Databasing for a Presidential Campaign

BIDEN
PRESIDENT

Michael Chen, Rebecca Driever, & Rayna Ji

* This is an exercise in creating a database for a campaign, **not** a reflection of any individual's political views and **not** an endorsement of any candidate for President

Table of Contents

Executive Summary

Each year hundreds of millions of dollars are donated by individuals, corporations, and organizations in an attempt to get their ideal candidate into the White House. Millions of dollars end up being donated to candidates who won't get the chance to be on the ballot in November, as they first need to win their parties' primary elections in order to secure the nomination.

While campaigns often rely heavily on data, there are many different sources that are tracked, including finances, media coverage, social media mentions, polling numbers, and more. With this in mind, our team set out to create a comprehensive database for Former Vice President Joe Biden's campaign to use throughout the remainder of the primaries as well as perhaps the remainder of his presidential campaign. Using publically available data from the New York Times, the FEC, and delegate counts up through Super Tuesday, our team scraped, cleaned, and organized information into a MongoDB database. Due to wanting to track large amounts of text information such as complete articles from the New York Times in addition to API information that is already in a JSON format, our team chose MongoDB for its scalability and flexibility. With no need to pre-define a schema, importing articles of varying lengths is extremely simple.

Maintaining this database through the use of simple APIs and reproducible web-scraping will allow the Biden campaign to keep their data organized and ready for analysis. Rather than campaign staffers spending time finding the articles written about the campaign, they can spend their time analyzing what is said and making data-driven decisions.

Background & Context

On April 25, 2019, former Vice President Joe Biden announced his bid for the Democratic Party candidacy in the 2020 Presidential election[1]. Prior to the first ballots of the primaries being cast, Biden was thought to be a strong front-runner alongside Bernie Sanders[2]. However, disappointing finishes in Iowa and New Hampshire made the press question whether his campaign could make it to Super Tuesday -- a day when ⅓ of the total delegates are decided. As a former VP, Biden entered the race with a lot of name recognition. However, name recognition is not always the ticket to the nomination, as his former boss Barack Obama proved in the 2016 Presidential race.

Securing the nomination is said to be based on the three M's: money, momentum and movement (organization)[3]. Most campaigns keep track of their money very well - where it's coming from, who it's coming from, and where it's being spent. Movement is often determined by knowing one's own candidate base and knowing where a campaign should set themselves up ahead of primaries in order to have the greatest impact on votes. Momentum is difficult to predict, but can be inferred by monitoring news outlets and public opinion. These three important pillars of a campaign are most likely the responsibility of many individuals with data in many different places. The goal of this project is to create a single database that can store information on each of these components - as well as opponent information where available - so that campaign analysis can be done easily and important data is all together in one place.

Introduction to Data & Data Sources

New York Times API

When tracking media mentions in an effort to understand primary election performance, it is important to understand what sources of information potential voters are viewing. Many Democratic primary voters trust sources such as the New York Times and NPR for their political news. With the availability of an API to get metadata on all articles published, the New York Times is used as a datasource here to track official news. The API was used to get metadata for each article published in February and the first two days in March -- the days before Super Tuesday. Web-scraping was then performed to obtain the body text of each article.

Information available through the API includes:

- Abstract
- Byline (author(s))
- Headline
- Keywords
- Publication date & time
- Section
- Multimedia links (images, video, etc.)
- URL

FEC Election Filings

An important part of campaigning is fundraising to employ staff, cover travel costs, and advertise. All campaigns are periodically required to file reports with the Federal Election Commission about how much money they receive, who and where they receive it from, how much they spend, and where they spend it. A .csv file containing information on all individual contributions to each 2020 campaign was downloaded from the FEC website, based on the most recent filings which occurred January 31st 2020.

Delegate Counts

Presidential primary campaigns all boil down to one thing -- who can secure enough delegates to secure the party nomination for President. The methodology used by Democrats to determine delegate count is much more complex than Republicans, who use a “winner takes all” approach. Due to this calculation as well as delays in vote counts from things such as absentee ballots, official delegate counts may not be known until weeks after a primary election has occurred. Our team decided to use the New York Times’ “Democratic Delegate Count and Primary Election Results” page, which is updated frequently to display the number of delegates and votes for each candidate as official numbers become available.

For the purposes of this project, data prior to Super Tuesday is scraped and added to the database.

Explanation of Database Design Choice

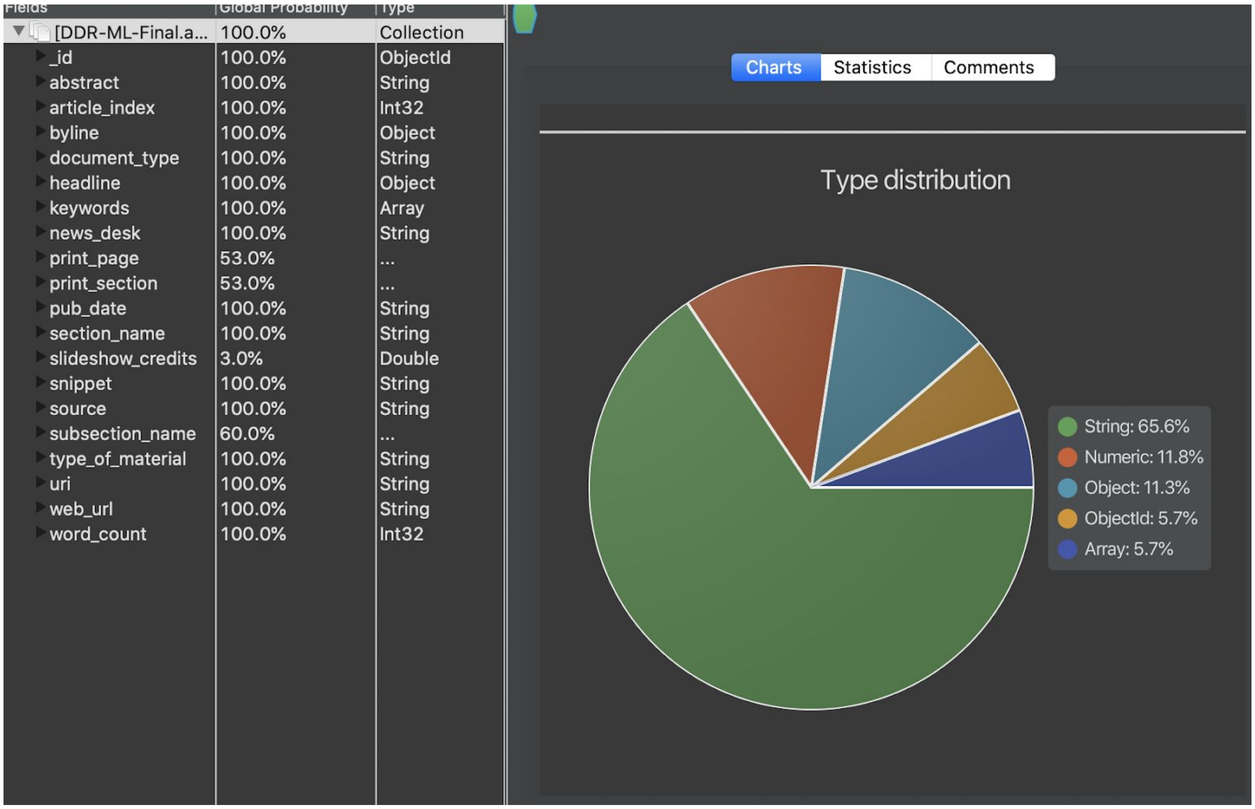
After acquiring the data from multiple data sources, where to store them so that it becomes quickly retrievable becomes the major decision to engineer our data pipeline. MongoDB and MySQL are the top two options proposed by the team.

MySQL is part of the relational database management system (RDBMS) family. MySQL databases generally use structured data and structured query language to define and manipulate data. The main advantage here is extremely quick and powerful, where tables within the database are often related to each other, secure and can be wrangled across from easy to very complex. However, it brings in the challenge that the data has to be prepared into the form of relations. This would require significant up-front preparation and data sanity check.

MongoDB is part of the Non-relational database family. Unlike MySQL, the tables and the data are not structured. They are essentially organized in a key-value fashion within each document. Similar to MySQL, writing queries in MongoDB can be super fast and reliable thanks to the built-in indexing. The main advantage here is the flexibility, where no predefined structure is needed, each document can have its own unique structure, and schemas can organically grow in both horizontal and vertical directions.

For our project, most of the data is text from different sources, which results in the challenge to effectively define schemas. News articles from the New York Times vary greatly in word count, so a SQL database where data types need to be predefined may present restrictions in adding articles of various lengths. In addition, by using an API, the New York Times provides a large amount of metadata for each requested article. Although the campaign funding data and delegates information of Bernie Sanders and Joe Biden is nicely tabulated and concrete, to work with the large amount of text data requires them to be on the same database for faster query and easier management. Lastly, in terms of study of raw text, natural language processing and modeling are the most common approaches. When running different algorithms, multiple features can be created, in which some are single points while others are lists or matrices. This would also require a flexible schema development to accommodate various analyses for such great amounts of data.

Given the complexity of text data, anticipated wrangling and feature engineering, and potential growth of the project, we chose to use MongoDB to store the data and integrate multiple data sources. We decided to have five collections in total and they will store the data from the NYT API, web scraping and the FEC .csv. For the New York Times articles, we stored main content and the metadata separately to have more clarity and to have the pipeline process clean. This two schemas have the common article index key-value pairs, which is the position of an article saved locally after request from API, so that they can be used for joint analysis in the future. The FEC filing information and the delegates status from each state are stored in two schemas to maintain clarity of respective data sources, whereas the candidate information is included in both. Below is an example of a schema for news articles' metadata.



Fields	Global Probability	Type
[DDR-ML-Final.a...	100.0%	Collection
_id	100.0%	ObjectId
abstract	100.0%	String
article_index	100.0%	Int32
byline	100.0%	Object
document_type	100.0%	String
headline	100.0%	Object
keywords	100.0%	Array
news_desk	100.0%	String
print_page	53.0%	...
print_section	53.0%	...
pub_date	100.0%	String
section_name	100.0%	String
slideshow_credits	3.0%	Double
snippet	100.0%	String
source	100.0%	String
subsection_name	60.0%	...
type_of_material	100.0%	String
uri	100.0%	String
web_url	100.0%	String
word_count	100.0%	Int32

Figure 1. Schema analysis for the collection of NYT news articles' metadata.

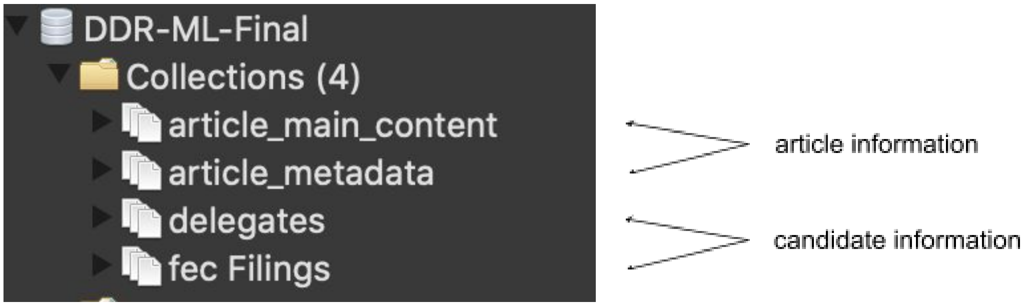

DDR-ML-Final
Collections (4)
article_main_content
article_metadata
delegates
fec Filings

Figure 2. Collections in MongoDB

Business Value Created

Building a database in MongoDB that tracks Joe Biden's campaign information such as delegates won and contributions secured, as well as media mentions, allows for a complete view of the campaign and its progression. Financial information is typically tracked in a structured fashion such as a database for easy analysis, however media coverage is not. Most likely, one individual is responsible for monitoring reporting sources such as the New York Times and NPR, while another is responsible for monitoring social media trends and opinions. These individuals will often need to go directly to the media outlets and manually search through a large amount of information. Maintaining this database through the ease of using APIs and web-scraping will allow for a large portion of this important information to be aggregated all in one place, saving individuals time and allowing them to focus more on what is being written.

More clearly, this database allows easy monitoring of candidate mentions in the news. There are over one thousand New York Times articles written each week, especially during campaign season -- in the month of February there were over 700 articles that mentioned either Joe Biden or Bernie Sanders. Connecting to the New York Times API and easily pushing the JSON response for only articles that mention one of the candidates into MongoDB creates no need for a campaign staffer to physically search and sort through the thousands of articles written every month. Even if a piece isn't a headline right away, if it mentions the candidate it will be added to the database where the content, author, and other information can be accessed within seconds.

With the aggregation of all of this information in one convenient place, our team was able to also include the same information for the Bernie Sanders campaign. With the time saved for staffers by not needing to physically sort through media output, they can use the information on the Bernie Sanders campaign to better understand their competition. Any analysis that can be performed on articles, delegates, or financial contributions regarding the Biden campaign can just as easily and efficiently be performed on the Sanders campaign. This can provide a great advantage to the Biden team if utilized.

Conclusion

By nature, political campaigns are a combination of multi-dimensional flows of information. To gauge the activities of campaigns, there are traditional sources like TV Ads, mail cards, and newspapers, online platforms such as digital newspapers and twitter, and operational data such as funding details, policies, rally details, etc. Running a campaign becomes harder and harder with the magnitude of information available. In this report, we tried to integrate the most accessible yet complete and distinct data sources, and designed the proper database with MongoDB to both manage and conduct analysis. The entire data pipeline is engineered with automation and customization. The advantages of the database design are that it is easily scalable, secure, reliable and fast. It is easy for campaign staff with limited technical background to use and support data-driven reporting, decision making and budgeting optimization. A more comprehensive overlook via understanding from major media, social media, donations, and campaign's performance would be of tremendous help to the candidate.

Sources

1. “Joe Biden announces he is running for president in 2020”. Arlette Saenz. 25 April 2019. https://www.cnn.com/2019/04/25/politics/joe-biden-2020-president/index.html

2. “Biden Is The Front-Runner, But There’s No Clear Favorite”. Nate Silver. 10 January 2020. https://fivethirtyeight.com/features/biden-is-the-front-runner-but-there-is-no-clear-favorite/

3. “Joe Biden Picks Up Substantial Winds In Super Tuesday Contests”. Scott Detrow, Mara Liasson. 4 March 2020. https://www.npr.org/2020/03/04/812000142/joe-biden-picks-up-substantial-wins-in-super-tuesday-contests