Programming Assignment #6

Plagiarism catcher support

CS 3358.751, Summer II 2013
Instructor: Jill Seaman

Due: Monday, 8/5/2012 (upload electronic copy by 11:30am)

Problem:

You will be writing **part** of a tool that could be used to catch plagiarists. The goal of the tool is to determine similarities between documents in a large set to find out if plagiarism is going on within the group.

For this assignment you will write a program that will be able to compare two documents and indicate how many n-word sequences they have in common.

Your program should take as input the name of two files in the working directory, and the value for n (the number of words in a sequence). Your program can prompt the user to enter this information. Then your program should output the number of n-word sequences that the files have in common (ignoring duplicates in either file).

You should use a hash table to compute the results for your program. Implement a hash table for containing strings using the class declaration given in hashtable_3358.h. Implementing the hash table is a requirement for this programming assignment. It is not a template, so you should put your function definitions in hashtable_3358.cpp.

You should also use the program you wrote for assignment 5 as a starter for reading in the two files and collecting the n-word sequences.

NOTES:

- I will be putting a zip file of text files on the class website soon. You can use these files to test your tool (see the file called "catchmeifyoucan.txt").

- Consider using these member functions from the string class:
  bool empty(): returns true if the string is empty
  void clear(): sets string content to empty string (size == 0)

- Write your own hashtable_test.cpp file to test your implementation before you start writing the main program. This will save you from headaches later.

Here is the output from my program. All output is optional except the number of sequences in common.

```
Please enter name of the first file to analyze:
sm_doc_set/catchmeifyoucan.txt
Please enter name of the second file to analyze:
sm_doc_set/ecu201.txt
Please enter the number of words per sequence:
5
***sm_doc_set/catchmeifyoucan.txt
***sm_doc_set/ecu201.txt
word count file1 = 1010
word count file2 = 1291
sequences in common: 289
```

Style:

See the Style Guidelines document on the course website.

Logistics:

Please submit the following files in a single zip file. You can call it assign6_xxxxxx.zip.

```
hashtable_3358.h hashtable_3358.cpp stop_plagiarism.cpp
```

The xxxxx is your TX State NetID (your txstate.edu email id).

Submit: an electronic copy only, using the Assignments tool on the TRACS website for this class.
