This is the documentation of the programming assignment for Big Data (CSU CS535).

Project: spark-hits-wikipedia
Distributed Framework: Apache Spark, Apache Hadoop, Apache Yarn
Programming Language: scala
IDE: IntelliJ
Primary namenode: boise
Secondary namenode: boston
Spark WebUI: spark://boise:31721
Jar file: spark-hits-wikipedia_2.11-0.1.jar
Dependency management: sbt

For improving performance, following steps has been applied.
1. Number of outgoing links per page p in RootSet is limited to less than or equal to 20.
2. Number of incoming links per page p in RootSet is limited to less than or equal to 20.
3. The maximum number of iteration is set to less than or equal to 40. In combination with this, the threashold of 0.000005 is set
  for sum of change in consecutive auth scores for base set.

```
    if(i>25){
    if(i>=40 | threshold(tempA, authScore, spark)<0.000005) {
       loop.break
    }
  }
```

4. RDD are cached as required.

Command Line arguments:
args(0) : query word (eg. "eclipse")
args(1) : size of root set (eg. "25")
args(2) : if saving output in hdfs? (eg. for true "Y")
args(3) : number of iteration (eg. "35")

Running the jar:
$SPARK_HOME/bin/spark-submit --class "com.csu.cs535.assignment.HitsImplementation" ~/spark-hits-wikipedia_2.11-0.1.jar "hardware" "20" "N" "30"

Output:
-For number of links of 187 in base-set and iterat, the execution time was: 2.09 min
-Sample Screen Shot attached.
-Sample output in txt format (Root Set, Auth Title, Hub Title) attached.

Bibek Raj Shrestha