W205-1 Spring 2016 Final Project Pipeline and Architecture

Alfred W. Arsenault

Project Overview

- Provide a graphical analysis of financial contributions to significant Presidential candidates
  - "Significant" = got more than 200 separate contributions
    - There are over 200 registered Presidential candidates, many obviously jokes
- Show who is giving money to what candidates

Data Source

- Files are on the Federal Election Commission's website, www.fec.gov
  - FTP'able directories: ftp.fec.gov/FEC/2016
    - ZIP'ed files that extract to pipe-delimited text files
    - Candidate master file – cn16.zip
    - Committee master file – cm16.zip
    - Individual donations file – indiv16.zip
    - Committee-to-committee, "pass-through" donations file – pas216.zip
    - Data dictionary for each file available, e.g.
      - http://www.fec.gov/finance/disclosure/metadata/DataDictionaryCommitteeMaster.shtml for Committee Master file

Pipeline

- Python script to FEC's FTP directory, download .zip files
- Extract .zip files, resulting in pipe-delimited text files
- Python script to read in pipe-delimited files, extract columns we need, and write them to .CSV files

Pipeline (2)

- OpenRefine used to clean .csv files
  - E.g., delete records for “non-serious” candidates; only retain records for candidates who have received more than X donations (initial X value = 100)
- Create .csv files of nodes and edges for passing to Neo4j

Graph Analysis

- Nodes:
  - 1. Candidates
    - Properties: Name, Candidate ID, Party, Committee ID(optional)
  - 2. Committees
    - Properties: Name, Committee ID, Candidate ID (optional)
  - 3. Individual Donors
    - Properties: Name, Employer
- Edges:
  - Donated_to (each donation is an edge)

Additional Work

- If I have time:
  - Edges would reflect the size of the donation. A donation of $10 shows up as identical to a donation of $2700.
  - Include past elections to show a trend of donations
  - Group employers into “industry” values to show donations by industry