dataBASED

Analyzing Press Coverage of the 2023 Chicago Mayoral Election

CAPP 30122

Winter 2023

Abe Burton (abejburton)

Kathryn Link-Oberstar (klinkoberstar)

Lee-Or Bentovim (bentovim)

Maddie Roberts (mkroberts)

Table of Contents

I. Project Abstract

The project aims to analyze the press coverage of the Chicago mayoral race and investigate how candidates are covered differently in the media. The study uses *word frequency* to identify which topics are brought up most often in relation to specific candidates, and *sentiment analysis* to examine the tone of the articles. We examine sentiment by candidate overall, by paper, overall and by the candidate in each paper.

The goal is to determine if specific candidates are mentioned more frequently in relation to specific topics than others and if there are any patterns in the press coverage that may indicate bias. The project will use data from various news sources that attempt to represent a cross-section of perspectives throughout the region to provide a comprehensive analysis of the press coverage of the mayoral race. The study examines articles from the day of the first candidate announcement (April 11, 2022) up to election day) that mention at least one of the 7 candidates who remained in the election until election day[1]. It uses articles from the Chicago Tribune, the Defender, Lawndale News, the Hyde Park Herald, the Triibe, and Crain's Business Journal.

[1] The following candidates remained in the race until election day. Kam Buckner, Jesús "Chuy" García, Ja'Mal Green, Brandon Johnson, Sophia King, Lori Lightfoot, Roderick Sawyer, Paul Vallas, Willie Wilson. Candidates who dropped out are not examined in this study.

II. Structure of Software

The project is structured in the following sections:

- Database with 3 tables: Candidate Info, Newspaper Info, and Candidate Names (/data)
- Data Collection (/data, /scrapers)
- Cleaning (/cleaning)
- Data Analysis (/analysis)
- Visualization (/visualization)

Data Collection

Article data was gathered through web scraping and the Proquest database. Articles were gathered for the time period from April 11, 2022, the date of the first candidate announcement, to February 28, 2023, election day.

Scrapers

The Scrapers module scrapes articles from Lawndale News, The Defender, The Hyde Park Herald, and The Triibe. The module uses candidate name tokens from the database to search web pages and scrapes all articles published during the study period.

Proquest API

Articles from the Chicago Tribune and Crain's Business Journal are gathered through the Proquest database. Files were compressed within the Proquest virtual environment to stay within the file download limits. The module assigned articles to candidates, and one article could be assigned multiple times if it applied to multiple candidates.

Cleaning

The Cleaning module is responsible for cleaning the gathered data. The module strips stop words, normalizes case, and selects only sentences that refer to the candidate that is the subject of the article. The module deduplicates the articles, so articles can exist more than once, but only once per candidate.

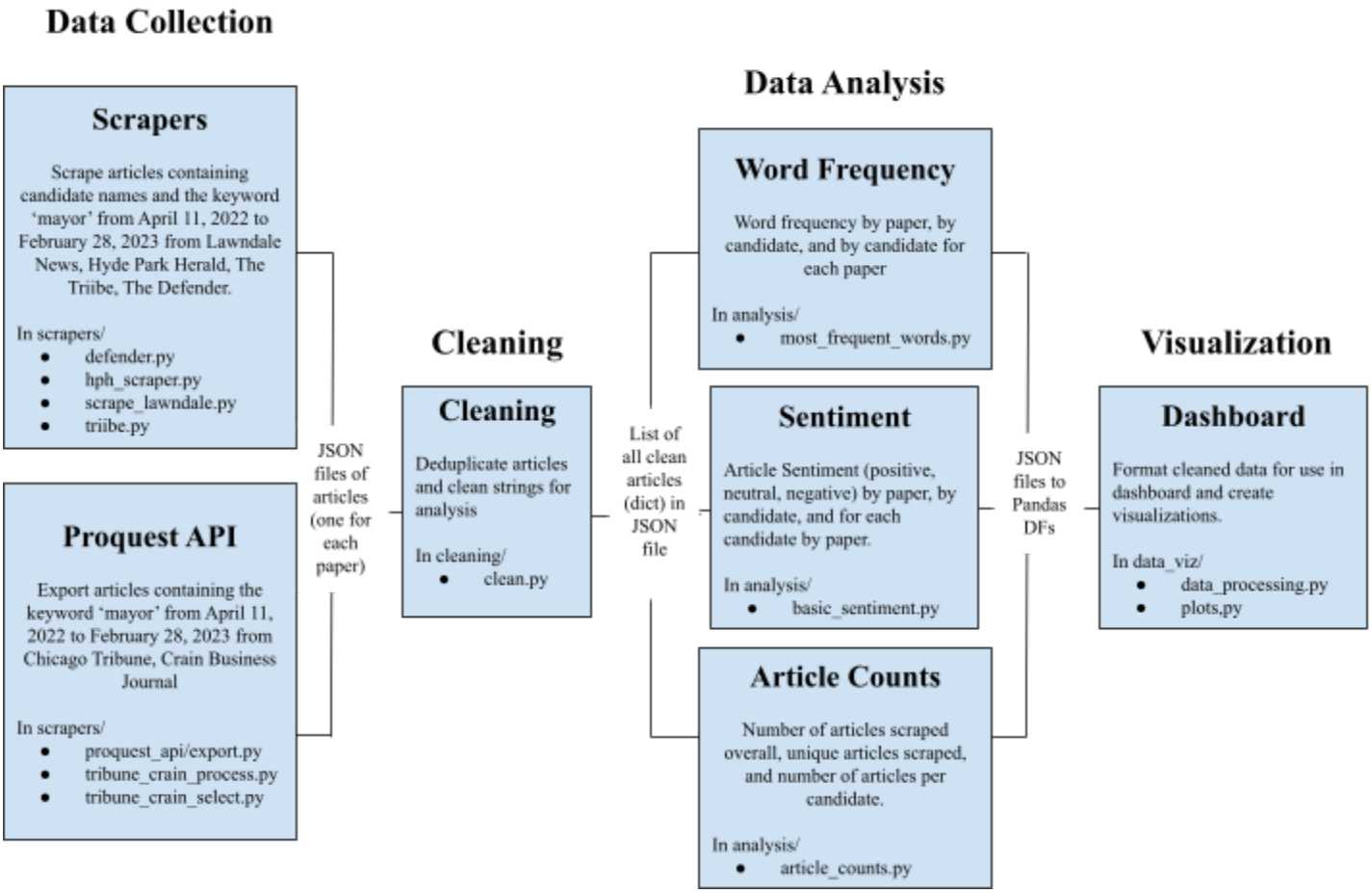
Analysis

The Data Analysis module conducts three types of analysis: word frequency, sentiment, and article counts. These three tiers of analysis are conducted by the candidate, by the newspaper, and by the candidate within each paper.

Visualization

Finally, the Visualization module creates a dashboard of findings using Dash and Plotly. The module presents the results of the analysis in an interactive and user-friendly way, allowing users to explore the data and insights easily.

III. Project Diagram



IV. Code Responsibilities

Database

- Abe set up the database and wrote helper functions to access the database

Data Collection

- The team worked together on the design for the inputs and outputs of the scraper
- Maddie wrote the scraper for Lawndale news.
- Abe wrote the scraper for Hyde Park Herald.
- Lee-Or wrote the scrapers for the Defender & the Triibe.
- Kathryn wrote the function to export and process data from the Proquest API for Crain and the Chicago Tribune.

Cleaning & Analysis

- Maddie and Kathryn worked together on the overall structure for cleaning and analysis.
- Kathryn wrote the cleaning module.
- Maddie wrote the analysis module.

Data Visualization

- Abe and Lee-Or designed the dashboard and graphs together.
- Abe and Lee-Or processed the clean data to be formatted for visualization.

Honorable Mentions

- Abe set up poetry, and the GitHub repository, and provided ongoing support resolving GitHub issues as they arose.
- Maddie set up Main so the project can run from the command line.

V. User Guide

Installation

Note can only be run with:

1. Install Poetry to Local Machine
2. Clone the Project Repository via SSH:
   git@github.com:uchicago-capp122-spring23/databased_project.git

Install Virtual Environment and Dependencies:
poetry shell
poetry install

Usage

Project must be run in the Poetry virtual environment. Upon completion of above installation requirements and within project terminal, and on each subsequent rendering of project, initialize virtual environment by running:
poetry shell

Execute the project by running:
python -m databased
This command may take a minute to load project to terminal.

You are then prompted to enter a singular-digit command to execute a portion or the entire project, as seen below.

To execute a desired aspect of the project please enter one of the following commands:
1 - Open Data Visualization
2 - Scrape All Newspapers
3 - Clean Scraped Data
4 - Conduct Data Analysis
5 - Run Entire Project Start to Finish (Scrape -> Clean -> Analyze -> Visualize)
6 - End Program

Please input the number of your desired command:
example: "1[Return]" will run the data visualization.

Command 1 - Opens Data Visualization

Renders a Dash to visualize the final results of the dataBASED project.

Notes:
This command will take about 1 minute to render Dash.
Dash will throw a warning "This is a development server," this error is fine.

Command 2 - Executes All Scrapers/Proquest API

Runs all scrapers and Proquest API to collect newspaper articles about Chicago's mayoral candidates. The retrieved data is then stored in JSON format and outputted to the databased/data folder.

Note: This command will take about 20 minutes to complete.

Command 3 - Executes All Data Cleaning

Runs data cleaning on all scraped data; strips stop words, normalizes case, and selects only sentences that refer to the candidate that is the subject of the article. The cleaned data is then stored in JSON format and outputted to the databased/data folder.

Note: This command will take about 1 minute to complete.

Command 4 - Execute All Data Analysis

Runs data analysis on cleaned candidate data to calculate word frequency, sentiment, and article counts for the candidate, the newspaper, and for the candidate within each paper. The results are outputted to JSON files within databased/analysis/data folder.

Note: This command will take about 12 minutes to complete. However, if you comment out lines 54 and 55 in basic_sentiment.py the command will execute in about 1 minute. The completion of the JSON for overall newspaper sentiment will be prevented as a result of this.

Command 5 - Execute Entire Project

Runs entire project start to finish. Runs scrapers/Proquest API, then cleans article data, conducts data analysis, and renders the visualization of results.

Note: this command will take about 45 minutes to complete.

Command 6 - Close Project

Terminates python scripts.

If you encounter issues with nltk or pyarrow please run the following commands within the poetry shell:

python3 -m pip install nltk

python3 pip install pyarrow

Upon extensive testing, sometimes a scraper will become blocked by servers. If this occurs, run program again and they should run completely.

VI. Goals and Findings

The goal of the project was to analyze the media coverage of different mayoral candidates in the 2023 Chicago Mayoral election. The hypothesis was that there would be greater differentiation in coverage between newspapers, indicating a preference for certain candidates based on the readership.

However, the actual results showed that there was not much differentiation in coverage between newspapers with different audiences. The coverage was highly standardized across papers, and there was no indication of differential coverage based on the readership of the paper.

Despite the lack of differentiation between papers with different audiences, the project uncovered interesting insights into how the candidates were covered overall. There were differences in tone, language, and topics discussed, providing useful information about the media's portrayal of the candidates.

To explore how the mayoral candidates were covered in the press, we scraped and cleaned data, analyzed the results, and provided visual exploratory analysis. We successfully completed all stated goals of the project. However, if more significant differences in coverage had been found, we would have liked to do further analysis to determine contributing factors.

Overall, this study provided valuable insights into how the local press covered Chicago's mayoral candidates. These insights have the potential to inform future research and enhance our understanding of how the media shapes public opinion in local elections.