CIS/STA 9665: Assignment 4

Applied Natural Language Processing

Due Date: 11:59 pm Oct 22

Guidelines:

- Use Python as a programming language and finish this assignment in Jupyter Notebook
- Work is to be done individually for this assignment
- Students handing in similar work will both receive a grade of 0 and will face disciplinary actions.

Chapter 5. Categorizing and Tagging Words

1. What are the most common adverbs in the brown corpus (categories="news")? Please output the 10 most frequent ones (Please use the universal tagset).

2. What are the part-of-speech tags before the word “religion” in the brown corpus (categories="religion")? (Please use the universal tagset)

3. What are the words that are highly ambiguous as to their part-of-speech tags ((i.e. the word has more than 3 pos tags) in the brown corpus (categories="adventure") (Please use the universal tagset).

4. Train a unigram tagger on the brown corpus (categories="humor"). a) Split the data into training and testing dataset- training on the 95% of data and testing on the remaining 5%. b) Evaluate the performance of this tagger on the test dataset. c) Use this tagger to tag some new text ['this','is','a','NLP','class']. d) Observe that some words are not assigned a tag. Explain why not? (Please do **not** use the universal tagset)

5. Explore the nps_chat corpus and find out what part-of-speech tags occur before a noun, with the most frequent ones first (Please use the universal tagset).

6. Explore the brown corpus (categories="romance") to find out all tags starting with VB and its associated (word, frequency) pairs (no more than 6 pairs). (Please do not use the universal tagset)

For example, one of the outputs should look like:

VBG [('going', 59), ('looking', 36), ('trying', 23), ('thinking', 21), ('watching', 20), ('taking', 19)]

7. Write programs to process the brown corpus (categories="editorial") and find answers to the following questions (Please do **not** use the universal tagset):

a. Which nouns are more common in their plural form (e.g. tag='NNS'), rather than their singular form (e.g. tag='NN')? (Only consider regular plurals, formed with the -s suffix).

b. What do the 10 most frequent tags represent in the Brown Corpus? Please output the tags and explain the meaning for each tag.

8. Write code to search the brown corpus (categories="hobbies") for particular words and phrases according to tags, to answer the following questions (please do **not** use the universal tagset):

a. Produce an alphabetically sorted list of the distinct words tagged as MD.

b. Identify three-word prepositional phrases of the form IN + AT + NN (eg. in the lab).

9. Use a default dictionary and itemgetter (n) to sort the most frequent tags used in the brown corpus (categories="reviews"). Please first convert the tags into the universal tags.

10. Explore the brown corpus (categories="learned") to find out the most 200 frequent words and store their most likely tags. We can then use this information as the model for a "lookup tagger" (an NLTK UnigramTagger). If the words are not among the 200 most frequent words, we would like to assign the default tag of "NN" to them. Then use this lookup tagger to tag a new sentence of your own.

What to Submit

a. Use Python as a programming language and finish this assignment in Jupyter Notebook
b. I have created an ipynb file with questions. Please add your code and answers in this ipynb file
c. After completion, please save your finalized ipynb file as a PDF file
d. Submit both **PDF file** and **ipynb file** to Blackboard
e. Please answers questions clearly, concisely, and completely. To answer some questions, the code is not sufficient. You should complement your answers in words by using **comments (#)**
f. The assignment will be graded on the correctness of the answers, comprehensiveness of the analysis, clarity of results' presentation and neatness of the report.