Homework #4

DS 3001, D-term 2018

100 points total [9% of your final grade]

Due: April 22, 2018 by 11:59pm
[no submission will be accepted after April 25, 2018 at 11:59pm]

Delivery: Submit via canvas

Overview

In this homework, you will write Map and Reduce functions to perform following two tasks 1 and 2:

Task 0: Setup a Single-node Hadoop Cluster

Based on the Hadoop tutorial, load the VM image file and run a word count example.

Task 1: Palindrome Count

Given the provided file (Tolstoy's War and Peace), report a frequency of each distinct palindrome that occurs in the text. (e.g., madam:5 \n bob:3 \n). Which palindrome occurs most often?

Task 2: Election Fraud

In this task your job is to investigate whether there was election fraud in 2018. You have 2016 and 2018 election data files: (i) 2016 data file; and (ii) 2018 data file. The files are of the format where each line is a vote in the election.

The format of the text file is:

VoterID \t CountyID \t PartyID

A) Which party won the election in 2018?

B) In 2016, which county was the most monolithic in the manner in which they voted? (i.e. which county came closest to voting 100% for a single party).

Optional Task: No additional points. If you want to have more fun with Hadoop, you may solve the following problems. Do not submit your solution for the following problems. We will not grade yours.

A) Studies have shown if a political party gains more than 50% in voting percentage from one election cycle to the next, then most likely fraud has occurred. (Example, if party A received 100 votes in 2016 in county B, then received 200 votes in 2018, fraud may have occurred). In which counties in 2018 did voter fraud likely occur?

B) From 2016 to 2018 how many voters changed which party they voted for? What is the most common type of change?

What to turn in:

- You should turn in **a PDF report** containing your answers, your **source codes** including Map and Reduce functions, and a **readme file** pointing which code is for what problem. Compress all the files to make a single file, and submit it (e.g., hw4_yourname.zip)
- This is an individual assignment, but you may discuss general strategies and approaches with other members of the class (refer to the syllabus for details of the homework collaboration policy). At the top of your report, please write the names of classmates you consulted and the nature of your discussion.