# CSCI 3333 Project Two

# Finding Common Patterns from Mark Twin’s Classics

## Problem Description

For this problem, I would like you to use some combination of hash maps, lists, and sorting to discover the most common phrases and common paragraphs between two classics by Mark Twain:

1. “The Adventures of Tom Sawyer,” and
2. “Adventures of Huckleberry Finn.”

I will upload the plain text files of the above two novels to Blackboard, so that you can down these two for input files to your program.

Specifically, I would like your program to do the following two parts: To find the most common phrases and the most common paragraphs, as detailed below.

**Efficiency requirements:**

* Given the large sizes of the two novels, the design and implementation of your program must consider the run time efficiency, that is, the practical running time complexity of your program.
* Under the usual computing environment, e.g., a typical laptop or a desktop, your program shall be able to complete the work in no more than 1 minute of running time.
* **Hints:** Design your data structures and algorithms well so that you just need to perform one pass of linear scanning/reading of both novel text files, in order to discover the desired patterns.

## Basic Part (100 points): Finding most common phrases

For N=1, 2, 3, 4, 5, 6, 7, 8, 9, and 10, find 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 outputting the top 10 most frequent phrases of N words shall be in a tabular form as shown below:

top 10 most frequent phrases of length 2

phrase frequency in Tom Sawyer frequency in Huckleberry Finn

speckled trout 100 150

red snapper 600 95

To save time and space, you shall let your program to output the top 10 phrases into a text file.

## Optional Part (for possible extra credits): Finding most similar paragraphs

Below isa 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 Basic Part to do the following:

* Find 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 textual information retrieval method: count the number of words appearing in two given paragraphs and use that number as the similarity measure.

**Efficiency requirements:**

* Under the usual computing environment, e.g., a typical laptop or a desktop, your program shall be able to complete the work in no more than 1 minute of running time.

## Notes for both Basic Part and Optional Part

1. When you do Basic Part and Optional Part, 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”.
2. A word shall not include punctuations, but may include hyphens, apostrophes, and underscores.
3. Only English letters and digits (0, 1, 2, …, 9) are considered.

## Plain Text Files

The text files of these novels are given here:

* [The Adventures of Tom Sawyer](http://faculty.utrgv.edu/zhixiang.chen/cs3333/3333/dic/hw2_tomSawyer.txt)
* [Adventures of Huckleberry Finn](http://faculty.utrgv.edu/zhixiang.chen/cs3333/3333/dic/hw2_huckleberryFinn.txt)

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