Data Handling: Import, Cleaning and Visualisation

Lecture 4: Research Insights

Philine Widmer
Prof. Dr. Ulrich Matter
(University of St.Gallen)

14/10/2021

---

1 Part I: A Spatial Model of Online News Provision

1.1 Project abstract (broader background)

Increasingly, news are consumed online. At the same time, we know little about the key players (i.e., news outlets) that provide news content online. For example, how concentrated is the online media market in different countries? There is a broad consensus in the social sciences that a pluralistic media landscape is of great importance in a democratic society.

1.2 Research questions and approach

The overall research questions we pose in this project are *Who are the key providers of online news and how large are their market shares?*. In our research paper, we also analyze how diverse the content provided by different news outlets is (but we do not discuss this part of the research project in this class).

1.3 Challenges and key data handling concepts

One main challenge in this project is the semi-structured nature of the raw WHOIS registrant data. This involves a lot of cleaning, e.g., using regular expressions regular expressions. Another challenge is the matching of the registrant data with the established firm data from Orbis (often, companies are referred to in slightly different ways, e.g., Ringier AG vs. Ringier). Also, there may be different firms with the same name. In such cases, the data cleaning also involves considering contextual knowledge.

1.4 Pipeline overview

1. Fetch raw data from `abyznewslinks.com`.

2. For every news website, find out who registered the website by querying WHOIS registrant information (see examples in video).
3. Clean and structure the WHOIS registrant data.
4. Match the cleaned registrant data with established firm data from the Orbis database (www.unisg.ch/en/universitaet/bibliothek/recherche/datenbanken/a-z/o/orbis).
5. Query reach data for each news outlet (i.e., for each domain) from Amazon's Alexa Web Information Service (proprietary data).
6. Compute the Herfindahl-Hirschman index (HHI) values.

The key data handling skills/knowledge needed for this pipeline are: basic programming (loops, inexact string matches, etc.), data analysis in R, basic web technologies (scraping, API, JSON).

1.5 Economic concepts used in this project

- Herfindahl-Hirschman index (HHI)

2 Part II: Web Search Personalization and Political Polarization

2.1 Project abstract (broader background)

There are popular concerns that the internet is a major driving force behind increasing political polarization, as citizens are increasingly likely to solely consume information from like-minded sources, fueling their prejudices and grievances. We design a field experiment based on 'bots'(programs that simulate internet users) to test whether and how the algorithmic personalization of web search results can lead to online segregation and political polarization. Like human users, each bot has its own browser fingerprint, residential IP address, and browser settings. In the experiment, we initiate a sample of several hundred bots and randomly treat bots with varying political preferences and varying degrees of privacy affinity (which privacy invading technologies bots block when browsing, e.g., cookies). While the bots are building up browsing and search histories over several weeks, we test (i) whether bots using the exact same search terms get systematically different search results, (ii) whether these results are politically flavored (in line with the bots' political preferences) and (iii) to what degree a potential political bias in the search results is diminished by the bots' privacy affinity.

2.2 Research questions and approach

The overall research question we pose in this project is *Can web search personalization contribute to political polarization?*. We think about the overall question by splitting the question up into three sub-questions:

1. Is the supply of online news polarized (in the US)?
2. Does web search personalization generate polarized results?
3. Do consumers change their ideological stance due to what they see online?

In this project we focus on sub-questions 1 and 2 with the following approach:

1. In a first step, we index hundreds of US online news outlets on a liberal-conservative scale.
2. Building on the result in the first step, we conduct a field experiment with simulated web users (bots) in the US. (This includes human user emulation: residential IPs, browser fingerprints, human-like typing/cursor movements.)
3. In order to make point 2 possible, we develop a web application to conduct large-scale bot-based field experiments.

2.3 Challenges and key data handling concepts

While the data analysis in step 1 is based on rather simple statistics/econometrics (χ^2-Test, linear regression; see Gentzkow and Shapiro (2010)), the data handling necessary to make these analytical steps possible is rather complex for the following reasons: raw data is collected from various sources in various formats, the main data input is *raw text* from congressional speeches and tweets (unstructured data). The project illustrates how organizing and processing raw text in order to use text as data is in many ways linked to key data handling concepts such as storage in csv files, data cleaning, data filtering etc.

The *data pipeline* concept is a very helpful tool to organize the different parts of this part of the research project, covering (in several parts and many individual R-scripts), the entire process from gathering the raw data to the visualization of the ideological distribution off online news outlets in the United States.

2.4 Pipeline overview

The following lists of tasks illustrate what steps are involved in the overall data pipeline and points to individual R scripts that handle some of these steps and create intermediary results in the form of csv-files. Each step is using the output of the previous step as an input and creates an output for the next step.

2.4.1 First part

1. Fetch raw data from `congress.gov`.
2. Organize fetched speech data in CSV-file. (still raw, uncleaned data)
3. Process raw speech data (add party info, basic cleaning; `code/cr_speeches.R`)
4. Compute bigram frequencies per congressperson (and party) `code/cr_byspeaker_2gram.R`.
5. Clean bigrams, compute bigram partisanship `code/cr_partisan_bigrams.R`.
6. Select the 500 most partisan bigrams.

The key data handling skills/knowledge needed for this part are: data structures, csv, cleaning/mutating, filtering, basic programming (loops), data analysis with R.

2.4.2 Second part

1. Fetch bigram frequencies from news articles and news headlines.
2. Compute online news outlet ideology (`code/all_outlet_ideology_gdelt_gfg.R`)

The key data handling skills/knowledge needed for this part are: basic programming (loops etc.), data analysis in R, big data (Google BigQuery), basic web technologies (API, JSON).

2.5 Statistical methods used in this project

- Chi-squared Test
- Linear regression
- n-gram

References

Gentzkow, Matthew, and Jesse M Shapiro. 2010. "What Drives Media Slant? Evidence from US Daily Newspapers." *Econometrica* 78 (1): 35–71.