Devin Wu

**CS585 Fall 2024 Programming Assignment #02**

Due: **Sunday, October 13, 2024, 11:59 PM CST**

Points: **100**

**Instructions:**

1. Place **all your deliverables (as described below) into a single ZIP** file named:

LastName\_FirstName\_CS585\_Programming02.zip

1. Submit it to Blackboard Assignments section before the due date. **No late submissions will be accepted**.

**Objectives:**

1. (50 points) Calculate probability of a sentence.
2. (50 points) Language Model word prediction.

**Deliverables:**

Your submission should include:

* **Make sure your code is sufficiently commented! NO Jupyter Notebook files!**
* **Part A**: Python code file(s). Your py file should be named:

CS585\_P02A\_AXXXXXXXX.py

where AXXXXXXXX is your IIT A number (this is REQUIRED!). If your solution uses multiple files, makes sure that the main (the one that will be run to solve the problem) is named that way and others include your IIT A number in their names as well.

* **Part B**: Python code file(s). Your py file should be named:

CS585\_P02B\_AXXXXXXXX.py

where AXXXXXXXX is your IIT A number (this is REQUIRED!). If your solution uses multiple files, makes sure that the main (the one that will be run to solve the problem) is named that way and others include your IIT A number in their names as well.

* this document with your results and conclusions. You should rename it to:

LastName\_FirstName\_CS585\_Programming02.docx or pdf

MS WORD or PDF formats only, please.

**Part A [50 pts]:**

Use Python’s NLTK package along with the Brown corpus for the following tasks:

1. **[1 pts]** Ask the user to enter a sentence S from a keyboard.
2. **[1 pts]** Apply lowercasing to S.
3. **[45 pts]** Calculate P(S) assuming a 2-gram language model (**assume that probability of any bigram starting or ending a sentence is 0.25**)
4. **[3 pts]** Display the sentence S, list all the individual bigrams and their probabilities, and the final probability P(S) on screen. It is fine if it is zero.

**Part B [50 pts]:**

Use Python’s NLTK package along with the Brown corpus (after removing all stop words; use the stopwords corpora for that purpose) for the following tasks:

1. **[1 pts]** Start by asking the user for initial word/token W1. Apply lowercasing to W1 (and all future entries). If the word is NOT in the corpus offer two options:
   1. **ask again**
   2. **QUIT**
2. **[45 pts]** Assuming a 2-gram language model, a menu with TOP 3 “most likely to follow W1” words (according to the W1, NEXT WORD probability estimate). For example, if the user started with W1 = “good”, the following could be displayed (**NOTE: I made up this selection and corresponding probability estimates**):

good …

Which word should follow:

1. morning P(good morning) = 0.25
2. evening P(good evening) = 0.15
3. afternoon P(good afternoon) = 0.14
4. QUIT

**Repeat (and add subsequent word choices to the “sentence”) until user selects (4) and QUITs.**

If the user picks a number other than 1,2,3, and 4, **assume user choice is (1)**.