. Το εργαλείο weka προσφέρει την δυνατότητα αυτόματης δημιουργίας αυτών των vectors.

   Παράδειγμα:

   Για τα reviews:

   review1="a great movie"- positive, review2="excellent film" - positive

   review3="worst film ever" - negative, review4="a bad movie" - negative

   Τα διανύσματα που θα δημιουργηθούν είναι:

	a	bad	ever	excellent	film	great	movie	worst	Class
V1	1	0	0	0	0	1	1	0	+
V2	0	0	0	1	1	0	0	0	+
V3	0	0	1	0	1	0	0	1	-
V4	1	1	0	0	0	0	1	0	-

Οι αλγόριθμοι κατηγοριοποίησης επωφελούνται σε μεγάλο βαθμό αν από το dataset έχουν αφαιρεθεί λέξεις που δεν είναι χρήσιμες για την ανάλυση. Τέτοιες λέξεις είναι και τα stop words, π.χ. and, the, a, this, I, you, etc. Το εργαλείο weka προσφέρει την δυνατότητα αφαίρεσης των stop words πριν τη δημιουργία των vectors που θα χρησιμοποιηθούν για την εκπαίδευση και την εκτέλεση των αλγορίθμων κατηγοριοποίησης.

2) Στην ιστοσελίδα http://cgi.di.uoa.gr/~ys11/ θα βρείτε αναρτημένα δύο διαφορετικά λεξικά θετικών και αρνητικών λέξεων. Επιλέξτε όποια γλώσσα προγραμματισμού επιθυμείτε και υλοποιήστε έναν απλό αλγόριθμο χαρακτηρισμού ενός review ως θετικού ή αρνητικού. Για να βοηθηθείτε, τα αποτελέσματά σας μπορούν να έχουν τη μορφή:

Review1: 40% positive, 20% negative, 40% neutral

Review2: 32% positive, 28% negative, 40% neutral

Review3: 30% positive, 25% negative, 45% neutral

Review4: 10% positive, 10% negative, 80% neutral

Εκτελέστε τον αλγόριθμό σας στο σύνολο των reviews που σας δίνετε, παρουσιάστε, μελετήστε και σχολιάστε τα αποτελέσματά σας. Ποια λεξικά δουλεύουν καλύτερα; Επηρεάζουν την ποιότητα των αποτελεσμάτων σας; (Θυμηθείτε ότι για κάθε review σας δίνεται και ένας χαρακτηρισμός ως positive ή negative. Αυτόν τον χαρακτηρισμό δεν πρέπει να τον χρησιμοποιήσετε στον αλγόριθμό σας (εξάλλου αυτό ακριβώς προσπαθούμε να «μαντέψουμε») αλλά μπορείτε να τον χρησιμοποιήσετε για την αξιολόγηση των αποτελεσμάτων σας).

Δεν θα βαθμολογηθείτε για την αποτελεσματικότητα του αλγορίθμου, αλλά για την ανάλυση των αποτελεσμάτων σας.

Οδηγίες χρήσης του Weka:

a. Κατεβάστε και εγκαταστήστε το weka από την ιστοσελίδα: http://www.cs.waikato.ac.nz/ml/weka/ (Stable book 3rd ed. Version)
b. Εκτελέστε το weka και στη συνέχεια επιλέξτε “Explorer”. Για περισσότερες λεπτομέρειες σχετικά με το weka υπάρχουν στην παραπάνω ιστοσελίδα αρκετές καλές πληροφορίες σχετικά με τον τρόπο λειτουργίας του.
c. Επιλέξτε Open file… και φορτώστε το αρχείο με το σύνολο δεδομένων.
d. Επιλέξτε την καρτέλα Classify και στη συνέχεια τον ταξινομητή που επιθυμείτε.
e. Στην επιλογή test options επιλέγετε:
    i. Cross-validation και εισάγετε τον αριθμό των folds που επιθυμείτε. Περισσότερες πληροφορίες σχετικά με το k-fold cross-validation μπορείτε να βρείτε εδώ:

http://en.wikipedia.org/wiki/Cross-validation_(statistics)#K-fold_cross-validation

ii. Percentage-split και εισάγετε το ποσοστό του dataset που επιθυμείτε να χρησιμοποιήσετε ως training set (το υπόλοιπο θα χρησιμοποιηθεί ως test set)

f. Πατήστε Start και θα δείτε τα αποτελέσματα

Στον παρακάτω πίνακα ακολουθεί η αναλυτική βαθμολογία ανά ερώτηση:

Ερωτήματα	**Βαθμολογία**
Προεπεξεργασία Δεδομένων	
Α) Δημιουργία του αρχείου δεδομένων .arff από το σύνολο δεδομένων που παρέχεται σε μορφή .txt	12.5%
Β) Προεπεξεργασία δεδομένων: Αφαίρεση των stop words από το σύνολο δεδομένων	12.5%
Κατηγοριοποίηση Δεδομένων	
Α) Εκτέλεση του αλγορίθμου κατηγοριοποίησης Decision Trees και παρουσίαση των αποτελεσμάτων.	12.5%
Β) Εκτέλεση του αλγορίθμου κατηγοριοποίησης Naïve Bayes και παρουσίαση των αποτελεσμάτων.	12.5%
Γ) Αξιολόγηση και σύγκριση των αποτελεσμάτων που δίνουν οι δύο διαφορετικοί αλγόριθμοι κατηγοριοποίησης (με και χωρίς αφαίρεση των stop words)	12.5%
Δ) Αξιολόγηση και σύγκριση των αποτελεσμάτων που δίνουν οι δύο διαφορετικοί αλγόριθμοι κατηγοριοποίησης ανάλογα με τις	12.5%

παραμέτρους που καθορίζουν το μέγεθος του training και του test set. (Π.χ. Πώς επηρεάζει το μέγεθος του training set την ποιότητα της κατηγοριοποίησης των δεδομένων του test set;)	
Αλγόριθμος Ανάλυσης Συναισθήματος Α) Υλοποίηση αλγορίθμου (Μπορείτε να χρησιμοποιήσετε όποια γλώσσα προγραμματισμού επιθυμείτε. Κώδικας που δεν τρέχει δεν θα βαθμολογηθεί καθόλου) Β) Μελέτη και σχολιασμός των αποτελεσμάτων ανάλυσης συναισθήματος. Γ) Bonus: Υλοποίηση αλγορίθμου αξιολόγησης αποτελεσματικότητας αλγορίθμου (tip: χρησιμοποιήσετε τα labels - positive και negative – που έχετε για κάθε review).	12.5% 12.5% 12.5%
Σύνολο	112.5%

E-mail επικοινωνίας: Δημήτρης Κωτσάκος (dimkots [at] di.uoa.gr).

Σε μια από τις επόμενες διαλέξεις θα γίνει παρουσίαση της άσκησης και επίλυση απορίων.