CS 696 Intro to Big Data
Fall Semester, 2019
Assignment 4

San Diego State University -- This page last updated 1/29/19

Due April 30 11:59 pm

2016 Election Data

You are going to look donation data from the 2016 presidential campaign.

This a assignment uses two datasets. A partial dataset for developing and the full dataset. The partial dataset can be downloaded from the assignment page. The full dataset can be accessed on AWS at s3://rw-cs696-data/itcont.txt. The format of the files is described at:

https://classic.fec.gov/finance/disclosure/metadata/DataDictionaryContributionsbyIndividuals.shtml

The data contains information about each donation made to candidates in 2016.

Hillary Clinton committee was HILLARY FOR AMERICA which has the id C00575795. Bernie Sanders committee as called BERNIE 2016 with id C00577130. Trump's committee was called DONALD J. TRUMP FOR PRESIDENT, INC. with id C00580100. You need those id to find contributions. Read the file formats to determine how to find which donations were for which campaign and how much was donated

1. How many donations did each campaign have?

2. What was the total amount donated to each campaign?

3. What percentage of the each campaign's donations was done by small contributors?

4. Produce a histogram of the donations for each campaign?

You are to use AWS Spark to answer the first three questions. For the fourth question you need to process the data on AWS and download the result so you can use Python plotting tools to produce the histograms.

Instructions

You are to use AWS Spark to answer the first two questions. For the fourth question you need to process the data on AWS and download the result so you can produce the histogram locally.

Turn in a jupyter notebook with all the code used with the answer to the questions. The code for problems 1 - 3 should be in a function that you run on AWS. To show that you ran the code on AWS include in your notebook the AWS CLI export command for each job that you need to run on AWS.

Grading

15 points per problem.

What to turn in

You need to turn in the jupiter notebook and the files that you download from AWS to answer problem 3. Put them in the same directory so the notebook can read the files as local files. Create a zip file of the directory and turn that zipped directory in.

Late Penalty

An assignment turned in 1-7 days late, will lose 5% of the total value of the assignment per day late. The eight day late the penalty will be 40% of the assignment, the ninth day late the penalty will be 60%, after the ninth day late the penalty will be 90%. Once a solution to an assignment has been posted or discussed in class, the assignment will no longer be accepted. Late penalties are always rounded up to the next integer value.