Άσκηση 2

Σκοπός της συγκεκριμένης άσκησης είναι εξοικειωθείτε με τις βασικές τεχνικές που χρησιμοποιούνται για τη δημιουργία πίνακα συμπτώσεων, posting lists, vector space model και την εύρεση ομοιότητας μεταξύ δύο εγγράφων.

Διαβάστε την παράγραφο 2.2 του βιβλίου [1]. Για την πρόταση του βιβλίου “Thomas Jefferson….” (sentence1) αλλά και για μία δική σας πρόταση (sentence2), κάντε κατακερματισμό (tokenization) τόσο α) με την λύση που προτείνει το βιβλίο όσο και β) με την χρήση πακέτου του NLTK που κάνατε στην προηγούμενη άσκηση.

Ερώτηση 1: Φτιάξτε τους πίνακες συμπτώσεων για τις δύο προτάσεις sentence1, sentence2 για κάθε μέθοδο κατακερματισμού που κάνατε παραπάνω. Παρουσιάστε τα αποτελέσματά σας και σχολιάστε τα.

Χρησιμοποιείστε το Pandas για την καλύτερη εμφάνιση των παραπάνω αποτελεσμάτων.

Ερώτηση 2: Τι αναπαριστά η κάθε γραμμή; Εκτός της εμφάνισης, τι άλλα πλεονεκτήματα προσφέρει η χρήση του Pandas.

Δείτε το παράδειγμα του βιβλίου όπου στον πίνακα συμπτώσεων φαίνεται η χρήση μίας λέξης (π.χ. Monticello) σε παραπάνω από μία προτάσεις και φτιάξτε άλλες 3 δικές σας προτάσεις (sentence3, sentence4, sentence5) ώστε να δείξετε αντίστοιχα αποτελέσματα (με χρήση κοινών λέξεων) για την πρώτη σας πρόταση (sentence2).

α) Εμφανίστε την ομοιότητα μεταξύ αυτών των προτάσεων.

β) Εμφανίστε την ομοιότητα των βιβλίων text4 και text7 από το NLTK για τις πρώτες 50 λέξεις και εμφανίστε ποιες είναι οι κοινές τους λέξεις.

Ερώτηση 3: Παρουσιάστε τα αποτελέσματά σας για όλα τα παραπάνω και σχολιάστε τι μπορεί να μας δείχνει μία μεγάλη ή μία μικρή ποσοστιαία ομοιότητα δύο εγγράφων κάνοντας χρήση παραδειγμάτων από τα παραπάνω.

Πάρτε βοήθεια από την ιστοσελίδα της GeeksforGeeks ώστε να:

Ερώτηση 4: Βρείτε και δείξτε τις 3 πιο συχνά εμφανιζόμενες λέξεις που εμφανίζονται στις πρώτες 50 λέξεις για το καθένα από τα δύο βιβλία και αναπαραστήσετε τις ως posting lists. Παρουσιάστε και σχολιάστε τα αποτελέσματα.

Διαβάστε το Κεφάλαιο 3 του βιβλίου [1] και ακολουθώντας τις οδηγίες αναπαραστήστε την δική σας πρόταση (sentence2) σε bag-of-words καθώς και τις πρώτες 50 λέξεις των βιβλίων text4 και text7 από το NLTK.

Ερώτηση 5: Υπολογίστε και παρουσιάστε την ομοιότητα συνημίτονων των δύο προτάσεων των δύο βιβλίων.

Ερώτηση 6: Υπολογίστε και παρουσιάστε την ομοιότητα συνημίτονων ολόκληρων των δύο βιβλίων. Τι παρατηρείτε από τα αποτελέσματα. Κάντε σύγκριση των αποτελεσμάτων από τις προηγούμενες ερωτήσεις.

Ακολουθώντας τις οδηγίες των παραγράφων 3.4.2 και 3.4.3 υπολογίστε την ομοιότητα TF-IDF των δύο προτάσεων των δύο βιβλίων με όποιον τρόπο θέλετε.

Ερώτηση 7: Δείξτε τα αποτελέσματά σας. Τι παρατηρείτε σε σχέση με την μέθοδο των συνημίτονων; Τι συμπεράσματα βγάζετε; Βρείτε από **δύο πιθανά** «ερωτήματα» που θα μπορούσατε να κάνετε **για κάθε μία** από τις δύο προτάσεις και επεκτείνετε τα συμπεράσματά σας.

Αναφορές-Βιβλιογραφία

[1] Hobson Lane, Howard Cole, and Hannes Max Hapke. *Naturale Language Processing in Action*.