Angelo, Basconcillo, Dumaliang, Tan, Villanueva, Gamboa

CS 129.1 A

Final Project Proposal

Big data problem:

- Given the list of wine reviews globally, how do we determine which country offers the best wine?

Description of data source

The dataset to be used for the project will come from Kaggle. The list of wine reviews enumerates 150,930 different and unique wines. The said list includes the wine's country of origin, description, designation, points, price, region, variety, and winery. In order to determine if a wine review is positive or negative, the following keywords will be searched:

Positive	Negative
● Juicy	● Rustic
● Rich	● Unclean
● Ripe	● Yeasty
● Delicious	● Flat
● Balanced	● Forced
● Fresh	● Murky
● Impressive	● Harsh
● Smooth	● Bitter
● Silky	● Awkward
● Opulent	● Flabby
● Beautiful	● Weak
● Gorgeous	● Bland
● Wonderful	● Questionable
● Exceptional	● Not (Fresh/Convincing/etc)
● Polished	● Barely (Acceptable/Drinkable/etc)
● Top-notch	● Heavy
● Sophisticated	● Unripe
● Perfect	● Sour
● Decadent	● Dull
● Remarkable	● Unpleasant
● Oaky	● Dry
● Fruity	● Corky
● Crisp	● Coarse
● Soft	● Sulphury
● Noble	● Musty

*Note: More keywords will be used in the actual project

Link: https://www.kaggle.com/zynicide/wine-reviews

Procedure on how to obtain the dataset

1. Go to https://www.kaggle.com/zynicide/wine-reviews
2. Download the .csv file
3. Convert to JSON file

Expected Output

1. ~~Big Data problem proposal~~
2. Raw dataset which includes…
    a. Program used to fetch the dataset from its source, if applicable.
    b. Program used to transform the dataset into MongoDB format, if applicable.
    c. Program used to load the dataset into MongoDB, if applicable.
3. ~~Replicate sets configuration~~
4. ~~MapReduce functions~~
5. Sharding set configuration
6. ~~Short report paper which includes…~~
    a. ~~Approved Big Data problem~~
    b. ~~Description of source dataset~~
    c. Description of the output, explanation and discussion of the output in relation to the Big Data problem
    d. Visualization of the output
7. A Git repository (in Github) that contains all the previous items and a README file which includes
    a. How to load the dataset
    b. How to setup the Replicate sets
    c. How to execute the MapReduce functions
    d. How to run the MapReduce functions
    e. How to shard the MapReduce collection