Αρχές Γλωσσών Προγραμματισμού & Μεταφραστών
Προαιρετική Εργαστηριακή Άσκηση 2019

Τμήμα Μηχανικών Η/Υ & Πληροφορικής
Πανεπιστήμιο Πατρών
Εαρινό Εξάμηνο 2019
Διδάσκοντες: Ι. Γαροφαλάκης, Σ. Σιούτας

Ονοματεπώνυμο	Καλαματιανού Δήμητρα
Α.Μ.	1054406

Κώδικας σε Python για την εργαστηριακή άσκηση:

Παρακάτω παρατίθεται ο κώδικας με τον απαραίτητο σχολιασμό, όπου κρίθηκε αναγκαίο, τα σχόλια βρίσκονται δίπλα στις γραμμές κώδικα μετά την δίεση(‘#’) με **έντονο μαύρο** χρώμα.

```
import glob
import math
from operator import itemgetter

print('Hello, please enter the documents in the same file directory as the program and run again the program \n') #το πρόγραμμα ζητά από τον χρήστη να τοποθετήσει στον ίδιο φάκελο με αυτό
                # τα αρχεία κειμένου(.txt) και αφού το κάνει να ξανατρέξει το πρόγραμμα

big_array = []      # λίστα που περιέχει όλες τις λέξεις από όλα τα αρχεία, αφορά όλα τα
                    # αρχεία

array = []          # λίστα που περιέχει λέξεις και αφορά κάθε αρχείο ξεχωριστά

for filename in glob.glob('*.txt'): #ανακάλυψη κάθε αρχείου κειμένου στον φάκελο που βρίσκεται
                                    # το πρόγραμμα

   with open(filename, 'r') as f:

      for line in open (filename, 'r'):  # for loop για την δημιουργία της λίστας με όλες τις λέξεις
         for word in line.split():
            if word not in big_array:
               big_array.append(word)
big_array.sort()  # ταξινόμηση λίστας με αλφαβητική σειρά

namelist = glob.glob('*.txt') # λίστα με τα ονόματα των αρχείων .txt που βρίσκονται στον ίδιο
                              # φάκελο με το πρόγραμμα
```

```
print('Are these the documents? ', namelist, '\nyes or no?') # ρωτά τον χρήστη αν τα κείμενα είναι
                                                             # αυτά που έβαλε μέσα στον φάκελο και
                                                             # τυπώνει την λίστα με τα ονόματα του
                                                             # φακέλου
apantisi = str(input())

while apantisi != 'yes': # αν η απάντηση είναι διαφορετική από 'yes' τυπώνει μήνυμα να
                         # επαναλάβει την διαδικασία και τερματίζει το πρόγραμμα
```

ΠΑΡΑΔΕΙΓΜΑ 1

```
  print('Try again')
   exit()
```

```
try:
   max = int(input('Give me the K number for the TOP-K most similar documents,please: \n'))
except ValueError:
   print("Not a number") # σε αυτό το κομμάτι κώδικα το πρόγραμμα ζητά στον χρήστη να
   exit()                # δώσει το K για τα TOP-K ζευγάρια κειμένων με μεγαλύτερη
                         # ομοιότητα, αν δεν δώσει αριθμό τυπώνει το αντίστοιχο μήνυμα και
                         # τερματίζει το πρόγραμμα
```

ΠΑΡΑΔΕΙΓΜΑ 2

```
while (max > math.factorial(len(namelist))/(2*math.factorial(len(namelist)-2))) or (max == 0):
   max = int(input('wrong,try again:')) # σε αυτό το κομμάτι κώδικα ελέγχεται αν το K είναι >0 και
                                        # αν υπάρχουν αρκετοί συνδυασμοί κειμένων που να είναι
                                        # περισσότεροι από το K που έδωσε ο χρήστης, αν δεν
                                        # είναι τυπώνει αντίστοιχο μήνυμα και ζητά από τον
                                        # χρήστη να ξαναπροσπαθήσει, περιμένει μέχρι να δοθεί
                                        # σωστό K


globlist = [] # λίστα που περιέχει τις λίστες με τις συχνότητες των λέξεων στα κείμενα

for filename in namelist:
   with open(filename, 'r') as f:

      wordcount = {} # είναι dictionary που αποθηκεύει την συχνότητα εμφάνισης της κάθε
                     # λέξεις σε κάθε κείμενο με key την λέξη  και value αριθμό εμφάνισης της,
                     # πχ 'a': 3, δηλαδή η λέξη 'a' εμφανίζεται 3 φορές

      for line in open(filename, 'r'):
         for word in line.split():
            for i in big_array:
```

```
            if i not in array: # δημιουργία επιμέρους λιστών με λέξεις για κάθε αρχείο .txt
                array.append(i)
            if i not in wordcount: # όσες λέξεις δεν υπάρχουν σε κάποιο αρχείο αλλά υπάρχουν
                                   # στην λίστα με όλες τις λέξεις τις βάζω με συχνότητα 0
                wordcount[i] = 0

        if word not in wordcount: # εδώ μετριέται η συχνότητα εμφάνισης κάθε λέξης σε κάθε
                                  # αρχείο ξεχωριστά
            wordcount[word] = 1
        else:
            wordcount[word] += 1

    numbers = []  # λίστα που περιέχει μόνο τις συχνότητες των λέξεων σε κάθε κείμενο
    for key in wordcount:
        numbers.append(wordcount[key])
    globlist.append([numbers[i]  for i in range(len(numbers))])  # λίστα με λίστες των συχνοτήτων
                                                                 # των λέξεων

local_1 = [] # λίστα για να κρατηθεί προσωρινά μία εκ των δύο λιστών με συχνότητες που
             # συγκρίνονται κάθε φορά
local_2 = [] # λίστα για να κρατηθεί προσωρινά η δεύτερη εκ των δύο λιστών με συχνότητες
             # που συγκρίνονται κάθε φορά

dictionary = {} # dictionary με  key το ζευγάρι των κειμένων και value την ομοιότητα
                # συνημίτονου αυτών των κειμένων

for k in range(0,len(globlist)-1): # σε αυτό το κομμάτι κώδικα παίρνω από την λίστα με τις λίστες
    for p in range(k+1,len(globlist)): # των συχνοτήτων κάθε μία λίστα με τις υπόλοιπες ώστε να
                                       # υπολογιστεί το μήκος κάθε διανύσματος και το εσωτερικό
                                       # γινόμενο των δύο κειμένων που εξετάζονται κάθε φορά

        local_1 = globlist[k]      # λίστα για να κρατηθεί προσωρινά μία εκ των δύο λιστών
        local_2 = globlist[p]      # λίστα για να κρατηθεί προσωρινά μία εκ των δύο λιστών

        norm_1=0
        for a in range (len(local_1)): # εδώ υπολογίζεται το μήκος διανύσματος του πρώτου εκ των
            norm_1 = norm_1+(local_1[a]*local_1[a]) # δύο διανυσμάτων
        norm_1 = norm_1**0.5

        norm_2 = 0
        for a in range(len(local_2)): # εδώ υπολογίζεται το μήκος διανύσματος του δεύτερου εκ των
            norm_2 = norm_2 + (local_2[a] * local_2[a]) # δύο διανυσμάτων
        norm_2 = norm_2 ** 0.5

        esot = sum([a * b for a, b in zip(local_1, local_2)]) # υπολογισμός εσωτερικού γινομένου των
                                                              # δύο κειμένων

        cos = esot/(norm_1*norm_2) # υπολογισμός ομοιότητας συνημίτονου
        dictionary[str(k)+","+str(p)] =cos # εισαγωγή στο key του dictionary των νούμερων των
                                           # κειμένων που εξετάζονται κάθε φορά δηλαδή στην πρώτη
                                           # επανάληψη θα περαστεί στο dictionary ως key το '0,1'
                                           # αφού θα εξεταστούν οι λίστες συχνοτήτων 0 και 1, στην
                                           # επόμενη επανάληψη θα περαστεί το '0,2' κ.ο.κ. με τις
```

```
                                                    # αντίστοιχες  ομοιότητες συνημίτονου τους ως value

dictionary = sorted(dictionary.items(), key=itemgetter(1), reverse=True) # ταξινόμηση dictionary
                                                                          # με βάση τα values, γίνεται
                                                                          # λίστα με tuples, αν για
                                                                          # παράδειγμα ζητούσα
                                                                         # να το τυπώσει θα τύπωνε
                                                                         # ('0,1', 0.5345224838248487)
                                                                        # δηλαδή ζευγάρι κειμένων και
                                                                        # ομοιότητα συνημίτονου


cnt = 1
temp = []   # λίστα για να περαστούν τα νούμερα των κειμένων για να γίνει
            # αντιστοίχιση με την λίστα των ονομάτων των κειμένων
for tup in dictionary:
   if cnt <= max : # cnt counter που ελέγχει αν έχουν τυπωθεί όλα τα TOP-K most similar
                   # documents ζήτησε ο χρήστης

      temp = tup[0].split(',') # πέρασμα στην λίστα από τα tuples του tup[0] που περιέχει το
                               # ζευγάρι των texts files, χωρίζοντας το όπου βρει ',', δηλαδή στην
                               # πρώτη θέση της λίστας temp μπαίνει το νούμερο του πρώτου
                               # κειμένου και στην δεύτερη το νούμερο του δεύτερου κειμένου

      print('The text files are',namelist[int(temp[0])], 'and', namelist[int(temp[1])], '\nwith cos: ',
round(tup[1],4), ' or ', round(tup[1]*100,2), '% \n')
 # η λίστα με τα ονόματα των αρχείων επιστρέφει το όνομα του κάθε αρχείου για το
# temp[0]--> νούμερο του πρώτου κειμένου temp[1]--> νούμερο του πρώτου κειμένου
# γίνεται αντιστοιχία του νούμερου του κειμένου με το όνομά του

   cnt = cnt + 1

f.close()
```

ΠΑΡΑΔΕΙΓΜΑ 1

A) Απάντηση no από τον χρήστη ≠ yes

```
Hello, please enter the documents in the same file directory as the program and run again the program

Are these the documents?  ['Example_1.txt', 'Example_2.txt']
yes or no?
no
Try again

Process finished with exit code 0
```

B) Απάντηση 5 από τον χρήστη ≠ yes

```
Hello, please enter the documents in the same file directory as the program and run again the program

Are these the documents?  ['Example_1.txt', 'Example_2.txt']
yes or no?
5
Try again

Process finished with exit code 0
```

ΠΑΡΑΔΕΙΓΜΑ 2

A) Απάντηση notanumber από τον χρήστη ≠ αριθμός

```
Hello, please enter the documents in the same file directory as the program and run again the program

Are these the documents?  ['Example_1.txt', 'Example_2.txt']
yes or no?
yes
Give me the K number for the TOP-K most similar documents,please:
notanumber
Not a number

Process finished with exit code 0
```

B) Απάντηση από τον χρήστη > επιτρεπτού αριθμού

```
Hello, please enter the documents in the same file directory as the program and run again the program

Are these the documents?  ['Example_1.txt', 'Example_2.txt']
yes or no?
yes
Give me the K number for the TOP-K most similar documents,please:
3
wrong,try again:4
wrong,try again:5
wrong,try again:6
wrong,try again:7
wrong,try again:1
The text files are Example_1.txt and Example_2.txt
with cos:  1.0  or  100.0 %


Process finished with exit code 0
```

Παραδείγματα εφαρμογής:

- ✓ Για τα Document 1(XML).txt και Document 2(XHTML).txt από την εκφώνηση της άσκησης με

Document 1(XML)
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 machine-readable. 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.

Document 2(XHTML)
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.

Το αποτέλεσμα που επιστρέφει το πρόγραμμα είναι:

```
Hello, please enter the documents in the same file directory as the program and run again the program

Are these the documents?  ['Document 1(XML).txt', 'Document 2(XHTML).txt']
yes or no?
yes
Give me the K number for the TOP-K most similar documents,please:
1
The text files are Document 1(XML).txt and Document 2(XHTML).txt
with cos:  0.6584  or  65.84 %


Process finished with exit code 0
```

✓ Για τα Example_1.txt και Example_2.txt με

Example_1
Hello my name is Dimitra

Example_2
Hello my name is Dimitra

Το αποτέλεσμα που επιστρέφει το πρόγραμμα είναι:

```
Hello, please enter the documents in the same file directory as the program and run again the program

Are these the documents?  ['Example_1.txt', 'Example_2.txt']
yes or no?
yes
Give me the K number for the TOP-K most similar documents,please:
1
The text files are Example_1.txt and Example_2.txt
with cos:  1.0  or  100.0 %


Process finished with exit code 0
```

✓ Για τα Example_3.txt και Example_4.txt με

Example_3
a a a b b c c c c c d d

Example_4
a d e e

Το αποτέλεσμα που επιστρέφει το πρόγραμμα είναι:

```
Hello, please enter the documents in the same file directory as the program and run again the program

Are these the documents?  ['Example_3.txt', 'Example_4.txt']
yes or no?
yes
Give me the K number for the TOP-K most similar documents,please:
1
The text files are Example_3.txt and Example_4.txt
with cos:  0.315  or  31.5 %


Process finished with exit code 0
```

✓ Για τα Example_5.txt και Example_6.txt

Example_5
DevOps engineers have both management and computer programming experience. They work as part of a team to streamline the process of creating (development) and using computer software (operations) in an online environment where the website is always functional. Businesses that sell products online using software such as image recognition require a DevOps engineer to automate services on a cloud-based platform. They build ways to deliver products that are continuously integrated within the specific architecture of a company's website. Skills needed for this job include project management as well as systems analysis and computer programming.

Example_6
Software engineers may specialize in a particular application such as retail, banking, transportation or artificial intelligence. Regardless of the software's focus, the process a software engineer uses involves consultation with the client to determine their needs. Once the type of functions are established, the software engineer designs a program for the client and works with computer programmers who write the code for the software. Finally, the software engineer evaluates how the client uses their software to ensure it is not too cumbersome.

Το αποτέλεσμα που επιστρέφει το πρόγραμμα είναι:

```
Hello, please enter the documents in the same file directory as the program and run again the program

Are these the documents?  ['Example_5.txt', 'Example_6.txt']
yes or no?
yes
Give me the K number for the TOP-K most similar documents,please:
1
The text files are Example_5.txt and Example_6.txt
with cos:  0.4441  or  44.41 %


Process finished with exit code 0
```

✓ Για τα Example_1.txt, Example_2.txt, Example_7.txt και Example_8.txt με

Example_1
Hello my name is Dimitra

Example_2
Hello my name is Dimitra

Example_7
Hello I do not have a name

Example_8
Hello I can not tell you my name

Το αποτέλεσμα που επιστρέφει το πρόγραμμα είναι:

Για K=1

```
Hello, please enter the documents in the same file directory as the program and run again the program

Are these the documents?  ['Example_1.txt', 'Example_2.txt', 'Example_7.txt', 'Example_8.txt']
yes or no?
yes
Give me the K number for the TOP-K most similar documents,please:
1
The text files are Example_1.txt and Example_2.txt
with cos:  1.0  or  100.0 %


Process finished with exit code 0
```

Το πρόγραμμα εμφανίζει τα κείμενα Example_1.txt και Example_2.txt μιας και είναι ακριβώς τα ίδια έχουν ομοιότητα συνημιτόνου 100%.

Για K=2

```
Hello, please enter the documents in the same file directory as the program and run again the program

Are these the documents?  ['Example_1.txt', 'Example_2.txt', 'Example_7.txt', 'Example_8.txt']
yes or no?
yes
Give me the K number for the TOP-K most similar documents,please:
2
The text files are Example_1.txt and Example_2.txt
with cos:  1.0  or  100.0 %

The text files are Example_7.txt and Example_8.txt
with cos:  0.5345  or  53.45 %


Process finished with exit code 0
```

Το πρόγραμμα εμφανίζει τα κείμενα Example_1.txt και Example_2.txt με ομοιότητα συνημιτόνου 100% και τα κείμενα Example_7.txt και Example_8.txt ως δεύτερο συνδυασμό κειμένων με την δεύερη μεγαλύτερη ομοιότητα συνημιτόνου. Έτσι συνεχίζει η εκτέλεση του προγράμματος όσο αυξάνει το K ο χρήστης.

Για K=3

```
Hello, please enter the documents in the same file directory as the program and run again the program

Are these the documents?  ['Example_1.txt', 'Example_2.txt', 'Example_7.txt', 'Example_8.txt']
yes or no?
yes
Give me the K number for the TOP-K most similar documents,please:
3
The text files are Example_1.txt and Example_2.txt
with cos:  1.0  or  100.0 %

The text files are Example_7.txt and Example_8.txt
with cos:  0.5345  or  53.45 %

The text files are Example_1.txt and Example_8.txt
with cos:  0.4743  or  47.43 %


Process finished with exit code 0
```

Για K=4

```
Hello, please enter the documents in the same file directory as the program and run again the program

Are these the documents?  ['Example_1.txt', 'Example_2.txt', 'Example_7.txt', 'Example_8.txt']
yes or no?
yes
Give me the K number for the TOP-K most similar documents,please:
4
The text files are Example_1.txt and Example_2.txt
with cos:  1.0  or  100.0 %

The text files are Example_7.txt and Example_8.txt
with cos:  0.5345  or  53.45 %

The text files are Example_1.txt and Example_8.txt
with cos:  0.4743  or  47.43 %

The text files are Example_2.txt and Example_8.txt
with cos:  0.4743  or  47.43 %


Process finished with exit code 0
```

Για K=5

```
Hello, please enter the documents in the same file directory as the program and run again the program

Are these the documents?  ['Example_1.txt', 'Example_2.txt', 'Example_7.txt', 'Example_8.txt']
yes or no?
yes
Give me the K number for the TOP-K most similar documents,please:
5
The text files are Example_1.txt and Example_2.txt
with cos:  1.0  or  100.0 %

The text files are Example_7.txt and Example_8.txt
with cos:  0.5345  or  53.45 %

The text files are Example_1.txt and Example_8.txt
with cos:  0.4743  or  47.43 %

The text files are Example_2.txt and Example_8.txt
with cos:  0.4743  or  47.43 %

The text files are Example_1.txt and Example_7.txt
with cos:  0.3381  or  33.81 %


Process finished with exit code 0
```

Για K=6

```
Hello, please enter the documents in the same file directory as the program and run again the program

Are these the documents?  ['Example_1.txt', 'Example_2.txt', 'Example_7.txt', 'Example_8.txt']
yes or no?
yes
Give me the K number for the TOP-K most similar documents,please:
6
The text files are Example_1.txt and Example_2.txt
with cos:  1.0  or  100.0 %

The text files are Example_7.txt and Example_8.txt
with cos:  0.5345  or  53.45 %

The text files are Example_1.txt and Example_8.txt
with cos:  0.4743  or  47.43 %

The text files are Example_2.txt and Example_8.txt
with cos:  0.4743  or  47.43 %

The text files are Example_1.txt and Example_7.txt
with cos:  0.3381  or  33.81 %

The text files are Example_2.txt and Example_7.txt
with cos:  0.3381  or  33.81 %


Process finished with exit code 0
```

Για K>7

```
Hello, please enter the documents in the same file directory as the program and run again the program

Are these the documents?  ['Example_1.txt', 'Example_2.txt', 'Example_7.txt', 'Example_8.txt']
yes or no?
yes
Give me the K number for the TOP-K most similar documents,please:
7
wrong,try again:8
wrong,try again:9
wrong,try again:10
wrong,try again:11
wrong,try again:12
wrong,try again:
```

Το πρόγραμμα δεν δέχεται Κ μεγαλύερο από τους δυνατούς συνδυασμούς των κειμένων. Σε αυτή την περίπτωση τα κείμενα είναι 4, οι συνδυασμοί είναι (4!/(2*(4-2)!)=6 οπότε σωστά το πρόγραμμα ζητά από τον χρήστη έναν άλλο αριθμό για να υπολογίσει τα TOP-K όμοια κείμενα.