ΠΑΝΕΠΙΣΤΗΜΙΟ ΜΑΚΕΔΟΝΙΑΣ
ΣΧΟΛΗ ΕΠΙΣΤΗΜΩΝ ΠΛΗΡΟΦΟΡΙΑΣ
ΤΜΗΜΑ ΕΦΑΡΜΟΣΜΕΝΗΣ ΠΛΗΡΟΦΟΡΙΚΗΣ
ΠΜΣ ΣΤΗΝ ΤΕΧΝΗΤΗ ΝΟΗΜΟΣΥΝΗ ΚΑΙ ΑΝΑΛΥΤΙΚΗ ΔΕΔΟΜΕΝΩΝ

ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ ΚΑΙ ΕΠΕΞΕΡΓΑΣΙΑ ΦΥΣΙΚΗΣ ΓΛΩΣΣΑΣ
1^{Ο} εξάμηνο (2022-23)

Εργασία 10^{ης} εβδομάδας

Για την εργασία αυτή θα πρέπει να χρησιμοποιήσετε τη βιβλιοθήκη NLTK της Python. Μπορείτε να συμβουλευτείτε το παρακάτω βιβλίο, στο οποίο θα βρείτε πολλά σχετικά παραδείγματα:

https://www.nltk.org/book/

Η εργασία έχει δύο ερωτήματα.

Α) Χρησιμοποιώντας τη βιβλιοθήκη NLTK της Python, φορτώστε 10 βιβλία της επιλογής σας από το corpus του project Gutenberg (υποστηρίζεται από το NLTK). Δημιουργείστε ένα λεξικό από όλες τις λέξεις που εμφανίζονται στα βιβλία (χρησιμοποιείστε tokenizer της επιλογής σας, διαθέτει αρκετούς η NLTK).

Χωρίστε τα κείμενα σε προτάσεις (sentence tokenizer) και σε tokens. Υπολογίστε συχνότητες μονογραμμάτων (unigrams), διγραμμάτων (bigrams) και τριγραμμάτων (trigrams). Για τα διγράμματα και τα τριγράμματα θα χρειαστεί να ορίσετε ένα token αρχής πρότασης, όπως και ένα token τέλους πρότασης. Για τα τριγράμματα θα χρειαστεί να ορίσετε δύο tokens αρχής πρότασης και ένα token τέλους πρότασης (δεν χρειάζεται δεύτερο token για το τέλος της πρότασης).

Παράγετε και αναφέρετε στην εργασία σας από 10 προτάσεις χρησιμοποιώντας:

α) Τα διγράμματα

β) Τα τριγράμματα

Θα πρέπει να υποβάλλετε τον κώδικα που γράψατε καθώς και ένα έγγραφο κειμένου με μια σύντομη τεκμηρίωση του κώδικά σας, τις προτάσεις που παρήχθησαν και αναφορές στις πηγές που ενδεχομένως χρησιμοποιήσατε.

Β) Από τη βιβλιοθήκη NLTK της Python φορτώστε το dataset movie_reviews, το οποίο περιλαμβάνει 2.000 αξιολογήσεις ταινιών, εκ των οποίων 1.000 είναι θετικές (pos) και οι υπόλοιπες 1.000 είναι αρνητικές (neg).

Εκπαιδεύστε έναν ταξινομητή (π.χ., Naïve Bayes classifier, νευρωνικό δίκτυο, κλπ) για να ταξινομεί σωστά τα θετικά και τα αρνητικά reviews. Επιλέξτε ποια χαρακτηριστικά θα χρησιμοποιήσετε ως είσοδο (π.χ., μεμονωμένες λέξεις, ngrams, κλπ).

Καταγράψτε σε ένα έγγραφο τι κάνατε και τι αποτελέσματα πήρατε.

Σε περίπτωση που χρησιμοποιήσετε πηγές από το διαδίκτυο (επιτρέπεται), θα πρέπει να τις αναφέρετε.

Ως παράδειγμα, αναφέρεται η παρακάτω πηγή:

https://blog.chapagain.com.np/python-nltk-sentiment-analysis-on-movie-reviews-natural-language-processing-nlp/

Εισηγητής: Γιάννης Ρεφανίδης

yrefanid@uom.edu.gr