Distributed Computing Project 2 Logbook

Group Members:

- Josh
- Fehmi
- Paul
- Aaron
- Nick

Main Communication:

- Google Space

Canvas: https://ncf.instructure.com/courses/7333/files/666389?module_item_id=172068
Github: https://github.com/FNeffati/DCP2
Adobserver: https://adobserver.org/ad-database

Communication Expectations:

-

Availability:

High-level plan:

-

Group structure:

-

Logs

Friday, April 28, 2023

Group Meeting with Matt
Notes

- Use pyspark to get random samples for visualizations of data

Monday, April 24, 2023

Group Meeting Agenda
Online meeting
Updates

- Josh
    - Submitted group project plan assignment 1
- Paul
    - Used R to run through EDA for propublica and adobserver data
    - Posted general summary in Google Drive
- Fehmi
    - Looked at targets column in sample data to create new columns (age, gender, etc.)
    - Going to try to write in pyspark for the full dataset
- Nick
    - Preliminary EDA/lit review
- Aaron
    - Not sure what to do with certain columns in data
    - Do we want to join the two datasets?

Topics

- What is our research question?
    - Is there a change in political advertisements between these two time periods?
        - Does the frequency of political advertisements change over time?
        - Do advertising agencies change their target demographics over this time period?
        - Who are the top advertisers over time?
        - Do any major groups suddenly stop advertising?
        - What time of year are ads most frequently posted? (pride month, BHM, etc.)
- Data preparation
    - Fehmi will complete targets column
- Analysis
    - Aaron: who are the top advertisers over time?

        - Josh: Does the frequency of political advertisements change over time?
        - Nick: Who are the most common donors?
        - Fehmi and Paul: Do advertisers and target demographics change over time?
- Project presentation
    - Everyone is fine to present.

Moving Forward

- Aaron will merge two datasets on similar columns and include a column to indicate which dataset the ad came from
- Aaron: who are the top advertisers over time?
- Josh: Does the frequency of political advertisements change over time?
- Nick: Who are the most common donors?
- Fehmi and Paul: Do advertisers and target demographics change over time?
- Meeting with Matt Lepinski:
    - Friday 3-4pm, alternative time 10am-11pm
    - Josh will reach out and send calendar invites

Notes

- Group will have considerable progress done by next Monday meeting

Monday, April 17, 2023

Group Meeting

Notes

- Data description:
  2020 dataset:
  https://www.propublica.org/datastore/dataset/political-advertisements-from-facebook

- Goals of the project:
    - Are there any interesting trends in political advertisement on Facebook?
- Initial ideas:
    - Do the frequency of ad political leanings change of time?
    - We can attempt to classify institutions that pay for the ads and see if there is a change in frequency?
    - How many ads per week/month?
    - How many ads are targeted? Who tends to be targeted? Does this change over time, by ad type, or by funding institution?
    - How do the two datasets differ?
    - Common donors, common areas of focus
- Initial Planning Document:
    - learned from Project 1.
        - Keeping scope of the project in mind
        - Be mindful of work bottlenecks

    - Rush Data collection
    - Have multiple people check the validity of the data
    - Use informative file/variable names
- List a few things that you are doing differently in this project than in the previous project in light of the lessons you have learned.
    - More regular, frequent meetings
    - Not spending too much time on “planning”
    -
- Initial high-level plan of attack.
    - Data munging
        - collecting data into one large dataset, matching variables
        - Extract targeted ad information from json
        - Check for missing values or bad data
        - Compare two datasets
    - Analysis
        - Time series analysis - categorical data, frequencies
- Who is doing what, initially.
    - Nick: Pyspark,
    - Aaron: data visualization,
    - Josh: time series analysis
    - Fehmi:Collecting and formatting data on targets
    - Paul: EDA
- Figure out what columns are useful

Issues with dataset:

Missing data
Missing time periods
Self selecting group (those who chose to install extension)
Only 3000 individuals
Inconsistency in data collection (Poor labeling on Facebook’s end)

Propublica:
Mid 2017 - mid 2019

Adobserver:
Dec 2020 - Feb 2022
Gap 3/2021 - 7/2021