CSE 708

Mid Term Exam

Robert Akinie

Question 1

Using cosine similarity compute the similarity between these two sentences

sentence1 and sentence2:

Sentence 1: "the cat sat on the mat."

Sentence 2: "the cat slept on the bed."

First step: Recording frequency of unique words that appear in each sentence

	the	cat	sat	slept	on	the	mat	bed
Sentence 1	1	1	1	0	1	1	1	0
Sentence 2	1	1	0	1	1	1	0	1

Second step is to determine the values of the cosine similarity equation, using the term-frequency as vectors

$$\cos(s_1, s_2) = (s_1 \cdot s_2)/||s_1||\,||s_2||$$

$$(s_1 \cdot s_2) = (1*1+1*1+1*0+0*1+1*1+1*1+1*0+0*1) = 4$$

$$||s_1|| = ((1*1+1*1+1*1+0*0+1*1+1*1+1*1+0*0)^{0.5} = 2.4495$$

$$||s_2|| = ((1*1+1*1+0*0+1*1+1*1+1*1+0*0+1*1)^{0.5} = 2.4495$$

$$\cos(s_1, s_2) = \frac{4}{2.4495 * 2.4495} = \mathbf{0.667}$$

Question 2

Use one of techniques we have learned in CSE708 to normalize the following data. Please indicate what technique you chose. Age= [12, 16,10, 20, 15]

From CSE 708, normalization can be performed in three ways: Min-max, z-score and decimal scaling

With **z-score normalization**, new normalized values are determined with the following equation:

$$x^i = \frac{x - \mu}{\sigma}$$

Where x = original value, μ = mean, σ = standard deviation

Mean is calculated as (12+16+10+20+15)/5 = 14.6

Standard deviation is calculated as

$$\sqrt{\frac{\sum(|12-14.6|^2+|16-14.6|^2+|10-14.6|^2+|20-14.6|^2+|15-14.6|^2}{5}} = 3.84$$

From the equation above, the new normalized values are

[-0.68, 0.36, -1.20, 1.40, 0.10]

This can also be done with Python, as shown below

[1]:

```
import math
import numpy as np
import pandas as pd
```

[3]:

```
ds = pd.Series([12, 16, 10, 20, 15])
norm=(ds-ds.mean())/ds.std()
norm
```

[3]:

```
0   -0.675838
1    0.363913
2   -1.195713
3    1.403663
4    0.103975
dtype: float64
```

Question 3

Given the following dataset.

Coffee	Day	Temperature	Course	Observation
No	Monday	Hot	CSE708	Sleepy
Yes	Tuesday	Hot	CSE708	Sleepy
Yes	Thursday	Hot	CSE708	Sleepy
No	Monday	Cold	CSE708	No sleepy
Yes	Friday	Cold	CSE708	No Sleepy

We can deduce the following:

The variables coffee, Day, Temperature and Course are predictor variables, and the Observation is the target variable.

Taking a look at the Day variable, Monday appears twice, with the same Coffee and Course variable values being “No.” From these two entries, it can be seen that the Temperature variable has an effect on Observation. When keeping the Temperature and Coffee variables constant, the Day variable does not have any impact on Observation. The same conclusion can be drawn for Coffee when the Temperature is kept constant.

It can be seen that only Temperature has a direct impact on Observation

a. The most likely observation from George would be that he is **No Sleepy**, as a result of earlier deductions showing that Temperature only affects the observation, with Cold predicting No Sleepy.
b. Marie’s observation would be **No Sleepy,** with the Temperature being Cold outside.