Kshitija Shelar | Text Analytics COMP4760 | Date: 01/11/20

20200155 | Practical 5

1) a. Analysing category of food in this question. Comparing features like taste, texture, ingredients etc. Following are the 5 instances used for analysis.
   i) Chicken nuggets = "spicy fried cheesy crispy meaty"
   ii) Cheesecake= "sweet baked cheesy creamy non-meaty"
   iii) Pie = "sour fried non-cheesy crumbly meaty"
   iv) Soup = "salty grilled cheesy creamy meaty"
   v) Burgers = "spicy fried cheesy crispy meaty"

   Jaccard distance is used to measure the dissimilarity between samples. It is calculated by subtracting 1 from Jaccard Index. Jaccard index is calculated as size of intersection divided by size of union of sample sets.

   The Jaccard distance between above mentioned points is measured. See matrix below.

	Chicken-nuggets	Cheesecake	Pie	Soup	Burgers
Chicken-nuggets	[0.0,	0.89,	0.75,	0.75,	0.0]
Cheesecake	[0.89,	0.0,	1.0,	0.75,	0.89]
Pie	[0.75,	1.0,	0.0,	0.89,	0.75]
Soup	[0.75,	0.75,	0.89,	0.0,	0.75]
Burgers	[0.0,	0.89,	0.75,	0.75,	0.0]

   The higher the distance, the more dissimilar 2 samples are. For example, Chicken nuggets and cheesecake has Jaccard distance 0.89, which means they are very dissimilar. From the matrix above, we can observe that Cheesecake and Pie are the most dissimilar pair (distance between them is 1). The most similar set is where distance is the least. In the matrix, chicken nuggets and burgers have the least distance.

b. Property of triangle of inequality states that distance of point x to z is always less that or equal to distance from x to y + distance of y to z.

dist(x,z) =< dist(x,y) + dist(y,z)

From the above matrix we can calculate is as follows

Dist(Chicken nugget, Soup) = 0.75

Dist(Chicken nugget, Cheesecake) = 0.89

Dist(Cheesecake, Soup) = 0.75

Therefore, 0.75=< 0.89 + 0.75 , 0.75<= 1.64, which holds true.

Triangle of inequality is used to define distance and find shortest path.

Hence, as shown above, it is observed that triangle of inequality rule is followed in Jaccard distance.

c. Dice Coefficient: It is used to calculate the similarity between sample sets. It is calculated as 2 times intersection of sample sets divided by the sum of number of species in each sample. Since we are calculating dissimilarity, we can subtract the result with 1 which will give us the dissimilarity measure. It is almost similar to Jaccard distance. In the following matrix we can observe that a few values and lesser than those in Jaccard distance.

i. We can observe that wherever the Jaccard distance was 0, the dice coefficient is also 0
ii. We can observed that wherever the Jaccard distance was 1, the dice coefficient is also 1
iii. Mostly there is no much difference observed in any other values other than 0 and 1. There is only point difference between other values. Like 0.75 is down to .60 in dice coefficent.

	Chicken-nuggets	Cheesecake	Pie	Soup	Burgers
Chicken-nuggets	[0.0,	0.8,	0.6,	0.6,	0.0]
Cheesecake	[0.8,	0.0,	1.0,	0.6,	0.8]
Pie	[0.6,	1.0,	0.0,	0.8,	0.6]
Soup	[0.6,	0.6,	0.8,	0.0,	0.6]
Burgers	[0.0,	0.8,	0.6,	0.6,	0.0]

2. a. Cosine similarity: Is used to measure how similar the documents are irrespective of the length of these documents. It is measured by the angle between 2 vectors. So if the angle is small, the vectors are more similar. Higher angles means less similarity.

The 2 similar topics I have chosen are 9/11 terrorist attack (New York) and 26/11 terrorist attack (Mumbai).

Here, Doc1, Doc2, Doc3 are related to 9/11 topic

And Doc4, Doc5, Doc6 are related to 26/11 topic.

In the matrix below we can observe that similarity between Doc1, Doc2 and Doc3 is more than similarity between Doc1 and Doc3 or any other Doc from 26/11 topic.

Vise versa is true as well.

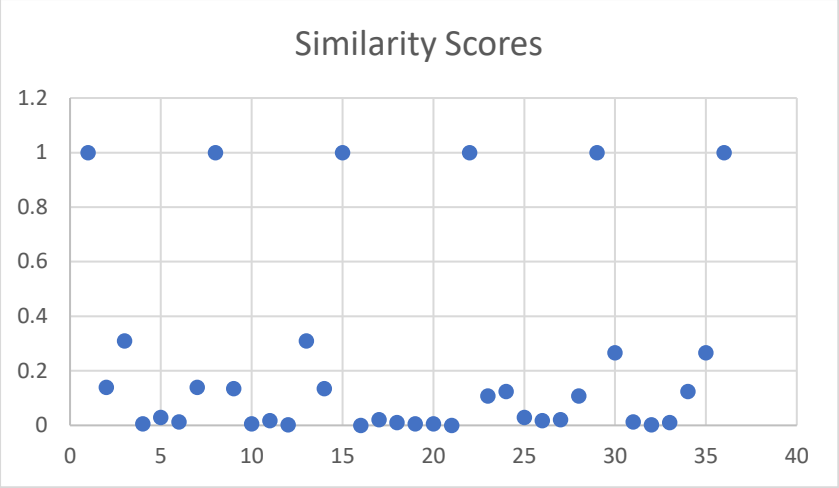
	Doc1	Doc2	Doc3	Doc4	Doc5	Doc6
Doc1	[1.0, 0.139, 0.31, 0.005, 0.029, 0.013]					
Doc2	[0.139, 1.0, 0.135, 0.006, 0.017, 0.002]					
Doc3	[0.31, 0.135, 1.0, 0.0, 0.021, 0.01]					
Doc4	[0.005, 0.006, 0.0, 1.0, 0.107, 0.124]					
Doc5	[0.029, 0.017, 0.021, 0.107, 1.0, 0.266]					
Doc6	[0.013, 0.002, 0.01, 0.124, 0.266, 1.0]					

For example, it is observed that doc5 and doc6 has 0.26 cosine similarity. It is because it has the following things in common .

“2008 Mumbai attacks” “places- Chhatrapati Shivaji Terminus, Mumbai Chabad House, The Oberoi Trident, The Taj Palace & Tower, Leopold Cafe, Cama Hospital, The Nariman House,the Metro Cinema, and in a lane behind the Times of India building and St. Xaviers College” “in South Mumbai at Chhatrapati Shivaji Terminus, Mumbai Chabad House, The Oberoi Trident, The Taj Palace & Tower, Leopold Cafe, Cama Hospital, The Nariman House, the Metro Cinema”

In both these articles, the places where the attack was done is named. Since the list is same, we can see a higher score.

b. Plotting these values gives the following graph.



Example of pairs is given in above i.e 2 a. part of the question.

From the graph we can observe few values are 1. These are documents similarity with itself. Other points are scattered across according to the scores.

c. We can observe a little bit higher values in case of cosine from sklearn. This could be because I have used different preprocessed data as input. The internal working of sklearn is not visible to us so not really sure why so much difference is seen. However, one common thing we can observe is the similarity of a document with itself is 1 in both cases.

	Doc1	Doc2	Doc3	Doc4	Doc5	Doc6
Doc1	1	0.500295	0.586857	0.066774	0.133688	0.119624
Doc2	0.500295	1	0.409341	0.0582354	0.114492	0.091827
Doc3	0.586857	0.409341	1	0.0440744	0.104399	0.0951141
Doc4	0.066774	0.0582354	0.0440744	1	0.447735	0.416125
Doc5	0.133688	0.114492	0.104399	0.447735	1	0.599201
Doc6	0.119624	0.091827	0.0951141	0.416125	0.599201	1

Cosine similarity from sklearn

Manhattan distance: The sum of absolute difference of coordinate points is called Manhattan distance.

Bellow we can see the matrix of Manhattan distance. Here we can observe that 0 is appearing instead of 1. That is the distance of one document from itself. Since it's the same, it is coming out to be 0. So in short, the higher the distance, the more dissimilar the points are.

We can see doc2 and doc4 are 14 points apart. That is because they do not have much things in common. Whereas, doc5 and doc6 are very close that means they have many words in common.

	Doc1	Doc2	Doc3	Doc4	Doc5	Doc6
Doc1	0	7.42548	6.71009	13.1796	12.2104	12.6376
Doc2	7.42548	0	8.96996	14.3441	13.3741	13.8072
Doc3	6.71009	8.96996	0	14.2789	13.2276	13.457
Doc4	13.1796	14.3441	14.2789	0	9.54566	9.8389
Doc5	12.2104	13.3741	13.2276	9.54566	0	6.98668
Doc6	12.6376	13.8072	13.457	9.8389	6.98668	0

Manhattan distance from sklearn