CS2613: Programming Languages Laboratory
Octave: Question 3

Overview:

The following question must be completed in Octave. Submissions must be made to GitHub. Late submissions will not be accepted. All online resources are available including official documentation, forums, videos, etc. If you are struggling with a concept and online research is not helping, then you should ask the course professor.

Learning Goals:

- File I/O
- Naïve Bayes
- Strings
- Cells

Question: Author Detector

For this question, you must write an Octave program that opens and reads text from two .txt files that each contain a single excerpt from the book: “Beowulf.txt” and “Vindication.txt”.

It is recommended that you read in the books and store them as frequency tables for each token for easiest access. For example, the text “Hello world world !” should be stored as:

Hello	1
world	2
!	1

For tokenization, you do not need to handle capitalization or special characters and can treat each token as unique. In other words, for the text “World, world WORLD world World,” would be stored in a frequency table as:

World,	2
world	2
WORLD	1

This should make the rest of the program easier.

After you have your frequency tables set up (one for each book), you will now create a Naïve Bayes algorithm using them to detect the author of a new piece of text. See the video on the following link to learn about Naïve Bayes:
https://www.youtube.com/watch?v=O2L2Uv9pdDA&ab_channel=StatQuestwithJoshStarmer

Write a function that takes a line of text and determines if it is more likely to be written by the author of Beowulf (Unknown) or the author of A Vindication of the Rights of Woman (Mary Wollstonecraft) using Naïve Bayes algorithm. If there is a tie, print “Could be from either author.”.

For example, the sentence “hello world”

$$\text{P(“hello world” | Beowulf.txt)} = \text{P(“hello” | Beowulf.txt)} + \text{P(“world” | Beowulf.txt)}$$

$$\text{P(“hello” | Beowulf.txt)} = \frac{\text{number of times "hello" appears Beowulf.txt}}{\text{Total number of tokens in Beowulf.txt}}$$

$$\text{P(“world” | Beowulf.txt)} = \frac{\text{number of times "world" appears Beowulf.txt}}{\text{Total number of tokens in Beowulf.txt}}$$

Then we must do the calculations for “Vindication.txt”. Whichever is higher is the more likely source.

Example:

```
    determineOrigin(dictionary1, dictionary2, "This is the message");
```

Output:

```
    More likely to written by Unknown Author
```

Hints:

I used cells for the frequency tables. The frequency tables may take a long time to be created. I recommend testing on shorter excerpts from the books first to make sure your functions work before moving on to longer excerpts of the book.