Γλωσσική Τεχνολογία

Εργασία 2013-2014

Εισαγωγικά

Η εργασία θα πραγματοποιηθεί σε Python με χρήση του NLTK. Παραδοτέο θα είναι ένας φάκελος αρχείων που θα περιέχει τους 10 επιμέρους φακέλους που θα αντιστοιχούν στον κώδικα κάθε ερωτήματος. Ο κώδικας θα πρέπει να έχει σύντομα σχόλια πάνω του. Το υλικό που θα σας είναι απαραίτητο, μπορείτε να το βρείτε στις διαφάνειες και τις προτεινόμενες πηγές του φροντιστηρίου

Τα θεωρητικά και προγραμματιστικά θέματα που περιλαμβάνονται στην άσκηση θεωρούνται ύλη προς εξέταση στην τελική εξέταση του μαθήματος. Για την επίλυση των ασκήσεων επιτρέπεται να χρησιμοποιήσετε οποιοδήποτε εργαλείο από το NLTK. Λόγω των εργαλείων, όπου αναφέρεται κείμενο εννοούμε αγγλικό κείμενο.

Τα προβλήματα που περιλαμβάνονται στο πεδίο δεν έχουν επιλυθεί ακόμα στο 100% των περιπτώσεων. Για παράδειγμα ένας tokenizer ή ένας tagger δουλεύουν μεν με μεγάλη ακρίβεια, αλλά δεν λειτουργούν απόλυτα σωστά σε όλες τις περιπτώσεις. Γι' αυτό το λόγο, για τις ασκήσεις θα γίνουν δεκτές με άριστη βαθμολογία λύσεις των οποίων η απόδοση είναι εφάμιλλη με αυτή που επιτρέπουν τα εργαλεία που προσφέρει το NLTK και οι οποίες βασίζονται στις θεωρητικές παρατηρήσεις που έχουν γίνει στα πλαίσια του μαθήματος και των διαφανειών. Αν η λύση που προτείνετε για κάποια άσκηση συμπεριλαμβάνει παραδοχές απο μέρους σας ή απαιτεί τεκμηρίωση, να υπάρχει η ανάλογη εξήγηση.

Ασκήσεις

Άσκηση 1η

Υλοποιήστε ένα πρόγραμμα που δέχεται ένα κείμενο και το χωρίζει σε προτάσεις.

Άσκηση 2η

Υλοποιήστε έναν απλό tokenizer ο οποίος δέχεται ένα κείμενο ως είσοδο και παράγει μια ακολουθία tokens ως έξοδο. Εκτός από τον διαχωρισμό λέξεων και σημείων στίξης θα πρέπει να αναγνωρίζει αριθμούς, ηλεκτρονικές διευθύνσεις και διευθύνσεις ηλεκτρονικού ταχυδρομείου.

Άσκηση 3η

Υλοποιήστε ένα πρόγραμμα που δέχεται μια html σελίδα και απομονώνει το κείμενο από αυτή.

Άσκηση 4η

Υλοποιήστε ένα πρόγραμμα που δέχεται μια html σελίδα και παράγει από αυτή μια λίστα των URLs στα οποία δείχνει και για το κάθε URL το κείμενο του συνδέσμου.

Άσκηση 5η

Υλοποιήστε ένα πρόγραμμα που δέχεται ένα κείμενο και μετράει τις συχνότητες εμφάνισης των Μερών του Λόγου που περιέχει. Πχ πόσα ουσιαστικά, πόσα ρήματα κλπ.

Άσκηση 6η

Υλοποιήστε ένα πρόγραμμα που δέχεται ένα κείμενο και παράγει τη λίστα των μοναδικών λημμάτων των ουσιαστικών που περιέχει.

Άσκηση 7η

Υλοποιήστε ένα πρόγραμμα που δέχεται ένα κείμενο και παράγει τη λίστα των stems που περιέχει καθώς και τη συχνότητα εμφάνισης του κάθε stem.

Άσκηση 9η

Με τη βοήθεια του Brown Corpus, βρείτε τις 200 λέξεις με το μεγαλύτερο tfidf score στην κατηγορία επιστημονικής φαντασίας, όπου tfidf score=(tf/n)*idf, tf: συχνότητα εμφάνισης στην κατηγορία, n: συνολική συχνότητα εμφάνισης στο corpus και idf=log(n/tf). Βρείτε επίσης τις 200 λέξεις με τη μεγαλύτερη συχνότητα εμφάνισης στην κατηγορία επιστημονικής φαντασίας.

Άσκηση 10η

Υλοποιήστε ένα ανάστροφο ευρετηρίο α)με dictionary σε σχέση και β) με δισδιάστατη λίστα. Γεμίστε το ευρετήριο με δεδομένα κειμενων της επιλογής σας και πραγματοποιήστε τις απαραίτητες μετρήσεις για να μπορέσετε να συγκρίνετε α) μέσο χρόνο ένθεσης στο ευρετήριο και β) μέσο χρόνο απόκρισης σε αναζήτηση

Άσκηση 11η

Υλοποιήστε ένα πρόγραμμα το οποίο δέχεται τρία ουσιαστικά (που θεωρητικά συνυπάρχουν σε ένα κείμενο) και επιλέγει για το καθένα την καταλληλότερη έννοια από το wordnet, εκτυπώνοντας το synset και τον ορισμό της έννοιας.