Parham Rahimi

Iran University of Science and Tech.
www.iust.ac.ir

Assignment 5 Problems

Natural Language Processing : Fall 1401 : Dr. Minaei
Due Thursday, Dey 1st, 1401

Contents

Problem 1

Mention two common approaches to sequence modeling and discuss their advantages and disadvantages. **(10 points)**

Problem 2

In this problem, you should use the following sentences to implement POS tagging using HMM method on this sentence: "<S> Can Will hunt Mark? <E>"
<S> Will mark hunt. <E>
<S> Mark will hunt. <E>
<S> Can Will mark? <E>
<S> Mark can hunt. <E>
Use only "Noun", "Verb", and "Modal" tags. <S> tags the start of the sentence and <E> tags the end of it. Follow the following steps to achieve the required results:

(a)

First, apply POS tagging on the given sentences and create a table that shows the probability that each word is a noun, modal, or verb. You can use the first row of the following table as an example: **(5 points)**

	Noun	Verb	Modal	Total Count
will	2	0	1	3
mark				
hunt				
can				

(b)

Second, calculate the probability of two labels occurring together. You can use following table structure: **(5 points)**

	<S>	Noun	Verb	Modal	<E>
<S>					
Noun					
Verb					
Modal					
<E>					

(c)

Finally, use the results of previous sections to perform POS tagging on "<S> Can Will hunt Mark? <E>". You need to draw a graph of the sentence, then delete the edges that have zero probability to get the answer. **(10 points)**

Problem 3

In this problem, you should incorporate different ways to assess the similarity between two sentences using vector semantics. Use the Hugging Face stsb_multi_mt dataset, a sample of which is illustrated in the following table:

Problem 3 continued on next page...

sentence 1 (string)	sentence 2 (string)	similarity_score (float32)
”A plane is taking off.”	”An air plane is taking off.”	5
”A man is playing a large flute.”	”A man is playing a flute.”	3.8
”Three men are playing chess.”	”Two men are playing chess.”	2.6

The dataset consists of 3 columns, two strings (sentences), and one float (similarity score).

(a)

Set up your environment and visualize your dataset in a simple way. **(10 points)**

(b)

Use the number of common words between sentences as a measure of similarity. Compare your results with the dataset using the Spearman rank correlation coefficient.
Visualize the top 10 most similar and the least similar pair sentences that your model found using a table. **(20 points)**

(c)

Use the TF-IDF vectorizer to embed the dataset's sentences. Use cosine to measure the similarities of sentences' embeddings and compare your results with the dataset's score (Use Spearman rank correlation coefficient).
Visualize the top 10 most similar and the least similar pair sentences that your model found using a table. **(20 points)**

(d)

(Bonus section)
Use a deep learning model to embed the sentences and compare your results with the dataset (Use Spearman rank correlation coefficient).
Visualize the top 10 most similar and the least similar pair sentences that your model found using a table. **(10 bonus points)**

(e)

Compare your results from different methods and explain the advantages and disadvantages of each method. **(20 points)**

Notes

- Codes should be implemented in .ipynb format (notebooks)
- All Code cells should be executed before turning in the assignment (Make sure your outputs are there before you submit your assignment)
- Please explain the code and the results in the notebook
- If you have any questions, feel free to ask. You can ask your questions in the Telegram group.
- Please upload your assignments as a zipped folder with all necessary components. Upload your file in HW5-NLP-YourStudentID-YourName.zip format.