CSCI 59000 NLP Spring 2024 Homework 5

April 9, 2024

Instructions

We will be using Canvas to collect your assignments. Please read the following instructions to prepare your submission.

1. Submit your solution in a pdf file and a zip file (<yourLastName_FirstName>.pdf/zip). Your write-up must be in pdf. Your code must be in the zip file.
2. In your pdf file, the solution to each problem should start on a new page.
3. Latex is strongly encouraged to write your solutions, e.g., using Overleaf (https://www.overleaf.com/). However, scanned handwritten copies are also acceptable. Hard copies will not be accepted.
4. You need to add screen captures of your code and the output in your write-up.
5. You may discuss the problems and potential directions for solving them with another student. However, you need to write your own solutions and code separately, and not as a group activity. Please list the students you collaborated with on your submission.

Problem 1 (30 points)

Below is a table that represents the occurrences of words in documents.

	Doc1	Doc2	Doc3	Doc4
cup	0	87	78	3
summer	10	12	15	50
coffee	25	18	0	0
walk	3	6	32	20
vacation	20	23	18	15

Answer the following questions. Show your work.

(a) Compute TF-IDF for *coffee* in each document. You will get four results ($TFIDF_{coffee,doc1}$, $TFIDF_{coffee,doc2}$, $TFIDF_{coffee,doc3}$, $TFIDF_{coffee,doc4}$)

(b) Compute TF-IDF for *vacation* in each document. You will get for four results. ($TFIDF_{vacation,doc1}$, $TFIDF_{vacation,doc2}$, $TFIDF_{vacation,doc3}$, $TFIDF_{vacation,doc4}$)

(c) Compare TF-IDF values from (a) and (b). Discuss the results in terms of their discriminating power.

(d) Compute cosine similarity between pairs of documents using TF-IDF representations of the documents. Which documents are most similar? Which documents are least similar?

Problem 2 (30 points)

Below is a table that represents the occurrences of words and their context words. We are going to compute Positive Pointwise Mutual Information (PPMI) following what we learned in class. Answer the following questions. Show your work.

	keyboard	webcam	sheep	legs	racing
whiteboard	1434	512	8	42	3
computer	5290	2351	19	9	42
mouse	6102	3425	352	4251	532
horse	1	53	531	2542	3513
vacation	14	53	34	5	9

(a) First, fill out the counts of word and context words, $count(w)$ and $count(context)$.

	keyboard	webcam	sheep	legs	racing	$count(w)$
whiteboard	1434	512	8	42	3	
computer	5290	2351	19	9	42	
mouse	6102	3425	352	4251	532	
horse	1	53	531	2542	3513	
vacation	14	53	34	5	9	
$count(context)$						

(b) Compute $p(w, context)$, $p(w)$, and $p(context)$.

	$p(w, context)$					$p(w)$
	keyboard	webcam	sheep	legs	racing	$p(w)$
whiteboard						
computer						
mouse						
horse						
vacation						
$p(context)$						

(c) Compute PPMI scores, and fill out the table.

	keyboard	webcam	sheep	legs	racing
whiteboard					
computer					
mouse					
horse					
vacation					

(d) Compute cosine similarity between pairs of words (e.g., whiteboard-computer, whiteboard-mouse, horse-vacation) using the PPMI word representations. Which words are most similar? Which words are least similar?

Problem 3 (Programming involved, 40 points)

In this assignment, you will answer the questions using WordNet and . You will use the `NLTK` package to access WordNet. A NLTK WordNet tutorial is available here: `https://www.nltk.org/howto/wordnet.html`

(a) Write the code to find hypernyms of the first definition of the word *house*. Print out the hypernyms.

(b) Using the `path_similarity` function (see the tutorial), write the code to find the path similarity between the first senses of the two words, *mouse* and *horse*. What is the similarity? What is the path similarity between the first senses of *horse* and *vacation*? Which pair is more similar?

(c) Download the Glove word embedding vectors (glove.6B.zip) from this website: `https://nlp.stanford.edu/projects/glove/`. We will use glove.6B.50d.txt. Write the code to compute cosine similarity between two words using the Glove word embedding vectors. Don't use any pre-built function/package for cosine similarity. You can use `numpy`.

(d) What is the cosine similarity between *mouse* and *horse*? What is the cosine similarity between *horse* and *vacation*? Which one is bigger?

3 bonus points will be given if your homework is easy to review.