Cosine Similarity Project

Ονοματεπώνυμο: Αντόν Πάπα
Email: papa@ceid.upatras.gr
ΑΜ: 1054337
Έτος: Γ'

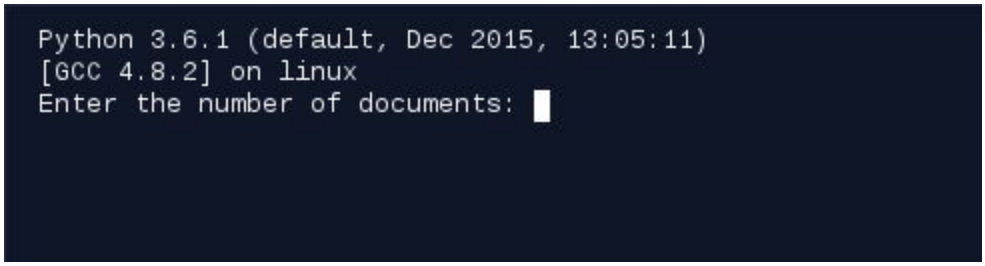
Το παρακάτω πρόγραμμα έχει την ικανότητα να συγκρίνει δύο αρχεία και να δείξει κατά πόσο είναι ίδια, με βάση τα όσα γράφονται σε αυτά. Αυτό γίνεται με την βοήθεια του cosine similarity, το οποίο ορίζεται ως το εσωτερικό γινόμενο δύο διανυσμάτων δια το γινόμενο των αναμεταξύ τους μέτρων.

*(cosine similarity = A*B/||A||*||B||), όπου τα A, B διανύσματα.*

Μερικά screenshots εφαρμογής του παρακάτω προγράμματος:

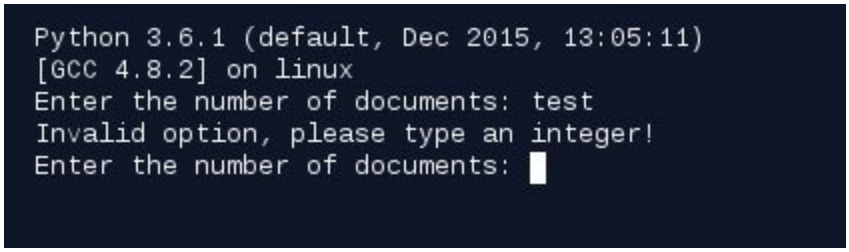
1.

Το πρόγραμμα ζητάει από τον χρήστη να δώσει τον αριθμό των αρχείων.



2.

Ο χρήστης πληκτρολογεί μη επιτρεπόμενο χαρακτήρα.



3.

Ο χρήστης πληκτρολογεί αρχείο που δεν μπορεί να εντοπιστεί ή δεν υπάρχει.

```
Python 3.6.1 (default, Dec 2015, 13:05:11)
[GCC 4.8.2] on linux
Enter the number of documents: test
Invalid option, please type an integer!
Enter the number of documents: 4
Enter the name (including file type) of document #1: something inv
alid
Invalid option, file not found, try again!
Enter the name (including file type) of document #1:
```

4.

Ο χρήστης πληκτρολογεί σωστά τα πάντα και το πρόγραμμα ζητάει τον αριθμό k.

```
Python 3.6.1 (default, Dec 2015, 13:05:11)
[GCC 4.8.2] on linux
Enter the number of documents: 2
Enter the name (including file type) of document #1: test.txt
Enter the name (including file type) of document #2: takest.txt
Enter the number(k) of most similar documents, you want to view:
```

5.

Ο χρήστης πληκτρολογεί αριθμό μικρότερο του 1, οπότε το πρόγραμμα παίρνει k=1.

```
Python 3.6.1 (default, Dec 2015, 13:05:11)
[GCC 4.8.2] on linux
Enter the number of documents: 2
Enter the name (including file type) of document #1: test.txt
Enter the name (including file type) of document #2: takest.txt
Enter the number(k) of most similar documents, you want to view: -21
1. <test.txt> <takest.txt> : 0.6923076923076924
>
```

6.

Ο χρήστης πληκτρολογεί τεράστιο αριθμό, οπότε το πρόγραμμα παίρνει k=len(dictionary).

```
Python 3.6.1 (default, Dec 2015, 13:05:11)
[GCC 4.8.2] on linux
Enter the number of documents: 3
Enter the name (including file type) of document #1: takis.txt
Enter the name (including file type) of document #2: test.txt
Enter the name (including file type) of document #3: takest.txt
Enter the number(k) of most similar documents, you want to view: 100
1. <takis.txt> <test.txt> : 0.7205766921228922
2. <takis.txt> <takest.txt> : 0.7205766921228922
3. <test.txt> <takest.txt> : 0.6923076923076924
>
```

7.

Ο χρήστης πληκτρολογεί τα πάντα σωστά και δίνει για την τιμή του k, k=2.

```
Python 3.6.1 (default, Dec 2015, 13:05:11)
[GCC 4.8.2] on linux
Enter the number of documents: 3
Enter the name (including file type) of document #1: takis.txt
Enter the name (including file type) of document #2: test.txt
Enter the name (including file type) of document #3: takest.txt
Enter the number(k) of most similar documents, you want to view: 2
1. <takis.txt> <test.txt> : 0.7205766921228922
2. <takis.txt> <takest.txt> : 0.7205766921228922
>
```

Κώδικας

```
#Εισαγωγή της βιβλιοθήκης numpy
import numpy as np

#Δήλωση κάποιων list/dictionary
documents_name = []
discrete_words = []
documents_dictionary = {}
common_words = {}
counter_list = []

#Δημιουργία συνάρτησης για τον υπολογισμό
#της cosine similarity, ανάμεσα σε δύο vectors
def cosine_similarity(document1, document2):
 norm_doc1 = np.linalg.norm(document1)
 norm_doc2 = np.linalg.norm(document2)
 dot_product = np.dot(document1, document2)
 vector_length_product = norm_doc1 * norm_doc2
 cos = dot_product / vector_length_product
 return cos

#Ο χρήστης εισάγει τον αριθμό των αρχείων
#που θέλει να συγκρίνει και υπάρχει ένα Exception
#το οποίο πετάει μήνυμα λάθους, αν ο χρήστης εισάγει
#κάτι άλλο εκτός από ακέραιο
x = 0
while x!=1:
 try:
   number_of_documents = int(input("Enter the number of documents: "))
   if number_of_documents < 2:
     number_of_documents = 2
   x = 1
 except ValueError:
   print("Invalid option, please type an integer!")
```

```
#Ζητείται από τον χρήστη να εισάγει τα ονόματα
#των αρχείων που θέλει να συγκρίνει και κάθε αρχείο
#εισάγεται σε μία λίστα η οπόια εισάγεται σε ένα dictionary
#ως key και έχει ως value το κείμενο που περιέχει
#Υπάρχει και εδώ Exception σε περίπτωση που ο χρήστης εισάγει
#αρχείο που δεν υπάρχει στον ίδιο φάκελο με τον φάκελο που
#βρίσκεται το πρόγραμμα
for i in range(number_of_documents):
 x = 0
 while x!=1:
   try:
     temp = input("Enter the name (including file type) of document #" +
str(i+1) + ": ")
     current_document = open(temp, "r")
     documents_name.append(temp)
     document_content = current_document.read().lower().split()
     for word in document_content:
       if word not in discrete_words:
         discrete_words.append(word)
     documents_dictionary.update({documents_name[i]: document_content})
     current_document.close()
     x = 1
   except FileNotFoundError:
     print("Invalid option, file not found, try again!")

#Κάνοντας ένα iteration στο documents_dictionary
#προστίθεται σε ένα άλλο dictionary ως value, μία λίστα με αριθμούς
#οι οποίοι δείχνουν πόσες φορές εμφανίζεται η κάθε λέξη στο
#συγκεκριμένο document.
for document in documents_dictionary:
 for word in discrete_words:
   counter_list.append(documents_dictionary[document].count(word))
 common_words.update({document: counter_list})
 counter_list = []
```

```
#Γίνεται έλεγχος κάθε πιθανού συνδυασμού χρησιμοποιώντας
#την συνάρτηση cosine_similarity και το αποτέλεσμα
#αποθηκεύεται ως value σε ένα καινούργιο dictionary
cosine_result_dictionary = {}
for doc1 in range(len(documents_dictionary)):
 for doc2 in range(doc1+1, len(documents_dictionary)):
   cosine_result =
cosine_similarity(common_words.get(documents_name[doc1]),
common_words.get(documents_name[doc2]))
   temp = ". <" + str(documents_name[doc1])+ "> <" +
str(documents_name[doc2]) + "> : "
   cosine_result_dictionary[temp] = cosine_result

#Τέλος ζητείται από τον χρήστη να επιλέξει
#πόσους συνδυασμούς αρχείων θέλει να προβληθούν.
#Αν ο χρήστης επιλέξει μη έγκυρο αριθμό,
#αυτόματα γίνεται προσαρμογή του k, ανάλογα
#με το αν αυτό είναι πολύ μεγάλο ή πολύ μικρό
#Οι συνδυασμοί τυπώνονται σε φθίνουσα σειρά.
max_k = len(cosine_result_dictionary)
k = int(input("Enter the number(k) of most similar documents, you want to
view: "))
if k < 1:
 k = 1
if k > max_k:
 k = max_k
for i in range(k):
 key = sorted(cosine_result_dictionary, key=cosine_result_dictionary.get,
reverse=True)
 value = cosine_result_dictionary.get(key[i])
 print(str(i+1)+key[i]+str(value))
```