Ονοματεπώνυμο: Δημήτριος Κωστορρίζος ΑΜ: 1054419

Email: up1054419@upnet.gr Έτος: Γ'

Περιγραφή Κώδικα

Αρχικά γίνονται import:

Το module string, για να χρησιμοποιηθεί η string punctuation.

Η δομή εξειδικευμένου λεξικού Counter του module collections.

Οι συναρτήσεις sqrt και pow του module math.

Η συνάρτηση combinations του module itertools.

Δημιουργούνται οι λίστες word_maps_list, filenames_list, οι οποίες αποθηκεύουν τα ατομικά για κάθε αρχείο λεξικά Counter και Filename αντίστοιχα. Παράγεται η translate_tab, η οποία όταν χρησιμοποιηθεί ως όρισμα στην συνάρτηση translate, αφαιρεί όλα τα σημεία στίξης από την λέξη. Δημιουργείται το λεξικό similarities, το οποίο θα έχει ως key ένα tuple της μορφής (αριθμός αρχείου, αριθμός αρχείου) και value το ποσοστό ομοιότητας, στρογγυλοποιημένο στο πρώτο δεκαδικό ψηφίο. Έπειτα, εξασφαλίζεται ότι η τιμή του αριθμού των εγγράφων που θα εισάγει ο χρήστης είναι τουλάχιστον 2, προκειμένου να επιτευχθεί η σύγκριση, ενώ δημιουργούνται όλα τα δυνατά, χωρίς επικαλύψεις, tuple(ζευγάρια) αρχείων, τα οποία θα μπορούν να συγκριθούν. Κατά την εισαγωγή του filename από τον χρήστη, ελέγχεται αν το αρχείο υπάρχει και μπορεί να ανοίξει, τερματίζοντας το πρόγραμμα, σε περίπτωση σφάλματος. Για κάθε αρχείο, ακολουθούνται τα εξής βήματα:

1. Αποθηκεύεται το όνομά του στην αντίστοιχη λίστα.
2. Ανοίγεται το αρχείο, με δυνατότητα εγγραφής.
3. Δημιουργείται ο ατομικός του Counter, ο οποίος αποθηκεύεται στην αντίστοιχη λίστα.
4. Αναγνωρίζεται η κάθε λέξη, κάθε γραμμής του αρχείου, από την οποία αφαιρούνται τα σημεία στίξης και μετατρέπονται όλα τα γράμματα σε πεζά. Μετά την επεξεργασία αυτή η λέξη προστίθεται στον Counter, ενημερώνοντας το value της ήδη υπάρχουσας εγγραφής, αλλιώς δημιουργείται νέα εγγραφή για την λέξη με value 0.
5. Μετά την ολοκλήρωση τους διαβάσματος του αρχείου, το αρχείο κλείνει.

Με βάση τα tuple, με τους αριθμούς των αρχείων, ο κάθε συνδυασμός (αριθμός αρχείου, αριθμός αρχείου) χρησιμοποιείται για να ξεκινήσει η σύγκριση των αντίστοιχων αρχείων. Χρησιμοποιώντας την τεχνική του unpacking πάνω στο συγκεκριμένο tuple, τα δύο μέλη του ζευγαριού, χρησιμοποιούνται για την εύρεση των απαιτούμενων Counter και την επιστροφή του αριθμού των εμφανίσεων κάθε λέξης μέσα στο έγγραφο. Ο υπολογισμός γίνεται για κάθε

λέξη του συνόλου, που δημιουργείται από την ένωση των συνόλων των λέξεων κάθε αρχείου. Αν η λέξη υπάρχει στο αρχείο, επιστρέφεται ο αριθμός των εμφανίσεων της, διαφορετικά η τιμή 0. Υπολογίζεται η επί της εκατό τιμή του συντελεστή ομοιότητας , στρογγυλοποιημένη στο πρώτο δεκαδικό ψηφίο. Μετά τον υπολογισμό της, χρησιμοποιείται η τεχνική του packing πάνω στο δεδομένο tuple, προκειμένου να χρησιμοποιηθεί ως key στο λεξικό similarities, δημιουργώντας έτσι εγγραφές με την μορφή [(αριθμός αρχείου, αριθμός αρχείου), similarity].

Διασφαλίζεται ότι η τιμή της μεταβλητής K, ανήκει στο [0,(πλήθος συνδυασμών N ανά 2)] και ταξινομείται το λεξικό similarities κατά φθίνουσα, από τα αριστερά, τιμή του value similarity. Στο τέλος, εμφανίζεται το πλήθος με τους συνδυασμούς των αρχείων που κατείχαν τις K μεγαλύτερες τιμές του συντελεστή ομοιότητας, που έχει ζητήσει ο χρήστης.

Screenshots παραδειγμάτων εκτέλεσης του προγράμματος

Χρησιμοποιήθηκαν τα εξής 5 κείμενα, για τον έλεγχο του προγράμματος:

Test1.txt

```
Extensible Markup Language (XML) is a markup language that defines a set of
rules
for encoding documents in a format that is both human-readable and
machinereadable.
It is defined in the XML 1.0 Specification produced by the World Wide Web
Consortium (W3C), and several other related specifications, all gratis open
standards.
```

Test2.txt

```
XHTML (eXtensible HyperText Markup Language) is a family of XML markup
languages that mirror or extend versions of the widely-used Hypertext Markup
Language (HTML), the language in which web pages are written. XHTML 1.0 became
a World Wide Web Consortium (W3C) Recommendation on January 26, 2000, for
encoding documents in a format that is both human-readable and machine-readable.
```

Test3.txt

```
JSON is a language-independent data format. It was derived from JavaScript, but
as of 2017 many programming languages include code to generate and parse JSON-
format data. The official Internet media type for JSON is application/json. JSON
filenames use the extension .json.
```

Test4.txt

A file format is a standard way that information is encoded for storage in a computer file. It specifies how bits are used to encode information in a digital storage medium. File formats may be either proprietary or free and may be either unpublished or open.

Some file formats are designed for very particular types of data: PNG files, for example, store bitmapped images using lossless data compression. Other file formats, however, are designed for storage of several different types of data: the Ogg format can act as a container for different types of multimedia including any combination of audio and video, with or without text (such as subtitles), and metadata. A text file can contain any stream of characters, including possible control characters, and is encoded in one of various character encoding schemes. Some file formats, such as HTML, scalable vector graphics, and the source code of computer software are text files with defined syntaxes that allow them to be used for specific purposes.

Test5.txt

A text file (sometimes spelled textfile; an old alternative name is flatfile) is a kind of computer file that is structured as a sequence of lines of electronic text. A text file exists stored as data within a computer file system. In operating systems such as CP/M and MS-DOS, where the operating system does not keep track of the file size in bytes, the end of a text file is denoted by placing one or more special characters, known as an end-of-file marker, as padding after the last line in a text file. On modern operating systems such as Microsoft Windows and Unix-like systems, text files do not contain any special EOF character, because file systems on those operating systems keep track of the file size in bytes. There are for most text files a need to have end-of-line delimiters, which are done in a few different ways depending on operating system. Some operating systems with record-orientated file systems may not use new line delimiters and will primarily store text files with lines separated as fixed or variable length records.

Αποτέλεσμα εκτέλεσης του προγράμματος

```
Enter the number of documents.5
Enter the filename of the document, number 1 :
test1.txt
Enter the filename of the document, number 2 :
test2.txt
Enter the filename of the document, number 3 :
test3.txt
Enter the filename of the document, number 4 :
test4.txt
Enter the filename of the document, number 5 :
test5.txt
How many TOP-K most similar documents you want to print? Enter the K number, it
has to be between 0 and 10 .
10
TOP - 10 most similar documents:
File:  test1.txt     File:  test2.txt  Similarity =  71.3 %
File:  test4.txt     File:  test5.txt  Similarity =  59.8 %
File:  test1.txt     File:  test4.txt  Similarity =  38.7 %
File:  test2.txt     File:  test4.txt  Similarity =  36.9 %
File:  test1.txt     File:  test5.txt  Similarity =  33.5 %
File:  test2.txt     File:  test5.txt  Similarity =  33.1 %
File:  test3.txt     File:  test4.txt  Similarity =  30.3 %
File:  test1.txt     File:  test3.txt  Similarity =  28.2 %
File:  test3.txt     File:  test5.txt  Similarity =  24.5 %
File:  test2.txt     File:  test3.txt  Similarity =  23.1 %
```