Northeastern University, Khoury College of Computer Science

CS 6120 — Assignment 4

Due: February 6, 2025 (100 points)

YOUR NAME + LDAP

In this assignment, we will implement an auto-correct algorithm that can be used (in real-time) to determine the most logical correct word substitute for a mispelled word.

- Ernest Hemingway (https://www.gutenberg.org/cache/epub/59603/pg59603.txt)
- William Shakespeare: (https://www.gutenberg.org/cache/epub/100/pg100.txt)
- Emily Dickinson (https://www.gutenberg.org/cache/epub/12242/pg12242.txt)

Delete the header and trailer (anything related to Project Gutenberg) so that you don't use unrelated text. Preprocess these files as you had previously done in Assignment 0, where each file was broken down into all the words. These authors are sufficiently different in tone and wording.

1. Discern which author the passage could have come from
2. Predict the next word to be written using a bi-gram model

Question 1: Create Bigram Distributions

```
def bigram_table(corpus):
  return table
```

Question 2: Infer on Bigrams

(Hint: With the preprocessed word frequencies, you should be able to make a rudimentary classifier that can take a single passage and determine which author produced it.)

Submit your code with the following function signature.

```
def write_in_style_bigram(passage, style_files):
  """
  Takes a passage in, matches it with a style, given a list of
  filenames, and predicts the next word that will appear
  using a bigram model.

  Args:
    passage: A string that contains a passage
    style_file: a list of filenames that will be used to determine the style

  Returns:
    a single word in the form of a string
  """
  % <YOUR-CODE_HERE>
  return []
```