Αρχές Γλωσσών Προγραμματισμού & Μεταφραστών

Τμήμα Μηχανικών Η/Υ & Πληροφορικής

Πανεπιστήμιο Πατρών

Εαρινό Εξάμηνο 2019

Διδάσκοντες: Ι. Γαροφαλάκης, Σ. Σιούτας

Προαιρετική Εργαστηριακή Άσκηση

Στα **Δεδομένα Εγγράφων**, κάθε έγγραφο μετατρέπεται σε ένα **διάνυσμα όρων**. Κάθε όρος είναι ένα συστατικό (**ιδιότητα**) του διανύσματος. Η **τιμή** κάθε συστατικού είναι ο **αριθμός εμφανίσεων** του όρου στο έγγραφο, π.χ.:

ΠΑΡΑΔΕΙΓΜΑ
D1: "Shipment of gold damaged in a fire"
D2: "Delivery of silver arrived in a silver truck"
D3: "Shipment of gold arrived in a truck

Doc ID	a	arrived	damaged	delivery	fire	gold	in	of	shipment	silver	truck
D1	1	0	1	0	1	1	1	1	1	0	0
D2	1	0	0	1	0	0	1	1	0	2	1
D3	1	1	0	0	0	1	1	1	1	0	1

Γενικά, η ομοιότητα μεταξύ δύο αντικειμένων (δεδομένα εγγράφων στην προκειμένη περίπτωση) ορίζεται ως η αριθμητική μέτρηση του πόσο μοιάζουν δυο αντικείμενα. Μεγαλώνει όταν τα αντικείμενα μοιάζουν. Συχνά εμπίπτει στο όριο [0,1].

Αντίστοιχα, η ανομοιότητα μεταξύ δύο αντικειμένων ορίζεται ως η αριθμητική μέτρηση του πόσο διαφορετικά είναι δυο αντικείμενα. Μειώνεται όταν τα αντικείμενα μοιάζουν. Η ελάχιστη ανομοιότητα είναι συχνά 0. Το ανώτατο όριο διαφέρει από περίπτωση σε περίπτωση.

Η **Ομοιότητα συνημίτονου για δεδομένα εγγράφων** ορίζεται ως εξής**:**

- Αν d_1 και d_2 είναι δυο διανύσματα εγγράφων τότε:

  $\cos(d_1, d_2) = (d_1 \bullet d_2) / ||d_1|| \, ||d_2||$,

όπου • σημαίνει εσωτερικό γινόμενο και $|| d ||$ το μήκος του διανύσματος d.

- Παράδειγμα:

  d_1 **= 3 2 0 5 0 0 0 2 0 0**

  d_2 **= 1 0 0 0 0 0 0 1 0 2**

$d_1 \bullet d_2 = 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5$

$||d_1|| = (3*3+2*2+0*0+5*5+0*0+0*0+0*0+0*0+2*2+0*0+0*0)^{0.5} = (42)^{0.5} = 6.481$

$||d_2|| = (1*1+0*0+0*0+0*0+0*0+0*0+0*0+0*0+1*1+0*0+2*2)^{0.5} = (6)^{0.5} = 2.245$

$\cos(d_1,d_2) = 0.3150$ ή 31,5%.

Ζητούμενα

Να υλοποιήσετε σε python πρόγραμμα που να δέχεται ως είσοδο N documents (το N το ορίζει ο χρήστης) και να υπολογίζει τα ποσοστά ομοιότητας των παραπάνω αρχείων. Τέλος, να εμφανίζει τα **TOP-K most similar documents**, όπου K το ορίζει ο χρήστης και προφανώς $K < \binom{N}{2}$.

Υπόδειξη: Χρησιμοποιείστε μικρά κείμενα με αρκετά keywords ομοιότητας, όπως τα παρακάτω Wikipedia Documents:

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.

Παραδοτέα

- **Γραπτή Αναφορά** που περιλαμβάνει
  - Τον κώδικα σε python εμπλουτισμένο με σχόλια
  - Screenshots παραδειγμάτων εφαρμογής
- Συμπιεσμένα σε **ένα αρχείο (zip)** τα:
  - Την πιο πάνω γραπτή αναφορά
  - Τον **ΤΕΛΙΚΟ** κώδικα σε python.

Το αρχείο zip** πρέπει να έχει όνομα τον αριθμό μητρώου του φοιτητή (π.χ. **3972.zip**), και **να ανεβεί (ΥΠΟΧΡΕΩΤΙΚΑ) στο e-class. Σε ξεχωριστό αρχείο word μέσα στο zip να αναφέρεται το ονοματεπώνυμο, το έτος, ο αριθμός μητρώου και η e-mail διεύθυνση του φοιτητή.

Διευκρινήσεις

- Η άσκηση είναι **ατομική**.
- Η άσκηση είναι **προαιρετική με bonus έως 1,5 μονάδα στον τελικό βαθμό, εφόσον ο βαθμός της είναι >=5.**
- Τελικές ημερομηνίες παράδοσης είναι οι ημερομηνίες γραπτών εξετάσεων περιόδου **Ιουνίου ΜΟΝΟ**!
- Η άσκηση **ΔΕΝ ΚΡΑΤΙΕΤΑΙ** για το επόμενο ΑΚΑΔΗΜΑΙΚΟ ΕΤΟΣ!
- Για τυχόν απορίες ή υποδείξεις μπορείτε να απευθύνεστε με e-mail στον Καθηγητή κ. Σ. Σιούτα (sioutas@ceid.upatras.gr)