Natural Language Processing

BSCS

Final TERM EXAM FALL 2020

DEPARTMENT OF COMPUTER SCIENCE,

SCHOOL OF SYSTEMS AND TECHNOLOGY,

UNIVERSITY OF MANAGEMENT AND TECHNOLOGY

Time: 240 Minutes Total Marks: 60 **weight: 30%**

Resource Person: Dr. Nabeel Sabir Date: 25-Feb -2021

Problem 1: There is a repository of documents. The total number of documents is 10000. Let's suppose you chose a document say P. Total number of words in P is 250 and the word" NLP" comes around 20 times in P. Also it is known that the same word "NLP" appears in 2500 documents. You are required to calculate the tfIdf for the word "NLP". Show complete working to get any credit. **(5)**

Problem 2: Suppose we have three documents P1, P2 and P3. **(10)**

P1=" Love is a world language.".

P2= "Love is a miracle"

P3= "Love is a world feature of the humanity life"

You are required to determine the similarities between these 3 documents by using TF-IDF and for measuring similarities use cosine similarity. Determine which pair of document is most similar. First you need to provide mathematical solution and then also provide the code.

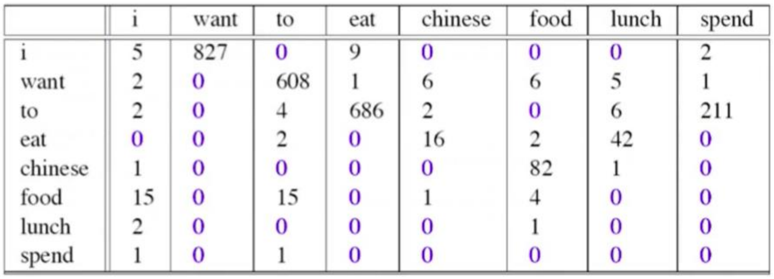
Problem 3 (a): What are the main problems in n-gram language model. Explain with the help of examples. Also discuss their possible solutions. **(3)**

Problem 3 (b): What are the draw backs of one hot vector representation. How these problems are solved? **(2)**

Problem 4: By using the unigram and bigrams counts in the table given below, find the following probabilities for bi-gram language model**. (5)**

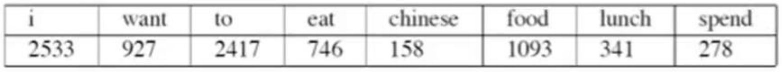
i. P(spend | to)
ii. P(food | eat)

Table 1: Bigram Counts



	i	want	to	eat	chinese	food	lunch	spend
i	5	827	0	9	0	0	0	2
want	2	0	608	1	6	6	5	1
to	2	0	4	686	2	0	6	211
eat	0	0	2	0	16	2	42	0
chinese	1	0	0	0	0	82	1	0
food	15	0	15	0	1	4	0	0
lunch	2	0	0	0	0	1	0	0
spend	1	0	1	0	0	0	0	0

Table 2: Unigram Counts



i	want	to	eat	chinese	food	lunch	spend
2533	927	2417	746	158	1093	341	278

Problem 5: Why NLP is hard. Explain with the help of examples. **(5) .**

Problem 6 (a): Write a complete note on different preprocessing techniques you have studied in the course. Explain each technique with example. In the end provide a jupyter notebook showing implementation of all preprocessing techniques**. (10)**

Problem 6 (b): Write a detailed description of different feature extraction techniques studied in the course. Explain with the help of examples. In the end provide their implementation as well in a Jupyter file. **(10)**

Problem 7: (10): “Don’t write anything from web”.

- Explain in your own words what NLP is. What are key application areas where NLP is being used? Try to avoid coping from web.
- Explain in 1 or 2 paragraphs about the topic and details of your NLP term project along with research article. The status of the article and final date when you submit the best version from your end. Make it sure best means best. The grade has a strong dependency on the article you may submit.
- Write in your own words how this course gives you an idea about NLP and what are the possible research areas where in future you can work and grow.