Γλωσσική Τεχνολογία
Ακαδημαϊκό Έτος 2012-2013
Τμήμα Μηχανικών Η/Υ και Πληροφορικής

Υποχρεωτική προγραμματιστική εργασία

Εξέταση Μαθήματος: για την επιτυχή ολοκλήρωση του μαθήματος προβλέπεται συμμετοχή στις διαλέξεις και τα φροντιστήρια του μαθήματος, ***εκπόνηση μιας υποχρεωτικής εργασίας*** και συμμετοχή σε προφορική εξέταση. Η επίδοση στην εργασία συμμετέχει κατά 70% στην διαμόρφωση του τελικού βαθμού. Η βαθμολόγηση της εργασίας θα γίνει σε 2 στάδια: η πρώτη εξέταση (πρόοδος) θα γίνει σε ημερομηνία που θα ανακοινωθεί στο φόρουμ του μαθήματος και θα αφορά το 30% του βαθμού της εργασίας, ενώ η δεύτερη εξέταση θα αφορά το υπόλοιπο 70% του βαθμού της εργασίας, και θα διεξαχθεί στην τελική εξέταση του μαθήματος, μαζί με την προφορική εξέταση.

Ημερομηνία Παράδοσης:

- Πρόοδος: Παράδοση κώδικα: **5/12/2012**. Μετά την παράδοση του κώδικα θα εξακολουθήσει σύντομη προφορική εξέταση (10-15 λεπτών) της άσκησης σε ημερομηνία που θα ανακοινωθεί στο φόρουμ. Η πρόοδος συμμετέχει κατά 30% στην διαμόρφωση της τελικής βαθμολογίας της άσκησης, και αναμένεται να έχει υλοποιηθεί το 30% της άσκησης.
- Τελική ημερομηνία παράδοσης της άσκησης: **1** ημερολογιακή **εβδομάδα** πριν την ημερομηνία **προφορικής εξέτασης** του μαθήματος.

Εξέταση: Προφορική, μεταξύ 10-14 Δεκεμβρίου 2012 (Πρόοδος), και στο τέλος της εξεταστικής. Θα βγει ανακοίνωση στο φόρουμ.

Ομάδες: **Τριών (3)** ατόμων.

Βαθμολόγηση: ΑΣΚΗΣΗ: 70% του τελικού βαθμού.
Η υλοποίηση της άσκησης θα αξιολογηθεί με βάση του κατά πόσο ανταποκρίνεται στην εργασία της εκφώνησης της άσκησης καθώς και στην έμφαση που έχει δοθεί στην ταχύτερη εκτέλεση των ζητούμενων εργασιών. Σημειώνεται ότι η βαθμολογία περιλαμβάνει και το κατά πόσο η ομάδα θα μπορεί να απαντήσει σε ερωτήσεις πάνω στην άσκηση, ενώ όλοι οι συμμετέχοντες της ομάδας πρέπει να έχουν ασχοληθεί προσωπικά για να λάβουν προβιβάσιμο βαθμό.

Άσκηση

Κάθε ομάδα μπορεί να επιλέξει και να υλοποιήσει **μία (1)** από τις ακόλουθες ασκήσεις (Α **ή** Β, **όχι** και τις δύο). Για την άσκηση, δεν απαιτούνται γραφικές διεπαφές χρήστη. Είναι αποδεκτό η είσοδος και η έξοδος του συστήματος να έχουν π.χ. την μορφή απλών αρχείων κειμένου ή αρχείων XML. Η άσκηση μπορεί να υλοποιηθεί σε οποιαδήποτε γλώσσα

προγραμματισμού, ενώ μπορούν να χρησιμοποιηθούν υπάρχοντα εργαλεία γλωσσικής τεχνολογίας κατόπιν συνεννόησης με τον διδάσκοντα.

Άσκηση Α – Κατηγοριοποίηση εγγράφων

Σκοπός της άσκησης Α είναι η υλοποίηση ενός συστήματος κατηγοριοποίησης κειμένων σε προκαθορισμένες θεματικές κατηγορίες. Η είσοδος του συστήματος αποτελείται από δύο σύνολα από έγγραφα (συλλογή Ε και συλλογή Α), και από ένα σύνολο θεματικών κατηγοριών ΘΚ. Η συλλογή Ε αποτελείται από έγγραφα ήδη κατηγοριοποιημένα στις προκαθορισμένες θεματικές κατηγορίες ΘΚ. Η συλλογή Α αποτελείται από έγγραφα τα οποία δεν είναι κατηγοριοποιημένα στις κατηγορίες του συνόλου ΘΚ, και τα οποία πρέπει να κατηγοριοποιηθούν στις θεματικές κατηγορίες ΘΚ.

Η κατηγοριοποίηση ενός εγγράφου Χ σε μια θεματική κατηγορία Κ γίνεται συγκρίνοντας το έγγραφο Χ με όλα τα μοντέλα των κατηγοριών ΘΚ, και επιλέγοντας την κατηγορία Κ από το ΘΚ που ταιριάζει περισσότερο με το έγγραφο Χ. Κάθε έγγραφο Χ θα αναπαρασταθεί από ένα διάνυσμα χαρακτηριστικών σταθερού μήκους. Κάθε κατηγορία Κ θα αναπαρασταθεί από ένα σύνολο διανυσμάτων χαρακτηριστικών σταθερού μήκους, το οποίο περιέχει όλα τα διανύσματα των εγγράφων της συλλογής Ε που ανήκουν στην κατηγορία Κ. Συνεπώς, η άσκηση αποτελείται από τρεις βασικές υπο-εργασίες:

- Ορισμός ενός χώρου χαρακτηριστικών S, πεπερασμένου μεγέθους (π.χ. 8.000 χαρακτηριστικά). Τα χαρακτηριστικά αυτά θα είναι θέματα (stems) από λέξεις που θα επιλεγούν από τα κείμενα της συλλογής Ε.
- Κατασκευή διανυσμάτων χαρακτηριστικών για όλα τα έγγραφα της συλλογής Ε. Το διάνυσμα του εγγράφου Χ θα περιέχει τα βάρη για όλα τα χαρακτηριστικά του χώρου S. Κάθε βάρος θα είναι το κανονικοποιημένο TF-IDF του χαρακτηριστικού στο κείμενο Χ.
- Κατηγοριοποίηση εγγράφου Χ από την συλλογή Α: δημιουργείται ένα διάνυσμα χαρακτηριστικών όπως στην περίπτωση των εγγράφων της συλλογής Ε, το οποίο συγκρίνεται με όλα τα διανύσματα των εγγράφων της συλλογής Ε, βάση συναρτήσεων σχετικότητας (similarity functions). Το έγγραφο Χ κατηγοριοποιείται στην κατηγορία του εγγράφου με το οποίο είναι πιο σχετικό.

Σαν συναρτήσεις σχετικότητας μπορούν να χρησιμοποιηθούν οι cosine similarity, οι συντελεστές Tanimoto, Jaccard, Dice, κλπ., καθώς και συναρτήσεις που αναφέρονται στο κεφάλαιο 4 του συγγράμματος του μαθήματος. Σαν συλλογή κειμένων θα χρησιμοποιηθεί ένα μέρος του σώματος κειμένων “20 newsgroups corpus”, το οποίο είναι διαθέσιμο από την διεύθυνση http://qwone.com/~jason/20Newsgroups/.

Μέρη του συστήματος

Προ-επεξεργασία των συλλογών Ε και Α

Η άσκηση απαιτεί την αναγνώριση λέξεων και την θεματοποίησή τους. Μπορούν να χρησιμοποιηθούν έτοιμα εργαλεία γλωσσικής τεχνολογίας για αυτές τις εργασίες για την Αγγλική γλώσσα.

Δημιουργία χώρου χαρακτηριστικών

Θα πρέπει να υπολογιστεί το TF-IDF για όλα τα θέματα όλων των λέξεων όλων των εγγράφων της συλλογής Ε, και να επιλεγούν τα Ν (ανάλογα με το επιθυμητό μέγεθος του χώρου χαρακτηριστικών) με την μεγαλύτερη τιμή σύμφωνα με το TF-IDF.

Δημιουργία διανυσμάτων χαρακτηριστικών

Θα πρέπει να είναι δυνατή η δημιουργία διανυσμάτων τόσο για έγγραφα της συλλογής Ε όσο και της συλλογής Α. Το IDF στην περίπτωση των εγγράφων της συλλογής Α, θα ταυτίζεται με το IDF του χαρακτηριστικού στην συλλογή Ε. Είναι επιθυμητό η αναπαράσταση των διανυσμάτων να γίνει με αραιό τρόπο (sparse vectors).

Σύγκριση διανυσμάτων χαρακτηριστικών

Θα πρέπει να υλοποιηθεί υποσύστημα σύγκρισης δύο διανυσμάτων σταθερού μήκους, χρησιμοποιώντας τουλάχιστον 2 μετρικές σχετικότητας.

Άσκηση Β – Εύρεση συνεκφερόμενων λέξεων

Σκοπός της άσκησης Β είναι η υλοποίηση ενός συστήματος για την εξαγωγή λέξεων που συνεκφέρονται πολύ συχνά μαζί (collocations ή χαλαρά συνώνυμα) μέσα σε κείμενα γραμμένα στην Ελληνική γλώσσα. Στο πλαίσιο της άσκησης πρέπει να υλοποιηθούν οι δύο μέθοδοι εύρεσης συνεκφερόμενων λέξεων που περιγράφονται αναλυτικά στο κεφάλαιο 5 του συγγράμματος του μαθήματος. Η πρώτη μέθοδος είναι η μέθοδος του μέσου και της διασποράς (mean and variance method), η οποία υποστηρίζει τις αποστάσεις μεταξύ των λέξεων σε ένα σώμα κειμένων, και ψάχνει για πρότυπα αποστάσεων με χαμηλή διασπορά (spread). Η δεύτερη μέθοδος βασίζεται στον X^2 έλεγχο, μια στατιστική προσέγγιση για την εκτίμηση του εάν ένα συμβάν είναι τυχαίο γεγονός. Η άσκηση αποτελείται από τρεις βασικές υπο-εργασίες:

- Εξαγωγή διγραμμάτων (bigrams): το σώμα κειμένων πρέπει να περάσει από την φάση προ-επεξεργασίας (αναγνώριση προτάσεων και εξαγωγή θεμάτων – stems), και στην συνέχεια να εξαχθεί η λίστα των 1000 πιο συχνά εμφανιζόμενων διγραμμάτων από θέματα στο σώμα κειμένων. Η λίστα αυτή θα αποτελέσει την βάση των λέξεων που θα εξεταστεί αν είναι συνεκφερόμενες ή όχι.
- Έλεγχος αν ένα δίγραμμα από θέματα είναι συνεκφερόμενο ή όχι με την μέθοδο του μέσου και διασποράς. Κάθε δίγραμμα εξετάζεται με την μέθοδο, χρησιμοποιώντας ένα παράθυρο 10 λέξεων: όλες οι εμφανίσεις των δύο θεμάτων του διγράμματος εντοπίζονται στο σώμα κειμένων, και μετράται η μεταξύ τους απόσταση σε λέξεις. Υπολογίζεται η μέση τιμή και η τυπική απόκλιση για το δίγραμμα. Η διαδικασία επαναλαμβάνεται για όλα τα διγράμματα, τα οποία ταξινομούνται ανάλογα με την τυπική απόκλιση, από το μικρότερο προς το μεγαλύτερο.
- Έλεγχος αν ένα δίγραμμα από θέματα είναι συνεκφερόμενο ή όχι με την μέθοδο του ελέγχου X^2. Για κάθε δίγραμμα $(stem_1, stem_2)$ δημιουργείται ο ακόλουθος πίνακας:

	$w_1 = stem_1$	$w_1 \neq stem_1$
$w_2 = stem_2$	f_1	f_2
$w_2 \neq stem_2$	f_3	f_4

όπου f_1, f_2, f_3, f_4 οι συχνότητες εμφάνισης των (w_1, w_2) στο κείμενο. Από τον παραπάνω πίνακα, η τιμή X^2 υπολογίζεται ως εξής:

$$X^2 = \frac{N(f_1 f_4 - f_2 f_3)^2}{(f_1 + f_2)(f_1 + f_3)(f_2 + f_4)(f_3 + f_4)}, N = f_1 + f_2 + f_3 + f_4$$

Μέρη του συστήματος

Προ-επεξεργασία του σώματος κειμένων

Η άσκηση απαιτεί την αναγνώριση λέξεων και την θεματοποίησή τους. Μπορούν να χρησιμοποιηθούν έτοιμα εργαλεία γλωσσικής τεχνολογίας για αυτές τις εργασίες για την Ελληνική γλώσσα.

Εύρεση διγραμμάτων

Θα πρέπει να υπολογιστούν οι συχνότητες εμφάνισης όλων των διγραμμάτων από θέματα στο σώμα κειμένων, και τα διγράμματα να ταξινομηθούν με βάση την συχνότητα εμφάνισης, από το συχνότερο προς το λιγότερο συχνό. Από την λίστα θα κρατηθούν τα 1000 συχνότερα διγράμματα.

Εντοπισμός διγραμμάτων σε κείμενο

Θα πρέπει να κατασκευαστεί υποσύστημα εντοπισμού της πρώτης λέξης του διγράμματος σε κείμενο, και κατά πόσο η δεύτερη λέξη του διγράμματος βρίσκεται πριν ή μετά την πρώτη λέξη (για την μέθοδο του μέσου και διασποράς). Για την μέθοδο X^2 απαιτείται η εύρεση του θέματος της επόμενης λέξης και κατά πόσο συμφωνεί με την δεύτερη του διγράμματος, αλλά και η εύρεση της δεύτερης λέξης του διγράμματος και της λέξης πριν από αυτό (και αν το θέμα της συμφωνεί με την πρώτη λέξη του διγράμματος).

Εξαγωγή μέσου και διασποράς και X^2 για κάθε δίγραμμα

Με βάση τις συχνότητες εμφάνισης για τα διάφορα συστατικά ενός διγράμματος, θα πρέπει να υπολογιστεί η διασπορά (τυπική απόκλιση μέσης τιμής αποστάσεων) και το X^2,και να ταξινομηθούν τα διγράμματα με βάση και τις δύο αυτές τιμές.