Program #4

CISC 3130

Fall 2016

Due date

This assignment is due at 11:59pm on Sunday, December 11. Four total late days are available to you for this course. One point will be deducted for each unexcused late day. Homework submissions will not be accepted after 11:59pm on Tuesday, December 13.

Description

Write a program to calculate bigram probabilities from a corpus.

A bigram is a pair of words. For example, the bigrams in the previous sentence are (A, bigram), (bigram, is), (is, a), (a, pair), (pair, of), and (of, words). Bigrams can also be word pairs that are unlikely or impossible to exist in English, such as (of, of).

Bigram probabilities are important in many natural language applications such as voice recognition, natural language generation, and machine translation. They can be interpreted as "how likely is it that word 2 will follow word 1?" They are usually computed from large corpora by the following formula:

$$\frac{freq(w1, w2)}{freq(w1)} \tag{1}$$

Create a file called calc_bigrams.cpp. Your main function should take the name of a file as a command line argument. It should call a function to calculate bigram probabilities for every pair of words in the file. You will need two maps: one to keep track of how many times each pair of words occurs (for the numerator in (1)) and one to keep track of how many times each word occurs (for the denominator). You may use the STL map.

After a map has been populated with probabilities for every pair of words in the input file, your program should prompt the user to input a pair of words. Your program should than output the probability of that pair of words. It should continue prompting until the user enters "q".

I will test your code on the documents in the "gutenberg" corpus provided by nltk (linked from the class website). You will know you've implemented the formula correctly if your program outputs `0.0369478` when run on `whitman-leaves.txt`, `0.0248344` when run on `austen-emma.txt`, and `0.0130719` when run on `shakespeare-macbeth.txt` with the input "with me".

Use the following function to remove capitals and punctuation from each string:

```
#include <algorithm>
#include <string>
std::string normalizeString(const std::string& str) {
   std::string res = str;
   std::cout << str << " ";
   std::transform(res.begin(), res.end(), res.begin(), ::tolower);
   res.erase(std::remove_if(res.begin(), res.end(), std::ptr_fun<int,int>(ispunct)), res.end());
   std::cout << res << "\n";
   return res;
}
```

Submission

Please use **Blackboard** to submit this homework. Your submission should consist of one file only: your `calc_bigrams.cpp`. Put your name in a comment in line 1 of the program. Make sure your code compiles

and runs without errors.