CSCI/CMPE 3333 Assignment 2
Instructor: Zhixiang Chen

Team Work Option: *We allow at most two students to form a team to work on this homework. If you choose this option, you and your team member need to submit one copy of your solution to BlackBoard. To do so, please include your name and your team member name in your program document part.*

Problem (100 points): For this problem, I would like you to use hash techniques to identify the most common phrases and common paragraphs between two classics by Mark Twain – *"The Adventures of Tom Sawyer"* and *"Adventures of Huckleberry Finn."*

Specifically, I would like your program to do the following:

Part 1 (100 points): Finding most common phrases. For N=1, 2, 3, 4, 5, 6, 7, 8, 9, and 10, list

- The top 10 most common N consecutive word phrases in these two novels and the frequencies of the phrases in each novel. The format for each phrase in the output shall be

  phrase frequency_in_Tom_Sawyer frequency_in_Huckleberry_Finn

Part 2 (70 points): Finding most similar paragraphs. Below is a paragraph from Mark Twain's another novel:

"THE Mississippi is well worth reading about. It is not a commonplace river, but on the contrary is in all ways remarkable. Considering the Missouri its main branch, it is the longest river in the world—four thousand three hundred miles. It seems safe to say that it is also the crookedest river in the world, since in one part of its journey it uses up one thousand three hundred miles to cover the same ground that the crow would fly over in six hundred and seventy-five. It discharges three times as much water as the St. Lawrence, twenty-five times as much as the Rhine, and three hundred and thirty-eight times as much as the Thames. No other river has so vast a drainage-basin: it draws its water supply from twenty-eight States and Territories; from Delaware, on the Atlantic seaboard, and from all the country between that and Idaho on the Pacific slope - a spread of forty-five degrees of longitude. The Mississippi receives and carries to the Gulf water from fifty-four subordinate rivers that are navigable by steamboats, and from some hundreds that are navigable by flats and keels. The area of its drainage-basin is as great as the combined areas of England, Wales, Scotland, Ireland, France, Spain, Portugal, Germany, Austria, Italy, and Turkey; and almost all this wide region is fertile; the Mississippi valley, proper, is exceptionally so."

I want you to expand your program for Part 1 to do the following:

- List top 10 paragraphs in each of the two novels, “The Adventures of Tom Sawyer” and “Adventures of Huckleberry Finn,” that are mostly similar to the above given paragraph.

One critical part is how to define similarity between two paragraphs. I would like you to use the simplest method: count the number of words appearing in two given paragraphs and use that number as the similarity measure.

A Note for both Part 1 and Part 2: When you solve Part 1 and Part 2, you shall convert any phrases or words into lower cases for comparison. You shall also use only one blank space symbol to separate any two words. For example, “cat dog” can be preprocessed into “cat dog”.

The text files of these novels are given here:
http://faculty.utrgv.edu/zhixiang.chen/cs3333/3333/dic/hw2_tomSawyer.txt (“The Adventures of Tom Sawyer”

http://faculty.utrgv.edu/zhixiang.chen/cs3333/3333/dic/hw2_huckleberryFinn.txt (“Adventures of Huckleberry Finn”)

Warning: Never try to print these files, because they are **VERY LARGE**.

Note:

Due Date:
The due date will be given via Blackboard.

Warning:
Any submission one week after the due date will not be accepted.

How to submit your work?
Please upload your source program files and your test results to Blackboard.