Department of Computing
The University of the West Indies, Mona

COMP3162 – Data Science Principles

Semester II 2020/2021

Project Overview

A large entertainment & bar franchise (Hard Knocks) would like to determine whether now is a good time to expand and which country/region of the world is best to open next. Ultimately, they would like to gain some insights on whether to open their next franchise, based on public sentiments generally and opportunities to maximize profits. They first want to test the pulse of persons in relation to the products they sell (Beverages) which would indicate the current public views towards these products. In addition, they have a large dataset with sales data by region and country for different types of consumer goods (household, cosmetics etc), food and beverage. You will be helping Hard Knocks to make their decision in this project! The tasks required have been broken down in several segments and you will be required to start this week!

Project Part 01 – Individual Tasks
DUE: March 17, 2021

[Weighted 5% of course marks]

1. Between **March 04 & March 09** (max. 1 word per day) connect to twitter on two separate days and retrieve 8,000 or more tweets containing one of the words from ***a*** and one of the words from ***b*** (total 16,000 tweets). Retrieve tweets for the word from ***a*** on a separate day from ***b***.
   a) "beverage" or "beer"
   b) "party" or "concert" **[4]**

2. For each set of tweets retrieved (a-b above), retain the following features only:
   text, screen_name, user_id, created_at, favourite_count, retweet_count, location, followers_count, friends_count, account_lang, lang.

   a) Remove all non-English tweets (you must indicate how many tweets were removed). **[1]**
   b) A tweet is considered a duplicate if the text is the same as another tweet. Remove all duplicate tweets (you must indicate how many tweets were removed). **[2]**
   c) Write the remaining tweets data to a file (.csv). The csv filename should have the format <keyword>_<date>_<myname>. For example, for tweets on "beverage" retrieved on March 07 by Anderson would be: beverage_2021Mar07_Anderson.csv **[1]**

   d) Write code to review and show details of tweets retrieved including number of tweets (after doing 2a-c), screen_name with the most followers, tweet with the most retweets, location from which the most tweets originate. **[7]**

3. Between **March 10 & 14**, repeat tasks 1 & 2 above. For task 1, use the other words that were not used during March 04-09 tweet retrieval. That is, if you retrieved tweets using "beverage", you should use "beer" and if you used "party" now use "concert" for part 1b.
   Save files using same format (the names will be different given that dates will be different). **[15]**

WHAT TO SUBMIT

Upload a pdf generated using **rmarkdown** to demonstrate that tasks 1-3 have been completed. It should show code used to retrieve tweets, view of a few rows of the retrieved tweets before and after changes are made and the code used to make the changes. Also include short comments where necessary. Marks will be deducted if you show the full dataset of tweets in the document! Only few rows of data and various summaries should be shown in the PDF to prove that each task is completed.