Twitter Project

Dataset description

The dataset you will be using is a collection of Twitter data. Roughly 1% of Twitter data is made available for free through the Twitter API. Larger portions of streaming Twitter data are available for a fee. The department has collected public data from August until February. This is the data we will be using for analysis.

Project Description

You and your teammates will search the collection for tweets related to a topic of interest. The topic of interest is up to you, but we suggest that you select something with a large amount of interest in social media. For example, trending news topics, trending political topics, major sporting events, movies, major influencers etc. These topics tend to have more tweets and make for more interesting analysis. Other topics are possible to study using free Twitter data but require more complex techniques that are beyond the scope of this project. Please inquire with the instructor if you are uncertain about a topic.

Searching the Twitter data for tweets concerning your topic of interest will be accomplished using AWS EMR in the same way as your Hadoop assignment. With the code you will be provided you will only be able to search tweets containing specific hashtags. So, in order to best understand your topic of interest, you may first want to search Google to find trending hashtags related to your topic. You can also explore tweets using Twitter to find other hashtags that people use because these hashtags could be related to the hashtag you're interested in. After searching the data, you will want to examine its content to make sure that the data is valuable and usable for analysis.

You will then analyze this data using Jupyter notebooks. Simple analysis techniques can examine hashtag occurrence or cooccurence, word frequency, time of tweets and analysis of user information (e.g. followers, likes, retweets, location). Visual analytics with Tableau could be a good option to start. NLP techniques can be used for sentiment analysis. Clustering techniques can be used to examine the types of tweets in the datasets. You can collect multiple datasets, with related hashtags, and analyze each dataset with these techniques, then compare the results. During class, the instructor may be able to help you develop tailored analysis techniques based on your topic.

Finally, you will create a PowerPoint presentation on your topic and make a presentation to the class. This presentation should explain your topic of interest, your research question, your analysis and your conclusions.

Instructions

You will find the data on Amazon S3 located in s3://bual5660-twitter-dataset-debug/. There are a number of different folders containing the data because we do not have the authorization from AWS to create very large clusters necessary to process it all at once. These folders are listed by date and can be accessed with the list on the final page. You will insert these file paths as the input for your EMR streaming jobs. Note, you must provide your instructor with your AWS ID number to get access to the data.

Supplementary materials will be provided for you. These include mapper and reducer files for searching the Twitter data. Jupyter notebooks for exploring and analyzing the data. A sample data file for debugging on your laptops and sample results from a query of the first folder in the directory. You will find these supplementary materials in the Dropbox folder at:
https://bit.ly/2DhP9gX

Your grade

You will be graded on your attendance (10%), team formation (10%), data collection (20%), your team presentation and your team evaluations (60%).

- Teams are to be sent to your instructor via email by midnight on March 21st.
- Initial data for your team is due to your instructor by midnight on April 2nd.
- Presentations will be held April 23rd and 25th. You will schedule the day on April 4th during class.

What to submit

- Your PowerPoint presentation
- Your collected Twitter data in JSON format
- The modified Jupyter notebooks used for your analysis
- Your group evaluations
- Your teammate evaluations

FILEPATHS TO DATA:

- s3://bual5660-twitter-dataset-debug/0808/
- s3://bual5660-twitter-dataset-debug/0815/
- s3://bual5660-twitter-dataset-debug/0823/
- s3://bual5660-twitter-dataset-debug/0828/
- s3://bual5660-twitter-dataset-debug/0905/
- s3://bual5660-twitter-dataset-debug/0911/
- s3://bual5660-twitter-dataset-debug/0913/
- s3://bual5660-twitter-dataset-debug/0920/
- s3://bual5660-twitter-dataset-debug/0926/
- s3://bual5660-twitter-dataset-debug/1001/
- s3://bual5660-twitter-dataset-debug/1004/
- s3://bual5660-twitter-dataset-debug/1008/
- s3://bual5660-twitter-dataset-debug/1011/
- s3://bual5660-twitter-dataset-debug/1017/
- s3://bual5660-twitter-dataset-debug/1021/
- s3://bual5660-twitter-dataset-debug/1025/
- s3://bual5660-twitter-dataset-debug/1029/
- s3://bual5660-twitter-dataset-debug/1101/
- s3://bual5660-twitter-dataset-debug/1105/
- s3://bual5660-twitter-dataset-debug/1108/
- s3://bual5660-twitter-dataset-debug/1112/
- s3://bual5660-twitter-dataset-debug/1118/
- s3://bual5660-twitter-dataset-debug/1124/
- s3://bual5660-twitter-dataset-debug/1129/
- s3://bual5660-twitter-dataset-debug/1203/
- s3://bual5660-twitter-dataset-debug/1210/
- s3://bual5660-twitter-dataset-debug/1214/
- s3://bual5660-twitter-dataset-debug/1218/
- s3://bual5660-twitter-dataset-debug/1221/
- s3://bual5660-twitter-dataset-debug/1228/
- s3://bual5660-twitter-dataset-debug/0103/
- s3://bual5660-twitter-dataset-debug/0107/
- s3://bual5660-twitter-dataset-debug/0110/
- s3://bual5660-twitter-dataset-debug/0114/
- s3://bual5660-twitter-dataset-debug/0118/
- s3://bual5660-twitter-dataset-debug/0123/
- s3://bual5660-twitter-dataset-debug/0129/
- s3://bual5660-twitter-dataset-debug/0201/