Προγραμματισμός για Γλωσσική Τεχνολογία Ι (M901)

Εργασίες

Τελική εργασία μαθήματος

Στοιχεία εργασίας

Τίτλος:
Τελική εργασία μαθήματος

Περιγραφή:
Για την τελική εργασία του μαθήματος θα πρέπει να δημιουργήσετε μια κλάση η οποία θα διαβάζει όλα τα αρχεία κειμένων (*.txt) που περιέχει ένας φάκελος στον υπολογιστή σας, θα τα αποθηκεύσει σε κατάλληλες δομές και θα παρέχει μια σειρά μεθόδων για την εξαγωγή διαφόρων πληροφοριών.

Πιό συγκεκριμένα θα πρέπει:

1. Να δημιουργήσετε μια κλάση με το όνομα **FrequenciesGenerator** σε ένα αρχείο **final_project.py**. Η κλαση σας θα πρέπει να υλοποιεί τα παρακάτω:
    - Έναν initializer ο οποίος θα δέχεται ως όρισμα το όνομα ενός φακέλου και θα το αποθηκεύει σε μια instance variable **source_folder**. Επίσης, θα αρχικοποιεί δύο λεξικά με τα ονόματα **file_frequencies** και **total_frequencies** (θα είναι instance variables της κλάσης).
    - Μια μέθοδο **tokenize**(**text**) η οποία θα παίρνει ως όρισμα ένα string, θα το χωρίζει σε tokens και θα τα επιστρέφει σε μια λίστα. Μπορείτε εσείς να επιλέξετε τη μέθοδο με την οποία θα χωρίσετε το κείμενο σε λέξεις, είτε επιλέγοντας ένα απλό split είτε χρησιμοποιώντας κάτι το οποίο θα χωρίσει τις λέξεις με μεγαλύτερη ακρίβεια. (π.χ. regular expressions).
    - Μια μέθοδο **generate_frequencies**(**word_list**) η οποία θα παίρνει ως όρισμα μια λίστα με λέξεις και θα αποθηκεύει σε ένα dictionary τη συχνότητα των λέξεων. Στο τέλος, η μέθοδος θα επιστρέφει το λεξικό συχνοτήτων που δημιουργήθηκε.
    - Μια μέθοδο **read_folder**() η οποία θα διαβάζει όλα τα αρχεία με κατάληξη .txt που περιέχει ο φάκελος **source_folder** με τον οποίο έχει αρχικοποιηθεί το instance της κλάσης σας. Για κάθε αρχείο που περιέχει ο φάκελος, θα πρέπει να εκτελούνται οι παρακάτω ενέργειες:
        - Θα κανει tokenize το κείμενο χρησιμοποιώντας τη μέθοδο **tokenize** που έχετε δημιουργήσει
        - Θα παίρνει τη συχνότητα των tokens του αρχείου χρησιμοποιώντας τη μέθοδο **generate_frequencies** που ορίσατε παραπάνω, και θα την αποθηκεύει στο λεξικό της κλάσης **file_frequencies** με κλειδί το όνομα του αρχείου
        - Όταν η μέθοδος έχει διαβάσει όλα τα αρχεία που περιέχει ο φάκελος, θα πρέπει να ενώνει τα λεξικά κάθε αρχείου σε ένα συνολικό λεξικό συχνοτήτων (το οποίο θα αποθηκευτεί στην **instance variable total_frequencies**).
    - Μια μέθοδο **get_frequency**(**token**) η οποία θα παίρνει ως όρισμα ένα token και θα επιστρέφει από το **total_frequencies** τη συχνότητα εμφάνισης του token σε όλα τα αρχεία που έχετε διαβάσει. Αν καλέσετε την ίδια μέθοδο με δύο παραμέτρους **get_frequency**(**token**, **file_name**), με τη δεύτερη να είναι δηλαδή το όνομα ενός αρχείου, θα πρέπει να επιστρέφει τη συχνότητα εμφάνισης του token μόνο σε εκείνο το αρχείο. Αν το αρχείο που έδωσε ο χρήστης δεν υπάρχει θα πρέπει να επιστρέφετε ένα μήνυμα λάθους.
    - Μια μέθοδο **calculate_similarity**(**file_a**, **file_b**) η οποία θα πρέπει να υπολογίζει πόσες κοινές λέξεις έχουν τα λεξικά συχνοτήτων των δύο αρχείων και να επιστρέφει ένα ποσοστό ομοιότητας των δύο αρχείων. Το ποσοστό ομοιότητας θα το υπολογίσετε ως εξής:

        file_similarity = common_tokens/(file_a_tokens+file_b_tokens)

Για να τρέξετε τον κώδικά σας, θα χρειαστείτε ένα δεύτερο python module **final_project_test.py** μέσα στο οποίο θα δημιουργήσετε ένα instance την κλάσης **FrequenciesGenerator**, και θα τρέξετε τον κώδικά σας χρησιμοποιώντας το φάκελο που περιέχει το συνημμένο αρχείο **test_data_M901_final_project.zip**. Ο φάκελος περιέχει 140 text files με ελληνικά κείμενα τα οποία προέρχονται από το παρακάτω αποθετήριο: https://zenodo.org/communities/tramooc-h2020/?page=1&size=20

Η καταληκτική ημερομηνία υποβολής της εργασίας είναι η Δευτέρα 25 Ιανουαρίου στις 23:00. Η ημερομηνία της εξέτασης θα ανακοινωθεί μετά τις 20 Ιανουαρίου.

Είμαστε στη διάθεσή σας για οποιαδήποτε διευκρίνιση. Καλή επιτυχία!

Αρχείο:
test_data_M901_final_project.zip

Μέγιστη βαθμολογία:
10

Τύπος Βαθμολογίας:
Αριθμός

Ημερομηνία έναρξης:
31-12-2020 10:59:43

Προθεσμία υποβολής:
25-01-2021 23:00:00
(απομένουν 24 ημέρες 16 ώρες 10 λεπτά)

Τύπος εργασίας:
Ατομική εργασία