Γλωσσική Τεχνολογία

Ακαδημαϊκό Έτος 2011-2012 - Project Σεπτεμβρίου

***Ημερομηνία Παράδοσης*: Στην εξέταση του μαθήματος**

***Εξέταση*: Προφορική, στο τέλος της εξεταστικής. Θα βγει ανακοίνωση στο forum.**

***Ομάδες 2 ατόμων*.**

***Βαθμολόγηση*: 80% του τελικού βαθμού. Αν η ομάδα κατ' εξαίρεση αποτελείται από 3 άτομα, τότε 70% του τελικού βαθμού.**

ΑΣΚΗΣΗ

Δημιουργία Ευρετηρίων Συλλογής Κειμένων

Σκοπός της άσκησης είναι η υλοποίηση ενός συστήματος επεξεργασίας μιας συλλογής κειμένων με στόχο τη δημιουργία ευρετηρίων για τη διαχείρισή της. Το σύστημα θα αποτελείται από ένα υποσύστημα προεπεξεργασίας των δεδομένων και κατασκευής των ζητούμενων ευρετηρίων καθώς και από υποσυστήματα τα οποία θα είναι σε θέση να χρησιμοποιούν τα ήδη δημιουργημένα ευρετήρια για να υλοποιήσουν τη ζητούμενη λειτουργικότητα. Προκειμένου να αξιολογηθεί η απόδοση του κάθε ευρετηρίου θα υλοποιηθεί μηχανισμός υποβολής ερωτημάτων στα δεικτοδοτημένα κείμενα της συλλογής, μέσω του οποίου θα υποβάλετε ερωτήματα προς τα ευρετήρια και θα ανακτάτε τα κείμενα της συλλογής που σχετίζονται με αυτά.

Η υλοποίησή σας θα αξιολογηθεί με βάση το κατά πόσο ανταποκρίνεται στις ανάγκες πραγματικών συστημάτων δεικτοδότησης. Η απόδοση των σχεδιαστικών επιλογών που θα κάνετε πρέπει να είναι τέτοια ώστε να δίνεται έμφαση στην ταχύτερη εκτέλεση των ζητούμενων εργασιών. Επίσης η δημιουργία των ζητούμενων ευρετηρίων θα πρέπει να συνοδεύεται από μηχανισμούς αποθήκευσης και επαναφόρτωσής τους έτσι ώστε η προεπεξεργασία να γίνεται μία μόνο φορά.

Θα χρησιμοποιήσετε τη συλλογή κειμένων της wikipedia που δίνεται στη σελίδα του μαθήματος και για τις μετρήσεις που ζητούνται θα χρησιμοποιήσετε το αρχείο με τη λίστα ερωτημάτων που δίνεται.

Εκτός από τη δεικτοδότηση της συλλογής, θα κληθείτε να μετρήσετε και κάποια στατιστικά για τη συλλογή, τα οποία θα παρουσιάσετε σε (σύντομη) αναφορά.

Διαβάσετε προσεχτικά την εκφώνηση και απαντήσετε με σαφήνεια στα ερωτήματα. Συνιστάται η χρήση Python, καθώς αρκετά από τα εργαλεία που θα χρειαστείτε είναι ήδη υλοποιημένα στο NLTK.

Σημειώνεται ότι <u>θα βαθμολογηθεί το κατά πόσο θα μπορείτε να απαντήσετε σε ερωτήσεις πάνω στο πρότζεκτ</u> και όχι το ίδιο το πρότζεκτ, επομένως <u>πρέπει να έχετε ασχοληθεί προσωπικά για να πάρετε προβιβάσιμο βαθμό</u>.

Σχεδιασμός των ευρετηρίων

Η απόδοση του συστήματος που θα υλοποιήσετε εξαρτάται σε μεγάλο βαθμό από την επιλογή της κατάλληλης δομής δεδομένων για την φόρτωση των ζητούμενων ευρετηρίων στη μνήμη. Η δομή δεδομένων που θα χρησιμοποιήσετε πρέπει να ελαχιστοποιεί σε κάθε περίπτωση το χρόνο αναζήτησης και κατασκευής των ευρετηρίων.

Στα πλαίσια της άσκησης θα κατασκευάσετε δύο τύπους ευρετηρίων, ένα κανονικό και ένα ανεστραμμένο, στα οποία θα αποθηκεύσετε τα διανύσματα των κειμένων της συλλογής που σας δίνεται, όπως αυτά κατασκευάζονται με βάση τη θεωρία του vector space model.

Αρχικά σε κάθε κείμενο που ανήκει στη συλλογή θα πρέπει να δίνεται ένα id. Το id ενός κειμένου είναι ένα μοναδικό αναγνωριστικό του. Μπορείτε να δώσετε δικό σας id σε κάθε κείμενο για να το χρησιμοποιείτε εσωτερικά ή να χρησιμοποιήσετε το path που έχει αποθηκευτεί τοπικά, δικαιολογώντας την επιλογή σας. Σε κάθε περίπτωση θα πρέπει με δεδομένο το id να μπορείτε να ανακτήσετε το full path.

Το ανεστραμμένο ευρετήριο είναι μια συλλογή εγγραφών της μορφής:
<λήμμα, {<id_κειμένου1, βάρος1>, <id_κειμένου2, βάρος2>,…}>
Για κάθε μοναδικό λήμμα που συναντάται στη συλλογή κειμένων δημιουργείται μια εγγραφή στο ανεστραμμένο ευρετήριο. Η εγγραφή περιέχει το λήμμα καθώς και το σύνολο των κειμένων στις οποίες εμφανίζεται. Για κάθε κείμενο όπου εμφανίζεται αποθηκεύεται επίσης το βάρος του λήμματος σε αυτό.

Το κανονικό ευρετήριο είναι μια συλλογή εγγραφών της μορφής:
<id_κειμένου, {<λήμμα1, βάρος1>, <λήμμα2, βάρος2>,…}>
Για κάθε κείμενο στη συλλογή δημιουργείται μια εγγραφή. Η εγγραφή περιέχει το id του κειμένου καθώς και το σύνολο των λημμάτων τα οποία εμφανίζονται σε αυτό. Για κάθε λήμμα αποθηκεύεται και το βάρος του για το συγκεκριμένο κείμενο.

Μέρη του Συστήματος

Προεπεξεργασία της συλλογής

Η προεπεξεργασία της συλλογής κειμένων είναι η ολοκληρωμένη διαδικασία επεξεργασίας των αρχείων, δημιουργίας των ευρετηρίων και αποθήκευσής τους σε κατάλληλη μορφή ώστε να μπορούν να «φορτωθούν» χωρίς να επαναλαμβάνεται η κατασκευή τους. Μπορείτε να οργανώσετε την υλοποίηση των επιμέρους εργασιών (tokenization, tagging, δημιουργία ευρετηρίων, υπολογισμό βαρών κλπ) με οποιοδήποτε τρόπο θέλετε έτσι ώστε να επιτύχετε την ορθότερη λειτουργία του συστήματος και τη μεγαλύτερη δυνατή ταχύτητα κατασκευής. Ακολούθως περιγράφονται οι βασικές απαιτήσεις για τις επιμέρους εργασίες κατά την κατασκευή των ευρετηρίων.

Tokenization

Στα πλαίσια του tokenization καλείστε να προετοιμάσετε τα κείμενα που σας δίνονται ώστε να δοθούν στον μορφοσυντακτικό αναλυτή ως είσοδος. Θα προετοιμάσετε την είσοδο αφαιρώντας μεταδεδομένα αν υπάρχουν (xml tags, html markup κλπ), χωρίζοντας το κείμενο

σε μια λέξη ανά γραμμή αν χρειάζεται και πραγματοποιώντας όποια άλλη τροποποίηση κρίνετε απαραίτητη.

Οι επιλογές σας στο tokenization θα πρέπει να στοχεύουν στην μεγαλύτερη δυνατή απόδοση του μορφοσυντακτικού αναλυτή (tagger). Ανάλογα με τον μορφοσυντακτικό αναλυτή που θα χρησιμοποιήσετε καλείστε να υλοποιήσετε και τον αντίστοιχο κώδικα ώστε να διαμορφώσετε σωστά την είσοδο που πρέπει να του δώσετε. Σε περιπτώσεις όπου δεν είναι προφανής ο τρόπος που πρέπει να χωριστεί μια πρόταση ή μια λέξη (πχ Teacher's), θα δοκιμάζετε τον tagger και θα επιλέγετε την λύση που τον οδηγεί σε σωστό αποτέλεσμα.

Μορφοσυντακτική Ανάλυση

Σχολιάστε μορφοσυντακτικά τις λέξεις του κάθε tokenized κειμένου. Για το μορφοσυντακτικό σχολιασμό χρησιμοποιήστε κάποιον από τους προτεινόμενους PoS-Taggers (ανάλογα με το περιβάλλον υλοποίησης που έχετε επιλέξει) που θα κατεβάσετε από το site του μαθήματος.

Επιλογή Όρων

Οι μορφοσυντακτικοί αναλυτές που σας δίνονται πραγατοποιούν και αναγωγή στον πρώτο κλιτικό τύπο (lemmatization). Για κάθε λέξη επιστρέφουν το μέρος του λόγου αλλά και το λήμμα στο οποίο αντιστοιχεί. Οι όροι που θα χρησιμοποιήσετε για την δεικτοδότηση των κειμένων είναι τα λήμματα που έχουν προκύψει και όχι οι αρχικές λέξεις.

Στο σύνολο των όρων που θα δεικτοδοτήσετε δεν θα πρέπει να λάβετε υπόψη τους τερματικούς όρους (stop words). Οι τερματικοί όροι είναι λέξεις που δεν έχουν σημασιολογικό περιεχόμενο και εμφανίζονται σε όλα τα κείμενα, με αποτέλεσμα να μην αποτελούν χρήσιμους όρους δεικτοδότησης. Στο link: http://www.infogistics.com/tagset.html θα βρείτε δύο πίνακες, έναν με τα PoS tags για open class categories και έναν με τα PoS tags για closed class categories. Τα open class categories είναι γραμματικές κατηγορίες των λέξεων που έχουν σημασιολογικό περιεχόμενο και άρα τις χρειαζόμαστε. Αντίθετα, τα closed class categories είναι γραμματικές κατηγορίες για λέξεις άνευ σημασιολογικού περιεχομένου, δηλ., stop-words. Συνεπώς, για να εξαλείψετε τους τερματικούς όρους από κάθε μορφοσυντακτικά σχολιασμένο κείμενο της συλλογής θα πρέπει να αφαιρέσετε τις λέξεις στις οποίες έχει ανατεθεί ένα closed class category tag.

Υπολογισμός Βαρών

Το βάρος του κάθε λήμματος για ένα κείμενο αντιπροσωπεύει το βαθμό σπουδαιότητας του λήμματος για το συγκεκριμένο κείμενο και θα το υπολογίσετε χρησιμοποιώντας τη μετρική TF-IDF. Ο παρακάτω τύπος δίνει το βάρος του λήμματος i για ένα συγκεκριμένο κείμενο:

$$\boldsymbol{weight_i} = \frac{\boldsymbol{tf_i} \times \boldsymbol{idf_i}}{\sqrt{\sum_{\boldsymbol{k} \in \boldsymbol{vector}} (\boldsymbol{tf_k} \times \boldsymbol{idf_k})^2}}$$

όπου **tf** (term frequency) είναι η συχνότητα εμφάνισης ενός όρου σε ένα κείμενο, **idf** (inverse document frequency) η αντίστροφη συχνότητα κειμένου στη συλλογή και παρονομαστής (παράγοντας κανονικοποίησης) **το ευκλείδιο μήκος** του διανύσματος για κάθε κείμενο, το οποίο υπολογίζεται με τη χρήση των παραγόντων $\boldsymbol{tf} \times \boldsymbol{idf}$.

Η **idf** για έναν όρο **i** υπολογίζεται από τον τύπο:

$$idf_i = \log \frac{N}{n_i}$$

όπου **N** είναι ο συνολικός αριθμός κειμένων της συλλογής και $\mathbf{n_i}$ ο αριθμός των κειμένων της συλλογής στα οποία εμφανίζεται ο όρος **i**.

Αποθήκευση και επαναφόρτωση των ευρετηρίων

Για κάθε ευρετήριο θα υλοποιήσετε τις απαραίτητες λειτουργίες έτσι ώστε να είναι δυνατή η αποθήκευση της αντίστοιχης δομής δεδομένων σε xml αρχείο και η επαναφόρτωση του xml αρχείου στην δομή δεδομένων. Ο σχεδιασμός της xml δομής για κάθε ευρετήριο πρέπει να είναι τέτοιος ώστε να διεκολύνει τη γρήγορη αποθήκευση και επαναφόρτωση. Επίσης τα XML αρχεία που θα προκύψουν πρέπει να είναι ορθά δομημένα (well-formed).

Στόχος είναι να μην επαναλαμβάνεται κάθε φορά η κατασκευή των ευρετηρίων, αλλά να είναι δυνατή η απευθείας φόρτωσή τους για να μπορούν να εκτελεστούν λειτουργίες πάνω σε αυτά.

Αξιολόγηση και σύγκριση των ευρετηρίων

Για την αξιολόγηση των ευρετηρίων θα υπολοποιηθεί μηχανισμός υποβολής ερωτημάτων στα δύο ευρετήρια που θα προκύψουν. Θα πρέπει να υλοποιήσετε τη διαδικασία αναζήτησης στο ανεστραμμένο και στο κανονικό ευρετήριο και πειραματικά να χρονομετρήσετε τα δύο ευρετήρια για να συγκρίνετε την απόδοσή τους κατά την αναζήτηση. Ακολουθεί η περιγραφή των βασικών σημείων της πειραματικής διαδικασίας.

Προδιαγραφές αναζήτησης

Η αναζήτηση θα πρέπει να είναι υλοποιημένη έτσι ώστε να δέχεται άγνωστο αριθμό όρων για κάθε ερώτημα. Υποθέτουμε για τη διευκόλυνσή σας ότι οι όροι του κάθε ερωτήματος είναι σε πρώτο κλιτικό τύπο. Επίσης η αναζήτηση θα πρέπει να μπορεί να γίνει ανεξάρτητα από την κατασκευή του ευρετηρίου. Θα «φορτώνονται» δηλαδή τα ευρετήρια από αρχεία στις αντίστοιχες δομές δεδομένων και μετά θα πραγματοποιείται η αναζήτηση.

Για κάθε ερώτημα ο μηχανισμός αναζήτησης θα ψάχνει στο ευρετήριο και θα εντοπίζει τα σχετικά κείμενα. Έπειτα θα τα ταξινομεί με βάση την ομοιότητα του διανύσματος του κάθε κειμένου με το διάνυσμα του ερωτήματος (cosine similarity ή εσωτερικό γινόμενο). Σημειώνεται ότι για το διάνυσμα του ερωτήματος η τιμή συντεταγμένης όταν ο όρος υπάρχει είναι 1 και όταν δεν υπάρχει 0. Τα αποτελέσματα θα επιστρέφονται όχι με id αρχείου αλλά με full path στο δίσκο.

Ο τρόπος αναζήτησης σε κάθε περίπτωση εξαρτάται προφανώς από το ευρετήριο στο οποίο ψάχνουμε, επομένως θα υλοποιήσετε διαφορετική αναζήτηση για κάθε τύπο ευρετήριου (ανεστραμμένο και κανονικό). Σκοπός είναι για κάθε τύπο ευρετηρίου να έχετε υλοποιήσει τον αποδοτικότερο τρόπο αναζήτησης.

Θα πρέπει να υποστηρίζεται υποβολή ερωτημάτων από χρήστη και παρουσίαση των αποτελεσμάτων, αλλά και υποβολή πολλαπλών ερωτημάτων από αρχείο.

Πειραματική μέτρηση

Για κάθε ευρετήριο θα μετρήσετε το μέσο χρόνο απόκρισης, υποβάλλοντας τα ερωτήματα που περιέχονται στο αρχείο queries.txt (ένα ερώτημα ανά γραμμή). Θα μετρήσετε το συνολικό χρόνο και θα διαιρέσετε με τον αριθμό των ερωτημάτων για να υπολογίσετε το μέσο χρόνο απόκρισης. Αν οι χρόνοι είναι πολύ μικροί για να μετρηθούν, επαναλάβατε πολλές φορές πριν υπολογίσετε το μέσο χρόνο και διαιρέστε το συνολικό χρόνο με τις φορές επανάληψης του πειράματος επί τον αριθμό των ερωτημάτων.

Προσοχή, δεν πρέπει να συμπεριλάβετε στους χρόνους που μετράτε την ανάγνωση από αρχείο και την εκτύπωση αποτελεσμάτων στην οθόνη.

Μετρήσεις Χρόνων

Κατά την κατασκευή αλλά και την πειραματική εφαρμογή του συστήματος που θα υλοποιήσετε θα μετρήσετε τους αντίστοιχους χρόνους. Θα πρέπει να προσέξετε κατά τη μέτρηση των χρόνων που ζητούνται να συμπεριλάβετε μόνο τις διαδικασίες που πρέπει να χρονομετρηθούν και όχι χρονοβόρες λειτουργίες όπως I/O, κλήσεις εξωτερικών διεργασιών ή εκτύπωση στο standard output.

Τα μεγέθη που καλείστε να μετρήσετε στα πλαίσια της άσκησης είναι τα ακόλουθα:

1. **Χρόνος κατασκευής ευρετηρίων:** Στο συγκεκριμένο χρόνο δεν συμπεριλαμβάνεται το tokenization και το tagging, καθώς είναι κλήση εξωτερικής διεργασίας και πολύ χρονοβόρο. Θεωρούμε σαν διαδικασία κατασκευής τη διαδικασία που αρχίζει μετά το tagging.
2. **Μέγεθος ευρετηρίων:** Αρκεί να μετρήσετε και να συγκρίνετε το τελικό μέγεθος των xml αρχείων στα οποία αποθηκεύονται τα ευρετήρια. Σχολιάστε τη διαφορά αν υπάρχει.
3. **Μέσος χρόνος απόκρισης ευρετηρίων:** Μετρήστε όπως περιγράφεται στην πειραματική εφαρμογή και σχολιάστε τη διαφορά αν υπάρχει.

Στατιστικά για τη συλλογή

Για τη συλλογή κειμένων που σας δίνεται, καλείστε να πραγματοποιήσετε μετρήσεις συγκεκριμένων στατιστικών στοιχείων, τις οποίες θα συμπεριλάβετε στην αναφορά σας. Οι μετρήσεις μπορούν να πραγματοποιηθούν σε όποιο σημείο της προεπεξεργασίας θεωρείτε καλύτερο. Οι μετρήσεις θα πραγματοποιούνται μια φορά κατά την κατασκευή των ευρετηρίων. Τα στοιχεία που θα μετρήσετε είναι:

1. **Αριθμός κειμένων**: ο συνολικός αριθμός κειμένων
2. **Αριθμός λέξεων στη συλλογή:** ο συνολικός αριθμός λέξεων σε όλα τα κείμενα. Αν υπάρχει αμφιβολία στο τι θεωρείται λέξη και τι όχι, κάντε τη σχετική παραδοχή και σημειώστε τη στην αναφορά.
3. **Αριθμός ουσιαστικών στη συλλογή:** Ο συνολικός αριθμός ουσιαστικών σε όλα τα κείμενα.
4. **Αριθμός λέξεων ανά άρθρο:** Πραγματοποιήστε τη μέτρηση σε συμφωνία με την παραδοχή που κάνατε στο 2.
5. **Αριθμός ουσιαστικών ανά άρθρο**

6. **<u>Αριθμός urls ανά άρθρο:</u>** Εννοούμε τις διευθύνσεις που περιέχονται στο άρθρο (δλδ της μορφής http:// ...) σε οποιοδήποτε σημείο του. Ενδιαφερόμαστε μόνο για πλήρεις διευθύνσεις και όχι σχετικές (άν υπάρχουν).

Παραδοτέα

1. Ο πηγαίος κώδικας όλων των παραπάνω υποσυστημάτων. Ο σχολιασμός του κώδικα να γίνει απαραίτητα σε επίπεδο συναρτήσεων (λειτουργία, ορίσματα, έξοδος) αλλά και εσωτερικά των συναρτήσεων όπου κρίνεται αναγκαίο για να γίνει κατανοητός.
2. Μια σύντομη αναφορά που θα περιέχει τα αποτελεσματα των μετρήσεων που απαιτούνται από την άσκηση καθώς και ο σχολιασμός που ζητάται.