Selecting and Cleaning Texts

HSS 510 / DS 518: NLP for HSS

Taegyoon Kim

Mar 4, 2025

Agenda

Things to be covered

- Approaches to and principles of selecting texts
- Basic terms in NLP/text-as-data
- Guided coding
  - String manipulation (Python)
  - Regular expression (Python)

Two Approaches to Selecting Texts

1. Question-driven

- Already have a well-defined question
- Identify and collect texts that can best answer the question
- Highly effective *if* you have a question already

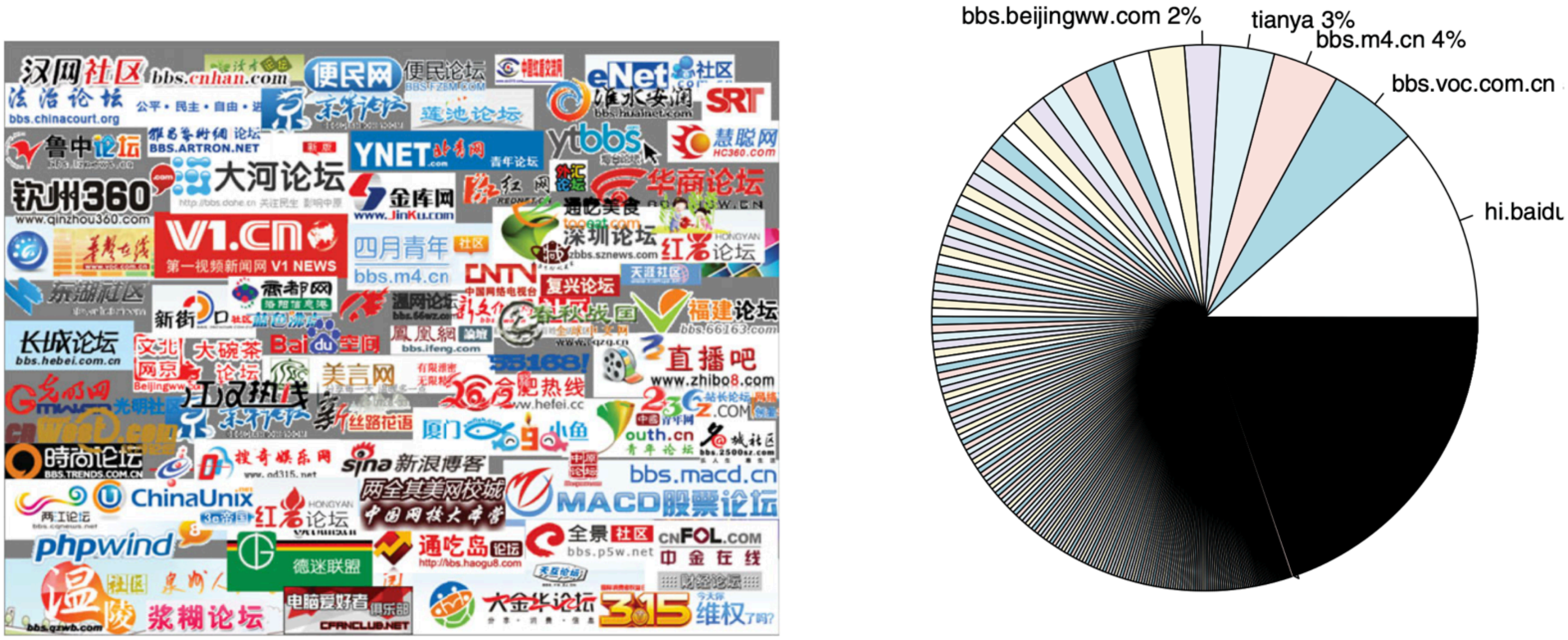
Two Approaches to Selecting Texts

1. Question-driven

- E.g., King et al. (2013)
  - Question: “what is the primary purpose of the online censorship program in China?
  - Data: millions of social media posts originating from +1,400 different social media services
  - “Users should be able to express themselves fully **prior to** potential censorship”

Two Approaches to Selecting Texts

Figure 1. The Fractured Structure of the Chinese Social Media Landscape



(a) Sample of Sites

(b) All Sites excluding Sina

All tables and figures appear in color in the online version. This version can be found at http://j.mp/LdVXqN.

Two Approaches to Selecting Texts

1. Question-driven

- E.g., King et al. (2013)
  - Method: keyword-based identification of pre-defined topics
  - Finding: they want to reduce the probability of collective action (not to suppress criticism)

Two Approaches to Selecting Texts

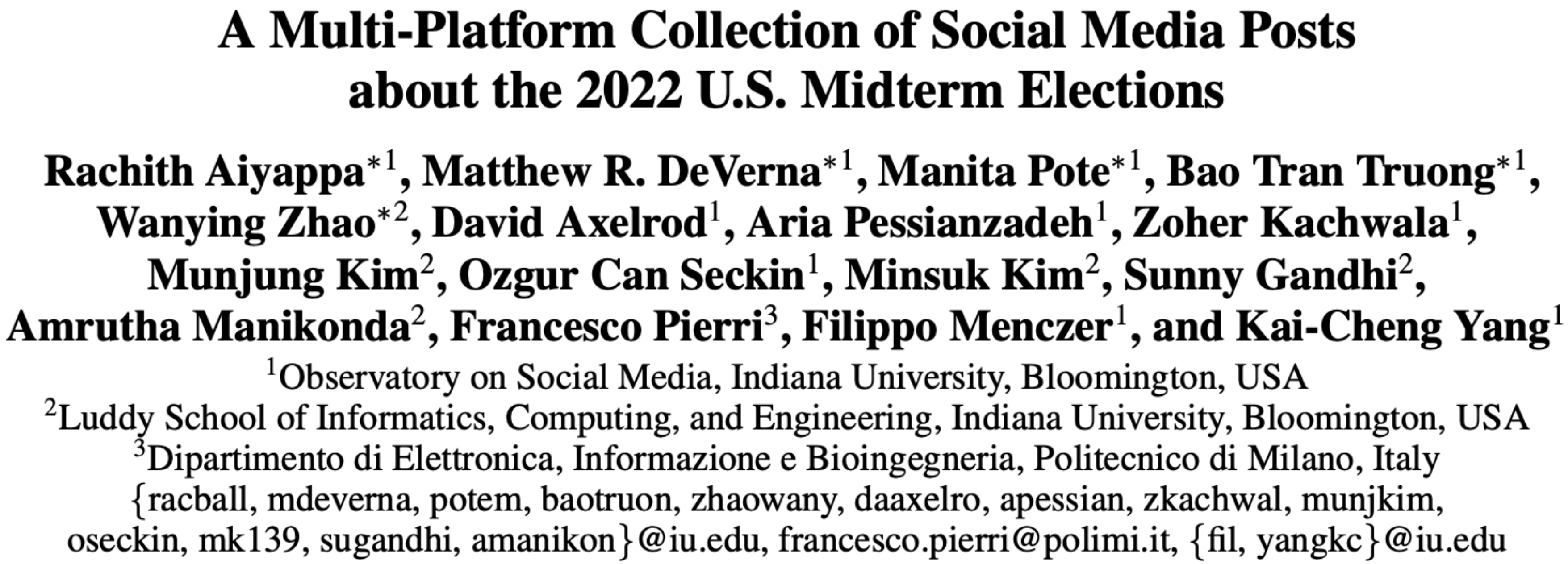
2. Data-driven

- We have interesting ideas that have not yet developed into specific questions
- Gather as much relevant data as possible
- These data can work as a “pool” for questions that arise later

Two Approaches to Selecting Texts

2. Data-driven

- E.g., Aiyappa et al. (2023)
  - 2022 U.S. Midterms Multi-Platform Social Media Dataset



A Multi-Platform Collection of Social Media Posts about the 2022 U.S. Midterm Elections

Rachith Aiyappa*[1], Matthew R. DeVerna*[1], Manita Pote*[1], Bao Tran Truong*[1], Wanying Zhao*[2], David Axelrod[1], Aria Pessianzadeh[1], Zoher Kachwala[1], Munjung Kim[2], Ozgur Can Seckin[1], Minsuk Kim[2], Sunny Gandhi[2], Amrutha Manikonda[2], Francesco Pierri[3], Filippo Menczer[1], and Kai-Cheng Yang[1]

[1]Observatory on Social Media, Indiana University, Bloomington, USA
[2]Luddy School of Informatics, Computing, and Engineering, Indiana University, Bloomington, USA
[3]Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Italy
{racball, mdeverna, potem, baotruon, zhaowany, daaxelro, apessian, zkachwal, munjkim, oseckin, mk139, sugandhi, amanikon}@iu.edu, francesco.pierri@polimi.it, {fil, yangkc}@iu.edu

Two Approaches to Selecting Texts

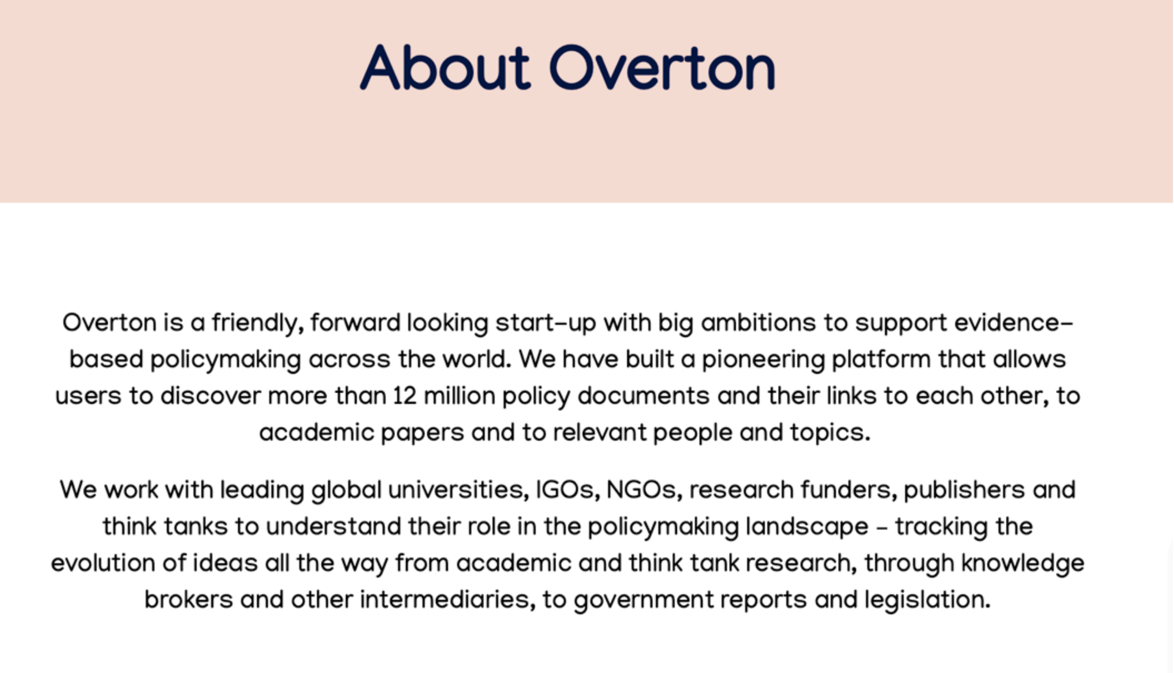
2. Data-driven

- E.g., Aiyappa et al. (2023)
  - 2022 U.S. Midterms Multi-Platform Social Media Dataset
    - “A collection of social media posts from Twitter, Facebook, Instagram, Reddit, and 4chan”
    - “Posts about the midterm elections based on a comprehensive list of keywords and tracks the social media accounts of 1,011 candidates from October 1 to December 25, 2022”
    - “We also publish the source code of our pipeline to enable similar multi-platform research projects”

Two Approaches to Selecting Texts

2. Data-driven

- E.g., Overton database
  - Policy documents and scientific publications cited in them
  - You might need to retrofit your questions to the data



About Overton

Overton is a friendly, forward looking start-up with big ambitions to support evidence-based policymaking across the world. We have built a pioneering platform that allows users to discover more than 12 million policy documents and their links to each other, to academic papers and to relevant people and topics.

We work with leading global universities, IGOs, NGOs, research funders, publishers and think tanks to understand their role in the policymaking landscape – tracking the evolution of ideas all the way from academic and think tank research, through knowledge brokers and other intermediaries, to government reports and legislation.

Two Approaches to Selecting Texts

2. Data-driven

- Alternatively, we are interested in specific texts
- Texts themselves hold intrinsic significance and merit study in their own right, independent of any specific research question
  - E.g., important literary works, historical texts, etc.

Two Approaches to Selecting Texts

Both are valid approaches; in the end, what matters is alignment between texts and research goals

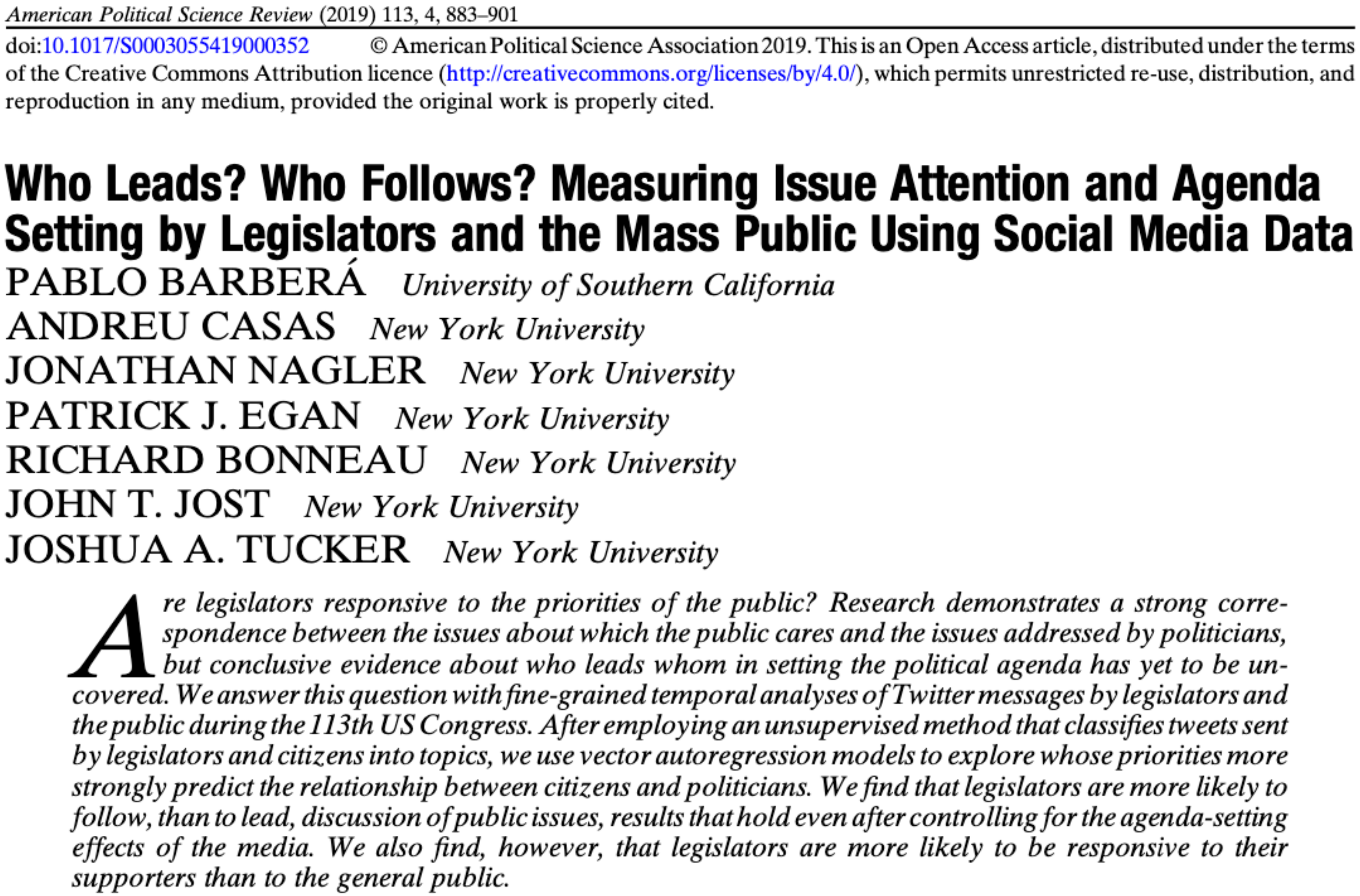
Population, Sample, Quantity of Interest

In the context of selecting texts

- Population: a set of entities that relate to a research question
- Sample: a subset of the population selected for analysis
- Quantity of interest: a numeric value that is of particular interest

Population, Sample, Quantity of Interest

E.g., Barbera et al. (2019)



American Political Science Review (2019) 113, 4, 883–901

doi:10.1017/S0003055419000352

Who Leads? Who Follows? Measuring Issue Attention and Agenda Setting by Legislators and the Mass Public Using Social Media Data

PABLO BARBERÁ *University of Southern California*
ANDREU CASAS *New York University*
JONATHAN NAGLER *New York University*
PATRICK J. EGAN *New York University*
RICHARD BONNEAU *New York University*
JOHN T. JOST *New York University*
JOSHUA A. TUCKER *New York University*

Are legislators responsive to the priorities of the public? Research demonstrates a strong correspondence between the issues about which the public cares and the issues addressed by politicians, but conclusive evidence about who leads whom in setting the political agenda has yet to be uncovered. We answer this question with fine-grained temporal analyses of Twitter messages by legislators and the public during the 113th US Congress. After employing an unsupervised method that classifies tweets sent by legislators and citizens into topics, we use vector autoregression models to explore whose priorities more strongly predict the relationship between citizens and politicians. We find that legislators are more likely to follow, than to lead, discussion of public issues, results that hold even after controlling for the agenda-setting effects of the media. We also find, however, that legislators are more likely to be responsive to their supporters than to the general public.

Population, Sample, Quantity of Interest

E.g., Barbera et al. (2019)

- Question: do politicians pay attention to (respond to) the issue priorities of the public?
- Population: tweets from politicians and the public
- Sample: all tweets sent by members of the 113th Congress (Jan 2013–Dec 2014)
- Quantity of interest: distributions of daily topics in tweets

Population, Sample, Quantity of Interest

(Some) people like to boast how big there data set is

- But consider how survey work is conducted!

"I think you'll find that mine is bigger..."

Population, Sample, Quantity of Interest

Uncertainty measures

- When we estimate a population parameter based on a sample, we quantify uncertainty to assess how close our estimate is likely to be to the true population value

We also consider generalizability

- We can better generalize our findings based on a sample that is random and representative of the population

Population, Sample, Quantity of Interest

Comprehensiveness / representativeness

- If your goal is not building a database per se, we do not necessarily have to (and cannot) collect *all* data
- In many applications, we only need data big enough for statistical inferences
- This is particularly the case for short-term, small-sized projects

Four Sources of Bias

(Try to) explore the data (text) generation process

- What you get is likely *the end product*
  - E.g., customer reviews, bill texts, historical newspaper articles, etc.
- Most text data is not generated for research purpose
  - Any exceptions?
- Please think about the process by which data was produced, stored, and put on your hands

Four Sources of Bias

1. Incentive bias

- The production/retention of documents is driven by strategic behavior
- The underlying incentives are diverse
- E.g., if you want to study online hate speech
  - Consider social media platforms’ content moderation

Four Sources of Bias

1. Incentive bias

- At a deeper level, we need to think about why certain things are said/written in the first place
- E.g., Schonhardt-Bailey (2013)
    - Analysis of transcripts of the Federal Open Market committees and Congressional committees on banking
    - Elite interviews to delve into what the transcripts represent
        - How candid the members felt they could be?
        - How transparency of the transcripts themselves affected their deliberation

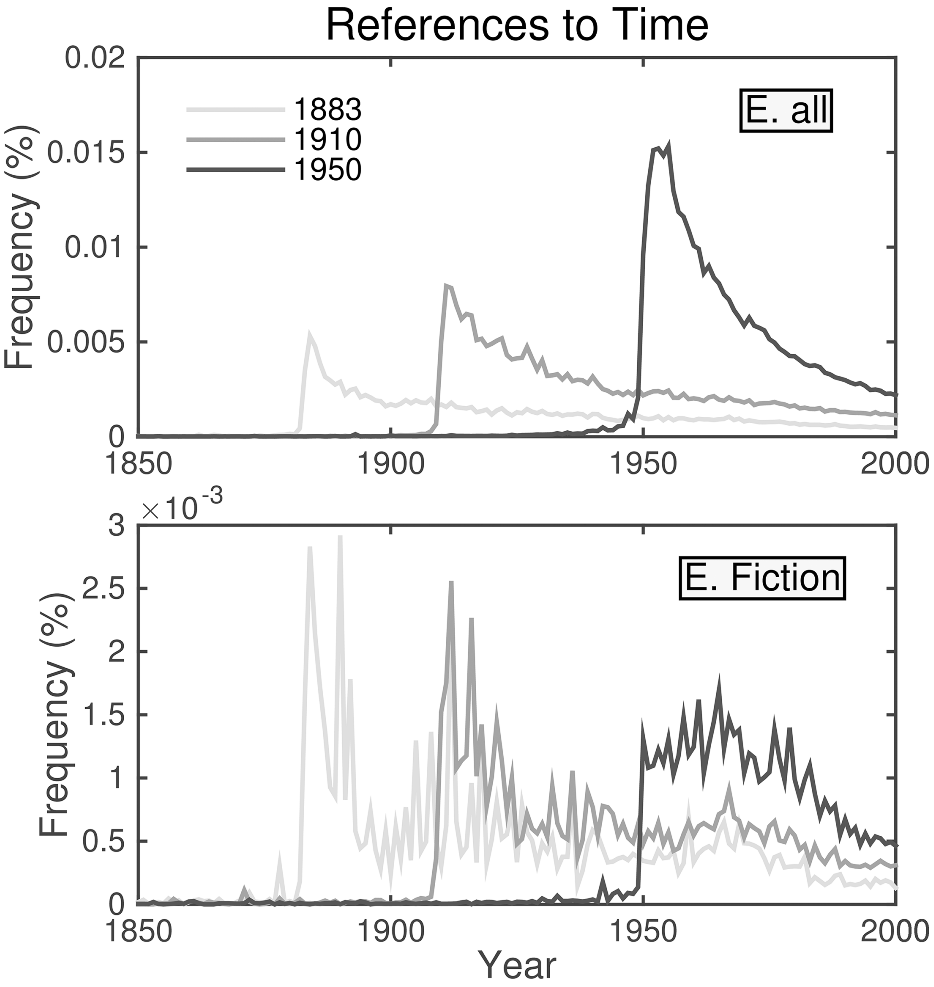
Four Sources of Bias

2. Resources bias

- Text often better reflect populations with more resources to produce, record, and store documents
- E.g., Google Books Corpus and literacy
    - Increasingly disproportionate amounts of scientific texts in the 20th century (Pechenick et al. 2015)

Four Sources of Bias

2. Resources bias



Four Sources of Bias

2. Resources bias

- E.g., Wikipedia: digital literacy
- E.g., Overton: government capacity (also related to incentive bias)

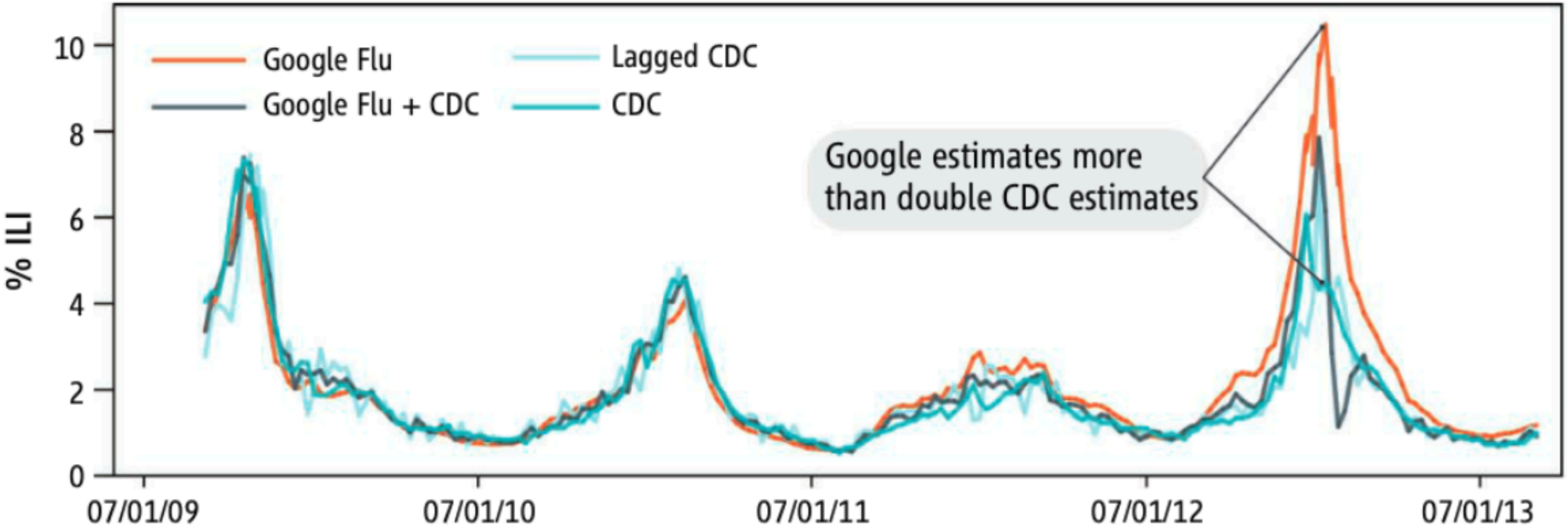
Four Sources of Bias

3. Medium bias

- Content can also be affected by the technologies/types of mediums
  - E.g., Jaidka et al. (2019)
    - The effect of limitations on character count on Twitter (previously 140)
    - “We show that doubling the permissible length of a tweet led to less uncivil, more polite, and more constructive discussions online.”
  - E.g., Lazer et al. (2014)
    - Google Flu tracks the prevalence of ILI (Influenza-like illness) using Google search texts data
    - Accurately predicted CDC estimates (ground truth) until changes in search algorithms

Four Sources of Bias

3. Medium bias



Four Sources of Bias

4. Retrieval bias

- Retrieval of data is often incomplete
- Systematic
  - E.g., biased keywords
  - Also note de-biasing methods (e.g., King, Lam, and Roberts 2017)
- Non-systematic (e.g., disrupted connection)

Four Sources of Bias

“Garbage in garbage out”

- Need to develop critical eyes for data quality
- No (text) data is perfect though
- Be aware of and acknowledge limitations and potential biases

Four Sources of Bias

What should we do with all the (un)known biases?

→ Collect good data

→ Ask how the bias affects the quantities of interest

→ Reframe/modify RQs

Iterative Process

1. Collect texts (before/after you identify questions)
2. Evaluate how well the texts are aligned with the questions
3. Augment texts
4. If not feasible, be sure to discuss limitations, including the direction and extent of bias.

Some NLP/text-as-data Terminology

Corpus

- Corpus (plural: corpora): a computer-readable collection of text or speech
- Size of corpus: the number of texts or words
- Punctuation: period, comma, apostrophe, quotation, question, exclamation, brackets, parenthesis, dash (—), hyphen (-), ellipsis (…), colon, semicolon

Some NLP/text-as-data Terminology

Sentence

- A set of words that is complete in itself
- Utterance is the spoken version of a sentence
    - Disfluences: fragment, filler/filled pauses
    - E.g., “I do uh main- mainly business data processing”

Some NLP/text-as-data Terminology

Lemma

- A set of lexical forms having the same stem ⟷ word forms
- Run (lemma)
    - Runs (third person singular present)
    - Ran (simple past)
    - Running (present participle)

Some NLP/text-as-data Terminology

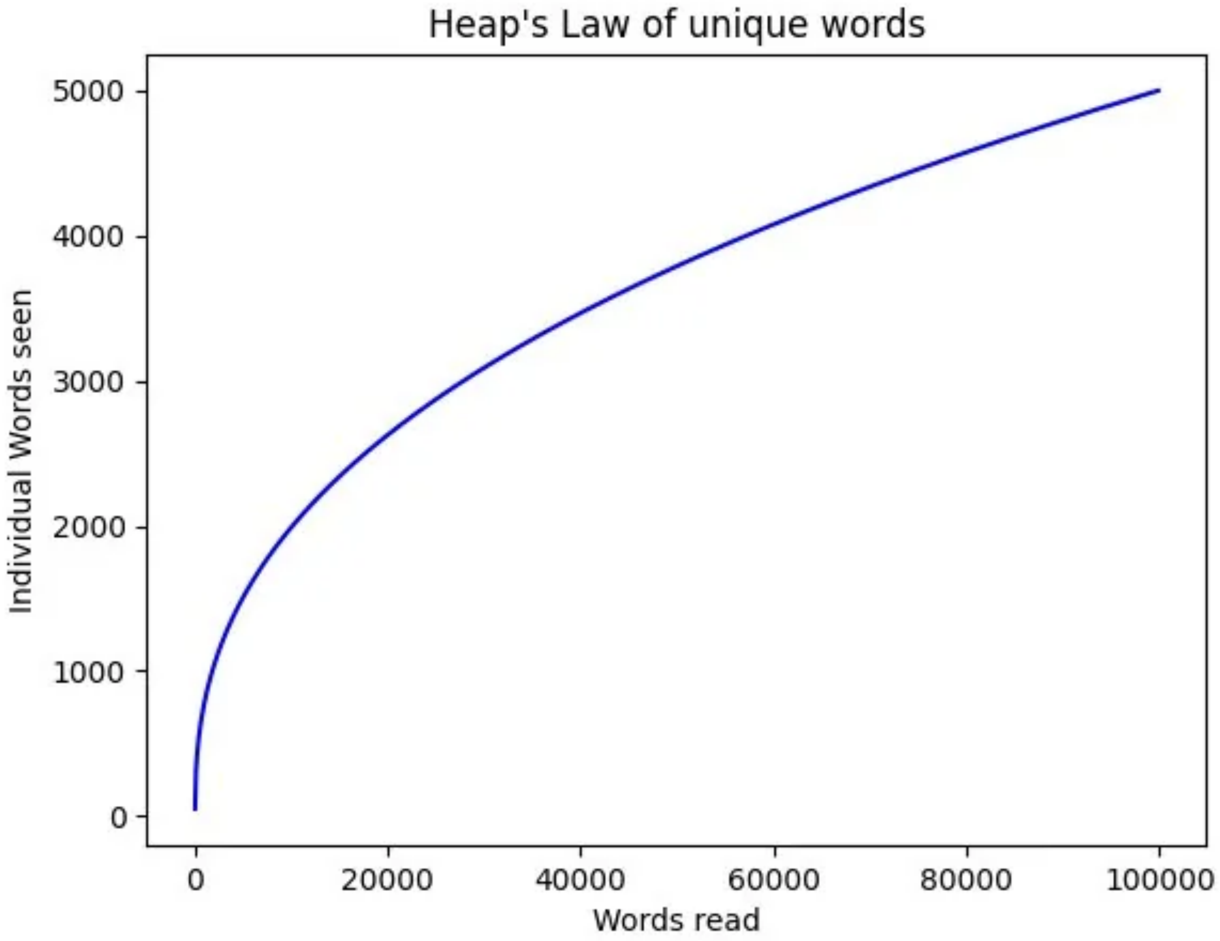
Tokens and types

- Token: the total number N of running words
- Type: the number of distinct words in a corpus (its size indicated with |V|)
- E.g, “they picnicked by the pool, then lay back on the grass and looked at the stars”

Some NLP/text-as-data Terminology

Tokens and types

- Heap’s Law



Some NLP/text-as-data Terminology

Strings

- A sequence of characters
  - “file upload complete”
  - “I got a new job today”
  - “100%”
  - “?action=edit”

Some NLP/text-as-data Terminology

Reg(ular) ex(pression)

- Notation for characterizing a set of strings
- Powerful way to search text based on certain patterns
- E.g., cellphone numbers in South Korea `010-\d{4}-\d{4}`

Summary

Before you dive into analysis, make sure to contemplate on

- The match between your data and questions
- Potential sources of bias
- How they can affect your findings
- How you can justify it

Guided Coding

String manipulation and regex in Python (Link)