Programming Assignment 6
Using Apache Spark

Due on August 11 before midnight

Description

The purpose of this project is to develop a simple spark program on Apache Spark to analyze data.

This project must be done individually. No copying is permitted. **Note: We will use a system for detecting software plagiarism.** That is, your program will be compared with the programs of the other students in class as well as with the programs submitted in previous years. This program will find similarities even if you rename variables, move code, change code structure, etc.

Note that, if you use a Search Engine to find similar programs on the web, we will find these programs too. So don't do it because you will get caught and you will get an F in the course (this is cheating). Don't look for code to use for your project on the web or from other students (current or past). Just do your project alone using the help given in this project description and from your instructor and GTA only.

Platform

You will develop this project on your local machine. Apache Spark can only be run on Linux/Unix environment. Follow steps from pre-requisite for this assignment to setup your laptop for Ubuntu if you are Windows user. Download the latest Apache Spark (Spark release: 2.0.0 pre-built for Hadoop 2.7) from here. Unzip it. Open terminal and go to the spark folder. Download the Word count example do the following:

For comiling:

```
$ javac -cp "jars/*" JavaWordCount.java
```

For making jar:

```
$ jar cvf JavaWordCount.jar *.class
```

For Running on Spark

```
$ ./bin/spark-submit --class JavaWordCount --master local[4] JavaWordCount.jar <input local file or
directory> <threshold integer>
```

Documentation:

1. Spark API
2. Apache Spark Prgramming Guide

Project Requirements

You will be implementing Project 5 (computing tf-idf) using Apache Spark.