CSE 487/587 Programming Assignment #5

Due Date: May 12, 2015
Weight: 12%

Description

In this assignment, you will use Twitter API to collect data of 30 NBA teams, and store the data in CSV format (comma separated) on your local machine. Accumulo will be utilized to create table and compute the popularity of NBA teams based on the collected data.

NBA Team	Hashtag	Twitter Account
Eastern Conference Atlantic Division		
Boston Celtics	#Celtics	@Celtics
New York Knicks	#Knicks	@NYKnicks
Philadelphia 76ers	#76ers	@Sixers
New Jersey Nets	#Nets	@NetsBasketball
Toronto Raptors	#Raptors	@Raptors
Eastern Conference Central Division		
Chicago Bulls	#Bulls	@ChicagoBulls
Indiana Pacers	#Pacers	@Pacers
Milwaukee Bucks	#Bucks	@Bucks
Detroit Pistons	#Pistons	@DetroitPistons
Cleveland Cavaliers	#Cavs	@Cavs
Eastern Conference Southeast Division		
Miami Heat	#MiamiHeat	@MiamiHEAT
Orlando Magic	#OrlandoMagic,	@Orlando_Magic
Atlanta Hawks	#Hawks	@Atlanta_Hawks
Charlotte Bobcats	#Bobcats	@Bobcats
Washington Wizards	#Wizards	@WashWizards
Western Conference Northwest Division		
Oklahoma City	#okcthunder	@OKCThunder
Denver Nuggets	#Nuggets	@DenverNuggets
Portland Trailblazers	#TrailBlazers	@PDXTrailBlazers
Utah Jazz	#UtahJazz	@Utah_Jazz
Minnesota Timberwolves	#TWolves	@MNTimberwolves
Western Conference Pacific Division		
LA Lakers	#Lakers	@Lakers
Phoenix Suns	#Suns	@PhoenixSuns
Golden State Warriors	#GSWarriors	@Warriors
L.A. Clippers	#Clippers	@LAClippers
Sacramento Kings	#NBAKings	@SacramentoKings
Western Conference Southwest Division		
San Antonio Spurs	#GoSpursGo	@Spurs
Dallas Mavericks	#Mavs	@DallasMavs
New Orleans Hornets	#Hornets	@Hornets
Memphis Grizzlies	#Grizzlies	@memgrizz
Houston Rockets	#Rockets	@HoustonRockets

Authorizations:

Creating two username in Accumulo,

- the username "**east**" is allowed to read team data from eastern conference.
- the username "**west**" is allowed to read team data from western conference.

Both users with **different permissions will generate different results**.

Twitter Data Analysis:

Use the Hashtag in the above table to fetch twitter messages of each NBA team. **A sample python script is given.** It will generate a CSV file using Hashtag as file name with 2 columns and 100 rows, the second column "Text" will be used as messages. You **might need to fix the data in CSV file** as there are some \n \r characters in messages to destroy the comma separated format.

. Created_at	Text

The condition: "count=100" must be used (do not change the value, for grading).

```
data = t.search(q=hashtag, count=100)
```

The above data variable will include the latest 100 messages with given Hashtag. Each request with same Hashtag will get different messages (real-time latest). So for grading purpose, **you have to store 100 messages of each team locally**, and submit all the collected data (30 CSV files) along with your Accumulo source code in zip file.

To evaluate the popularity of NBA teams, you will use a very simple strategy as follows:

1. Given the assumption that more keywords "**win**" appeared in the twitter messages of team K means that it has higher popularity.
2. Given the assumption that more keywords "**lose**" appeared in the twitter messages of team K means that it has lower popularity.
3. Calculate the frequency (total number of word appears) of keyword "**win**" (case insensitive) in 100 messages for each team using “east” and “west”, respectively.
4. Calculate the frequency (total number of word appears) of keyword "**lose**" (case insensitive) in 100 messages for each team using “east” and “west”, respectively.
5. The search for keywords is a complete match (whole word match).
6. Sort the two groups of teams (by “east” and “west”) by frequency in descending order.

Note:

- Sometimes, twitter API is **unstable**, so you may run the scripts multiple times to get data.
- The given python script works for version **2.7.9**
- “Example 3 – MapReduce, WordCount” is given as a reference in **AccumuloTutorial.pdf**
- The file "**API_hashtag.py**" to get Twitter data is given for a quick start.

Expected Result Should Looks Like Below(**examples only**)

Username: east

East Conference (15 teams)		
Team Name	**Hashtag**	**Win**
Milwaukee Bucks	# Bucks	34
Detroit Pistons	# Pistons	26
….	….	….
Boston Celtics	# Celtics	8
Chicago Bulls	# Bulls	0

East Conference (15 teams)		
Team Name	**Hashtag**	**Lose**
Cleveland Cavaliers	# Cavs	25
Philadelphia 76ers	#76ers	16
….	….	….
Atlanta Hawks	# Hawks	4
Detroit Pistons	# Pistons	1

Username: west

West Conference (15 teams)		
Team Name	**Hashtag**	**Win**
New Orleans Hornets	# Hornets	27
Houston Rockets	# Rockets	18
….	….	….
LA Lakers	# Lakers	5
Denver Nuggets	# Nuggets	4

West Conference (15 teams)		
Team Name	**Hashtag**	**Lose**
Utah Jazz	# UtahJazz	14
L.A. Clippers	# Clippers	11
….	….	….
New Orleans Hornets	# Hornets	6
Houston Rockets	# Rockets	2

What you need to do:

In your report, include following aspects:

- Your rationale for Accumulo computation.
- Discussion of all the experimental results and comparison results.

Specific Submission Guidelines: Assignment 5

1. Files should be strictly organized as following structure in your own directory (**/gpfs/courses/cse587/spring2015/students/username/hw5/**) and the naming of the directory should be followed exactly (case sensitive):

NOTE THAT YOU SHOULD NOT MAKE ANY CHANGES TO THE DIRECTORY AFTER THE SUBMISSION DEADLINE, AS THE TIME STAMP OF THE FILES WILL BE USED FOR TIMELY SUBMISSION.

hw5/accumulo/src/

(include the source code of your job, e.x. MapReduce code, Accumulo script and etc)

hw5/username.pdf

(your assignment report)

hw5/misc/ (optional)

(include any other files you may want to submit)

2. Use the Accumulo tutorial to install a VM image of Accumulo and run your programs on that:

Grading Criteria

- Program correctness (twitter and Accumulo): 80%
- Discussion and the report: 20%