Fortune 500 Campaign Influence

Alden Golab & Paul Mack

Final Report

CS 123 - Spring 2016

- **Dataset:** Sunlight Foundation State and Federal Campaign Contributions 1990 - 2012 (In Two Year Increments Corresponding to the Election Cycle); 15.1 GB with 50,733,830 lines corresponding to individual transactions.
    - *Data Dictionary*: http://data.influenceexplorer.com/docs/contributions/
- **Hypothesis:**
    - Corporate money is often believed to play an important role in political campaigns. Corporations give to political campaigns directly, and provide salaries for their employees some of whom use their salaries to donate to campaigns. Does the influence of corporations on election donations work in the same direction as that of their employees? When we think of corporate money in politics, should we consider that some of this money may be counteracted by the individual donations of the employees of these companies or should we consider that this may even be reinforced?
- **Algorithms Used:**
    - **Dataset Sampling:**
        - Mapper that randomly selects 1/100 of the entries in the full dataset and exports to a separate csv.
        - **Runtime**: Linear run time on 5.7 million records would have taken 10 hours locally on two nodes.
    - **Simple Aggregations:**
        - ***Total Donations from Individuals and PACs:*** Sum the donations for an entity across time filtering for whether the entity is type individual or type PAC
        - ***Unique Donor/Recipient Pairs****:* For each entry, takes the recipient and donor, then reduces across all nodes for unique entries.
            - Did not execute entity resolution on this step, but was planned for Network analysis until we ran out of time and AWS budget.
        - **Runtime**: Linear run time on 5.7 million records would have taken 10 hours locally on two nodes.
    - **Entity Matching:**
        - For each entry within the large dataset, we run matching between the transaction donor and/or recipient employer (depending on which is present) against the authoritative names of Fortune 500 companies. Once a match is hit (with a similarity score of over 85), the entry donor/employer is recorded in a JSON mapped to the correct Fortune 500 authoritative name. A JSON is then outputted by MapReduce to be used for data processing purposes. Due to time and money constraints, we decided to only run this algorithm on a 1% sample of the data.
        - **Runtime:** estimated to take 500 hours on two nodes locally for all 5.7 million records, assuming linearly scaled runtime. However, based on a few local runs, we determined that runtime grows non-linearly and would likely have been higher.
    - **Matched Entity Aggregation (one for corporate, one for individuals):**

        - We pass a JSON up to each MapReduce instance for name lookups against authoritative values. For each instance, a lookup is done on the JSON to match the entry with the corrected Fortune 500 name. The entry is then cleaned extensively, with special attention toward unicode elimination, so that it can be used locally for analysis.
        - Linear run time on 5.7 Million records would have taken 10 hours on two nodes locally.
- **Big Data Approaches Used:**
    - AWS EC2/S3
    - Elastic Map Reduce
- **Big Data Techniques Not Covered in Lecture:**
    - **MRJob Protocols**: formatting MapReduce output
    - **MRJob Configure Options**: exploited built-in MRJob functionality for JSON passing
    - **Edit Distance**: calculating the similarity for two strings using fuzzywuzzy's implementation of Levenshtein edit distance
    - Cleaning unicode encoded data
- **Challenges & Solutions**
    - Handling Unicode. Created runtime errors. Also, since we were doing entity resolution - we need to make sure we handled unicode consistently across scripts as if they were handled differently it could create further discrepancies in names. It was difficult to tell where unicode was being introduced - if it existed in the original input file or was being introduced as we used different MRjob output protocols
    - Balancing entity reduction. Removing common words increased performance for most of the entities, but could decrease performance for other companies like AT&T which were very short and by removing "Corporation" you were removing the majority of the string. Removing the common words also seemed to be computationally expensive.
- **Results:**
    - **Entity Resolution:** On a random sample of 1% of the data, we found 2,000 aliases for 444 unique companies out of the 500 listed in the fortune 500 list.
    - **Hypothesis Testing:**
        - **Corporations tend to give more to everyone in general, but their employees tend to give more to Right-aligned PACs and Campaigns, which means that Fortune 500 companies tend to agree more with their employees in supporting Right-aligned candidates.**
        - In general, Fortune 500 companies far out-spend their employees when it comes to the campaigns they support, of any stripe. We did find that, over time, Fortune 500 companies have consistently spent more in support of Right-aligned organizations and campaigns; however, this is not entirely true for their employees. In particular, the 2008 Obama campaign far out-raised the McCain campaign from Fortune 500 employees. There are clearly times when the spend by corporations are in opposition to the money donated by their employees in political campaigns.
        - When it comes to spend per common recipient, Fortune 500 companies tend to far outweigh their employees in support of any candidate (ratio tends to be 1:3 or 1:2 for each candidate in terms of corporate:employee support). Surprisingly, for Left-aligned campaigns and PACs, Fortune 500 companies tended to outspent their employees in higher proportion than for Right-aligned campaigns and PACs (approximately 1:4 employee:corporate spending on Left-aligned vs. 1:3.5 employee:corporate spending on Right-aligned).

        - This actually indicates that Fortune 500 corporate donations are far more likely to agree with their employee's dollars in support of Right-aligned candidates rather than Left-aligned candidates.
- **How did we find our dataset?**
    - Our original project was based on data we received from Professors from whom we work. After encountering issues with the kind of data encoding that was used (Microsoft Access is bad news) we moved to Sunlight Foundation campaign data. We knew of Sunlight prior to CS123 through various speakers that have come on campus.