ORIE 4741 Data Analysis Project Proposal

Joshua Caplan (jc4567), Benjamin Luckow (bjl93)

Repository Link: https://github.com/bjluckow/ORIE4741-Project

Problem, Techniques, and Goals

The purpose of this project is to gain insight into US electoral tendencies by county by building a national print and digital media dataset from the ground up. Using a proprietary webscraper and pre-trained Python sentiment analysis models, we seek to generate a large article database including text bodies, author, date, and other publishing metadata. We will then transform this text data into a form loosely suitable for modeling via feature engineering; by evaluating keywords, and leveraging NLP libraries for sentiment analysis and topic classification, we hope to categorize data by support for political issues and then correlate this data against county voting tendencies captured in US electoral and demographic datasets available through Kaggle.

We hope to explore potential uses of news data to predict counties' voting results: one such technique we hope to explore is the use of binary classification on county data trained on historical voting results to determine whether a set of articles from a given county is expected to vote Democrat or Republican in an election cycle. We will also explore how to engineer our article data to transform it into a form conducive for visualization and analysis, as well as various methods to classify counties (e.g., logistic regression).

Data Collection

While outside the scope of this project, we hope to use a highly-configurable news webscraper built in TypeScript to generate an initial dataset of articles capturing text content and metadata, including the region of the publication within the United States. This will involve manually configuring the scraper for each publication, which may number in the hundreds to thousands, but will allow for fault-tolerant mass data collection and provide consistent and complete text data for each article. After transforming and analyzing this initial dataset, we will also make use of datasets such as 2012 and 2016 voting data available on Kaggle.

While in-depth natural language processing is also outside the scope of the project, we hope to parametrize articles and counties into data points using more free-form techniques, such as evaluating the articles of a certain region containing certain text keywords using model libraries such as VADER sentiment analysis to determine support for certain policies and issues in a numerical form.