#Date: 11-12-23

#CSC 461 - Assignment 4 - NLP

Nosheen Azhar

FA20-BSE-061

#Description: In the bellow task we will calculate #TF, WOF, IDF and TF·IDF in question 1 and then the similarity between S_1, S_2 and S_3 using cosine, manhatten and euclidean distances.

Question 1.

S_1: The "Data is one of the most important courses in computer science".

S_2: "This is one of the best data science courses".

S_3: The data scientists perform data analysis.

BOW = 'data', 'science', 'is', 'one' 'of' 'the' 'most' 'important' 'courses', 'in', 'computer', 'this', 'best', 'scientists', 'perform' 'analysis'.

Vector R_1: [1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0]

Vector R_2: [1, 1, 1, 1, 1, 1, 0, 0, 1, 0, 0, 1, 1, 0, 0, 0]

Vector R_3: [2, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1]

Length of Vector R_1 = 12

Length of Vector R_2 = 9

Length of Vector R_3 = 6

Tf of Vector S_1

tf('data') = 1/12
tf('Science') = 2/12 = 1/6
tf('is') = 1/12
tf('one') = 1/12
tf('the') = 1/12
tf('most') = 1/12
tf('important') = 1/12
tf('courses') = 1/12
tf('in') = 1/12
tf('computer') = 1/12
tf('this') = 0/12 = 0
tf('best') = 0/12 = 0
tf('scientist') = 0/12 = 0
tf('perform') = 0/12 = 0
tf('analysis') = 0/12 = 0

← tf('of') = 1/12

Tf of Vector S_2

tf('data') = 1/9
tf('Science') = 1/9
tf('is') = 1/9
tf('one') = 1/9
tf('the') = 1/9
tf('most') = 0/9 = 0
tf('important') = 0/9 = 0
tf('courses') = 1/9
tf('in') = 0/9 = 0
tf('computer') = 0/9 = 0
tf('this') = 0/9 =
tf('best') = 1/9
tf('scientist') = 0/9 = 0
tf('perform') = 0/9 = 0
tf('analysis') = 0/9 = 0

← tf('of') = 1/9

Tf of Vector S_3

tf('data') = 2/6 = 1/3
tf('Science') = 0/6 = 0
tf('is') = 0/6 = 0
tf('one') = 0/6 = 0
tf('of') = 0/6 = 0
tf('the') = 1/6
tf('most') = 0/6 = 0
tf('important') = 0/6 = 0
tf('courses') = 0/6 = 0
tf('in') = 0/6 = 0
tf('computer') = 0/6 = 0
tf('this') = 0/6 = 0
tf('best') = 0/6 = 0
tf('scientist') = 1/6
tf('perform') = 1/6
tf('analysis') = 1/6

idf ( )

idf ('data') = log(3/3) = 0
idf ('Science') = log(3/2) = 0.18
idf ('is') = log(3/2) = 0.18
idf ('one') = log (3/2) = 0.18
idf ('of') = log (3/2) = 0.18
idf ('the') = log (3/3) = 0
idf ('most') = log (3/1) = 0.48
idf ('important') = log (3/1) = 0.48
idf ('courses') = log (3/2) = 0.18
idf ('in') = log (3/1) = 0.48
idf ('computer') = log (3/1) = 0.48
idf ('this') = log (3/1) = 0.48
idf ('best') = log (3/1) = 0.48
idf ('scientist') = log (3/1) = 0.48
idf ('perform') = log (3/1) = 0.48
idf ('analysis') = log (3/1) = 0.48

	tf*idf(R1)	tf*idf(R2)	tf*idf(R3)
data	0	0	0
Science	0.03	0.02	0
is	0.015	0.02	0
one	0.015	0.02	0
of	0.015	0.02	0
the	0	0	0
most	0.04	0	0
important	0.04	0	0
courses	0.015	0.02	0
in	0.04	0	0
computer	0.04	0	0
this	0	0.053	0
best	0	0.053	0
2) scientist	0	0	0.08
perform	0	0	0.08
analysis	0	0	0.08

Question 2: Cosine Similarity

Through BoW

$$\cos\theta = \frac{\bar{S}_1 \cdot \bar{S}_2 \cdot \bar{S}_3}{|S_1||S_2||S_3|}$$

$$\cos(S_1, S_2) = \frac{S_1 \cdot S_2}{|S_1||S_2|} = \frac{9}{(3.74)(3)} = 0.3017$$

$$\cos(S_2, S_3) = \frac{S_2 \cdot S_3}{|S_2||S_3|} = \frac{3}{(3)(2.4495)} = 0.4882$$

$$\cos(S_1, S_3) = \frac{S_1 \cdot S_3}{|S_1||S_3|} = \frac{8}{(3.74)(2.4495)} = 0.327$$

Through Tf*idf

$$\cos(S_1, S_2) = \frac{S_1 \cdot S_2}{|S_1||S_2|} = \frac{(0\times0)+(0.03\times0.02)+\cdots+(0\times0)}{(0.09)(0.087)} = \frac{1.8\times10^{-3}}{7.85\times10^{-3}}$$

$$= \frac{36}{157} = 0.229$$

$$\cos(S_2, S_3) = \frac{S_2 \cdot S_3}{|S_2||S_3|} = \frac{0}{(0.087)(0.138)} = 0$$

$$\cos(S_1, S_3) = \frac{S_2 \cdot S_3}{|S_1||S_3|} = \frac{0}{(0.09)(0.138)} = 0$$

Manhattan Distance through BOW.

Manhattan Distance $(S_1, S_2) = \sum_i |S_{1i} - S_{2i}|$

$= |0+1+(1-1)+(2-1)+\cdots+(0-0)|$

$= 7$

Manhattan Distance $(S_2, S_3) = \sum_i |S_{i2} - S_{32}|$

$= |(1-2)+(1-0)+\cdots\rightarrow(0-1)|$

$= 14$

Manhattan Distance $(S_1, S_3) = \sum_i |S_1, S_3|$

$= |[(1-2)+(2-0)+(\cdots+(0-1)|$

$= 11$

Euclidean Distance Using BOW.

$\sqrt{(S_1, S_2)^2} = \sqrt{(1-1)^2+(2-1)^2+\cdots+(0-0)^2}$

$= 2.6458$

$\sqrt{(S_2, S_3)^2} = \sqrt{(1-2)^2+(1-0)^2+\cdots+(0-1)^2}$

$= 14$

$\sqrt{(S_1, S_3)^2} = \sqrt{(1-2)^2+(2-0)^2+\cdots+(0-1)^2}$

$= 11.$

Manhattan Distance through tfidf

ManhattanDistance $(S_1, S_2) = \sum_i |S_{1i} - S_{2i}|$

$= |(0-0) + (0.03-0.02) + \cdots + (0-0)|$

$= 0.054$

ManhattanDistance $(S_2, S_3) = \sum_i |S_{2i} - S_{3i}|$

$= |(0-0) + (0.02-0) + \cdots + (0-0.08)|$

$= |-0.034| = 0.034$

ManhattanDistance $(S_1, S_3) = \sum_i |S_{1i} - S_{3i}|$

$= |(0-0) + (0.03-0) + \cdots + (0-0.08)|$

$= |0.01| = 0.01$

Euclidean Distance through tfidf

Euclidian $D(S_1, S_2) = \sqrt{\sum_i (S_{1i} - S_{2i})^2} = \sqrt{(0-0)^2 + (0.03-0.02)^2 + \cdots + (0-0)^2}$

$= 0.1105$

Euclidian $D(S_2, S_3) = \sqrt{\sum_i (S_{2i} - S_{3i})^2} = \sqrt{(0-0)^2 + (0.02-0)^2 + \cdots + (0-0.08)^2}$

$= 0.1637$

Euclidan $D(S_1, S_3) = \sqrt{\sum_i (S_{1i} - S_{3i})^2}$

$= \sqrt{(0-0)^2 + (0.03-0)^2 + \cdots + (0-0.08)^2}$

$= 0.1655$