M902
Βασικές Μαθηματικές Έννοιες
στη
Γλωσσική Τεχνολογία

Project 3

Κυλάφη Χριστίνα-Θεανώ

LT1200012

November, 2020

TABLE OF CONTENTS

Question 1 ____ 3

Question 2 ____ 4

Question 3 ____ 5

Question 4 ____ 6

Question 5 ____ 7

Question 6 ____ 8

Question 7 ____ 9

Question 8 ____ 10

Question 9 ____ 11

Question 10 ____ 12

Question **1**

Question **2**

Question 3

Question **4**

Question **5**

Question **6**

Question **7**

Question **8**

Question **9**

Question **10**

In order to find the similar documents based on the 3-dimensional vector representations, $d_n \in \mathbb{R}^3,\ n = 1, \ldots, 5$, a 5x5 matrix was constructed, with each row being each of the 3-D vectors d_1, d_2, d_3, d_4, d_5, normalised as follows:

$$\sum_{i=1}^{5} \sum_{j=1}^{3} \frac{d_{ij}}{\|d_i\|}$$

where d_{ij} the $j-th$ element of the $i-th$ vector and $\|d_i\| = \sqrt{d_{i1}^2 + d_{i2}^2 + d_{i3}^2}$ the norm of the respective vector i:

$$D = \begin{bmatrix} d_1 \\ d_2 \\ d_3 \\ d_4 \\ d_5 \end{bmatrix} \implies D_{5x3} = \begin{bmatrix} 8 & 6 & 0 \\ 0 & 6 & 8 \\ 6 & 0 & 8 \\ 2 & 3 & 0 \\ 9 & 6 & 0 \end{bmatrix} \implies D_{normal} = \begin{bmatrix} \frac{8}{10} & \frac{6}{10} & \frac{0}{10} \\ \frac{0}{10} & \frac{6}{10} & \frac{8}{10} \\ \frac{6}{10} & \frac{0}{10} & \frac{8}{10} \\ \frac{2}{3.6055} & \frac{3}{3.6055} & \frac{0}{3.6055} \\ \frac{9}{10.8166} & \frac{6}{10.8166} & \frac{0}{10.8166} \end{bmatrix} = \begin{bmatrix} 0.8 & 0.6 & 0 \\ 0 & 0.6 & 0.8 \\ 0.6 & 0 & 0.8 \\ 0.5547 & 0.8320 & 0 \\ 0.8320 & 0.5547 & 0 \end{bmatrix}$$

Finally, we multiply **normalized** matrix **D** with its **transpose**, to get the inner product of every pair of vectors and perform the comparisons between the document representations:

$$D_{normal} \cdot D_{normal}^T = \begin{bmatrix} 0.8 & 0.6 & 0 \\ 0 & 0.6 & 0.8 \\ 0.6 & 0 & 0.8 \\ 0.5547 & 0.8320 & 0 \\ 0.8320 & 0.5547 & 0 \end{bmatrix} \cdot \begin{bmatrix} 0.8 & 0 & 0.6 & 0.5547 & 0.8320 \\ 0.6 & 0.6 & 0 & 0.8320 & 0.5547 \\ 0 & 0.8 & 0.8 & 0 & 0 \end{bmatrix} =$$

$$\begin{bmatrix} 1 & 0.36 & 0.48 & 0.9429 & 0.9984 \\ \mathbf{0.36} & 1 & 0.64 & 0.4992 & 0.3328 \\ \mathbf{0.48} & \mathbf{0.64} & 1 & 0.3328 & 0.4992 \\ \mathbf{0.9429} & \mathbf{0.4992} & \mathbf{0.3328} & 1 & 0.9230 \\ \mathbf{0.9984} & \mathbf{0.3328} & \mathbf{0.4992} & \mathbf{0.9230} & 1 \end{bmatrix} = D_{normal\_dot\_product}$$

We keep only the **lower triangular matrix**, where each value represents the dot product of the $j-th$ vector (d_j) with the $i-th$ vector (d_j), where $i,\ j$ the rows and the columns of the matrix $D_{dot\_product}$ respectively.

The formula of the dot product of two vectors, using the included angle θ , is:

$$d_i \cdot d_j = \|d_i\| * \|d_j\| * cos\theta \implies cos\theta = \frac{d_i \cdot d_j}{\|d_i\| * \|d_j\|} = cos\theta_{d_i, d_j}$$

The process we followed above, calculates exactly the value of the cosine of the included angle θ, giving information on the relation between those vectors.

Therefore, the results are the follwing:

Inner Product (of normalized vectors)	**Similarity**
$cos\theta_{d_1,d_2} = 0.36$	Low
$cos\theta_{d_1,d_3} = 0.48$	
$cos\theta_{d_1,d_4} = 0.9429$	**High**
$cos\theta_{d_1,d_5} = 0.9984$	**High**
$cos\theta_{d_2,d_3} = 0.64$	
$d_2 \cdot d_4 = 0.4492$	
$d_2 \cdot d_5 = 0.3328$	Low
$d_3 \cdot d_4 = 0.3328$	Low
$d_3 \cdot d_5 = 0.4992$	
$d_4 \cdot d_5 = 0.9230$	**High**

Table 3.10

The **Documents 1**, **4** and **5** are **very similar**.

Also, for this exercise, some code was developed to validate the results (included in the uploaded .zip file).