Assignment 3: Parts-of-Speech Tagging (20 points)

CS 410/510 Natural Language Processing Fall 2021
Due Thursday, 11/11/2021, 11:59pm

If you think that NLP models and their associated documentation seem to be getting updated almost weekly, you are probably right. Reading library documentation thus is a good skill to practice. In this assignment, we will rifle through the **NLTK's documentation on Parts-of-Speech Tagging**[1] and work through the following exercises.

Instructions:

- Mandatory: Questions 1 to 6, and *any one* from 7, 8, or 9.
- Optional extra credit: If you attempt two or more from questions 7, 8 or 9.

1. (1pt) Search the web for 2 "spoof newspaper headlines", to find such gems as: *British Left Waffles on Falkland Islands*, and *Juvenile Court to Try Shooting Defendant.* Manually tag these headlines to see if knowledge of the part-of-speech tags removes the ambiguity.
2. (1pt) Tokenize and tag the following sentence: *They wind back the clock, while we chase after the wind.* What is the output?
3. (1pt) Pick 2 words that can be either a noun or a verb (e.g., *contest*). Predict which POS tag is likely to be the most frequent in the Brown corpus, and compare with your predictions.
4. (2pt) Use `sorted()` and `set()` to get a sorted list of tags used in the Brown corpus, removing duplicates.
5. (4pt) Write programs to process the Brown Corpus and find answers to the following questions: (i) Which nouns are more common in their plural form, rather than their singular form? (Only consider regular plurals, formed with the -s suffix.) (ii) List the top 20 tags in order of decreasing frequency - what do these most frequent tags represent?
6. (5pt) Generate some statistics for tagged data to answer the following questions: (i) What proportion of word types are always assigned the same part-of-speech tag? (ii) How many word types are ambiguous, in the sense that they appear with at least two tags? (iii) What percentage of word *tokens* in the Brown Corpus involve these ambiguous word types?

[1] NLTK Ch. 5 - Categorizing and Tagging Words `http://www.nltk.org/book/ch05.html`

7. (6pt) Write code to search the Brown Corpus for particular words and phrases according to tags, to answer the following questions: (i) Produce an alphabetically sorted list of the distinct words tagged as MD. (ii) Identify words that can be plural nouns or third person singular verbs (e.g. *deals, flies*). (iii) What is the ratio of masculine to feminine pronouns?

8. (6pt) How serious is the sparse data problem? Investigate the performance of n-gram taggers as n increases from 1 to 6. Tabulate the accuracy score.

9. (6pt) There are 264 distinct words in the Brown Corpus having exactly three possible tags. (i) Print a table with the integers 1..10 in one column, and the number of distinct words in the corpus having 1..10 distinct tags in the other column. (ii) For the word with the greatest number of distinct tags, print out sentences from the corpus containing the word, one for each possible tag.

Submission and Grading: Please upload your PDF containing your solutions as well as a zip file containing your code or a link to your Jupyter notebook on D2L.