CISC 489/689
Introduction to NLP
Homework 1 Contd.

Due Date: Monday, March 9th, 2020, 11:59 pm

Q1. DT question due on Monday Feb 24th.

Q2. **(Individual Work)** For this question, you need to run the sentence splitter from the NLTK package. Please run it on the file "sentence_splitter_input.txt" that is given to you. Examine the output and indicate all the under-splitting and over-splitting errors in the output. Under-splitting error is one where the sentence splitter didn't split the text into different sentences when it should have, whereas over-splitting error is one where the splitter splits where it shouldn't have.

Group the errors into different subtypes and in a sentence or two indicate what the reason for the error might have been. Also state for each type of error whether a sentence splitter based on your decision tree would have made this error.

Q3. **(Group Work)** You are given a "training set" which contains a set of words and the frequencies with which they appear in the Brown corpus text. For example, according to this file, the word "absence" appears 54 times in the Brown corpus.

Assume there is a blank symbol before and after each word and a tab before the frequency of that word. That is the line for the word "absence" looks like:
" absence	54"

Build a trigram character model using the "stupid" backoff. Please take the frequency of the words into account. For example, the dataset also specifies:
" absences	3"
" absent	34"

Your count for the number of times "s" follows "ab" from these three words should be 54+3+34 = 91. Then you will need to write a program that should say what the most likely next characters are for a given pair of characters (previous two characters). The input will be a set of words and you will need to say what happens as you type each word one character at a time.

For example, if given the word "hello", your program should say

a. what are the three most likely characters after the bigram " h" and the rank of the actual observed character "e".
b. what are the three most likely characters after the bigram "he" and the rank of the actual observed character "l".

c. …

d. what are the three most likely characters after the bigram “lo” and the rank of the actual character “ ”.

You will be provided (later) a set of words for which your program will need to say what the predictions will be.

Q4. **(Group Work)** Manually download the source files from Wikipedia pages for the following states:

1. California
2. Nevada
3. Arizona
4. Florida
5. Maryland
6. Pennsylvania
7. New Jersey
8. North Carolina
9. Vermont
10. Idaho

Note that the URL for the state of California is https://en.wikipedia.org/wiki/California and that the URLs for the other states follow a similar pattern.

Write a program that

1. Returns the gross state product for each state if the Wikipedia entry includes it. The output for each state should only contain the dollar amount. Points will be deducted otherwise.

   On a separate piece of paper, please include a regular expression that will capture the pattern for the dollar amounts for the gross state product.

2. Determines the nicknames of the states. Note: there might be multiple.

The output of your program should be a TSV file, where each line corresponds to a state. The first column should be the state name, the second column should be the gross state product and the third should be the nicknames. Multiple nicknames should be separated by a “;”.