Text Mining with R

Prof. Haohan Chen
The University of Hong Kong

Guest Lecture at National University of Singapore on Feb 2024

Today

A Hands-on Approach

- Why do we care about text mining?
- What is text mining?
- Overview of text mining workflow
- Lab: Text mining with politicians' public speech

Why text mining?

Identifying nonprofit organizations

Chen & Zhang (2021) Identifying Nonprofits

- Data: I-990 forms of millions of nonprofit organizations
- Mission: Identify organizations of interest for inquiry
- Challenges
  - IRS has a uninformative categorization system for nonprofits
  - No way to find some “niche”
- Solution: Quantify org’s self-reported missions and gets their relationship with keywords of interest

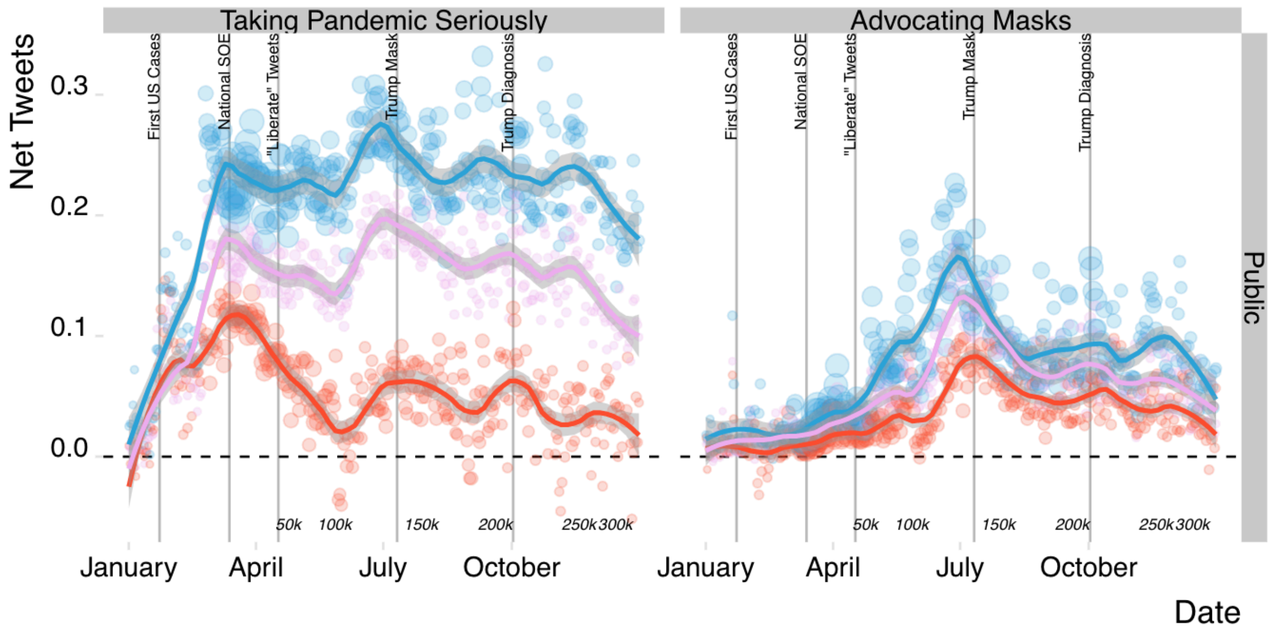
(a) Gay

(b) God

Tracking the time dynamics of COVID-19 Polarization

Bisbee, Chen, et al. (2021) Leaders and Followers during Covid-19

- Data: 1 million Twitter users' posts
- Mission: Understand how COVID gets politicized and polarized in the USA
- Challenges: Measuring public opinions on COVID
- Solution
  - Use tweets as opinion measures
  - Deep learning machine classifiers



Opinion, communication, media

Where text mining have found most applications

- Measuring public opinions — Supplement surveys with social media posts?
- Media bias — How outlets report different things and report the same thing differently
- Political polarization — Do people see and talk to each other? When they talk, what happens?
- Censorship and self-censorship — Do the state limit accessible information? Do people self-control what they say?

Political institutions

Zhu et al. (2017) “Big Tigers, Big Data”

- Data: Comments under Chinese news about corruption investigations
- Methods: Dictionary/ Keywords
- Findings: Anti-corruption campaign boost public support for the regime and legal institutions.

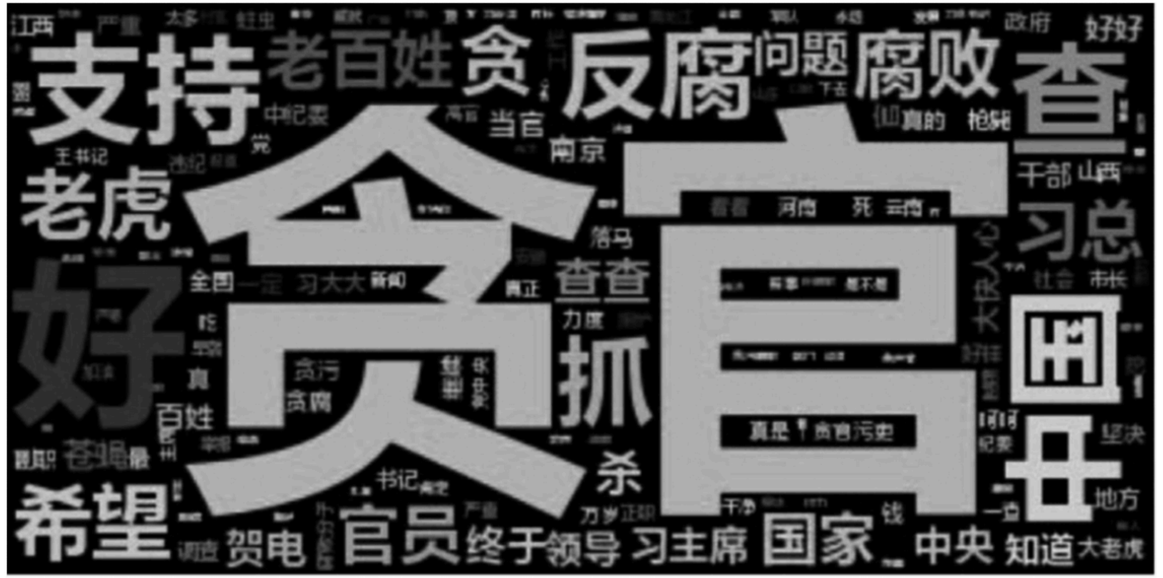


Figure 3.1 Word Cloud

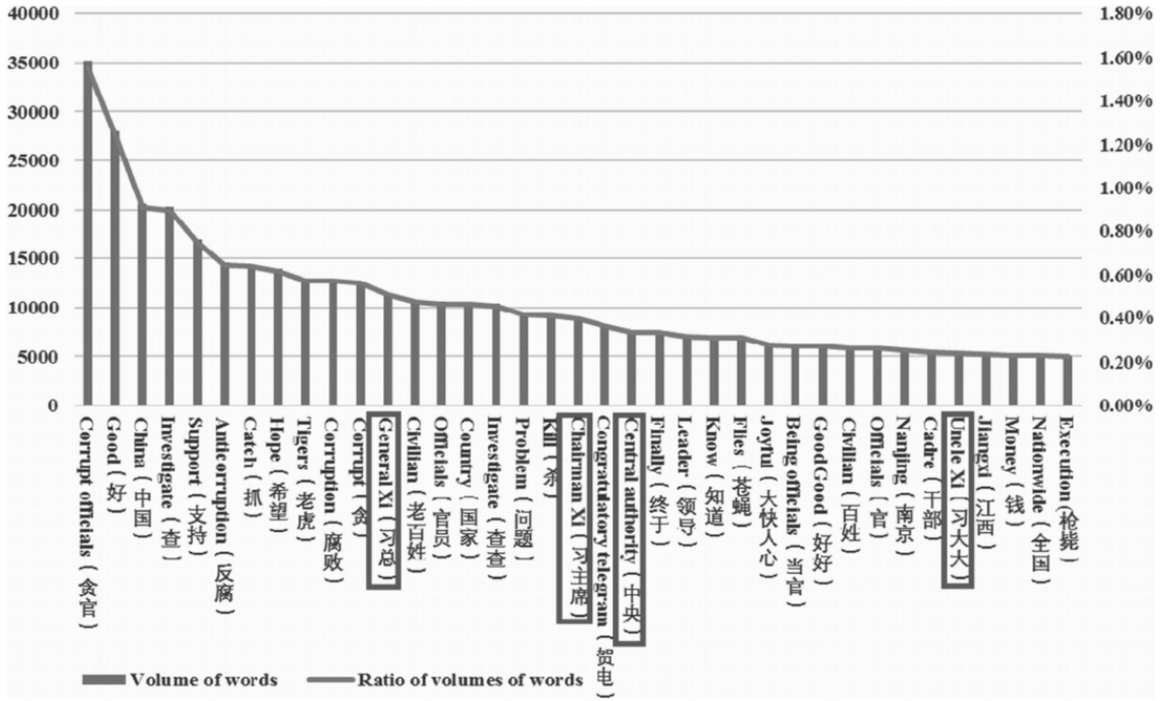
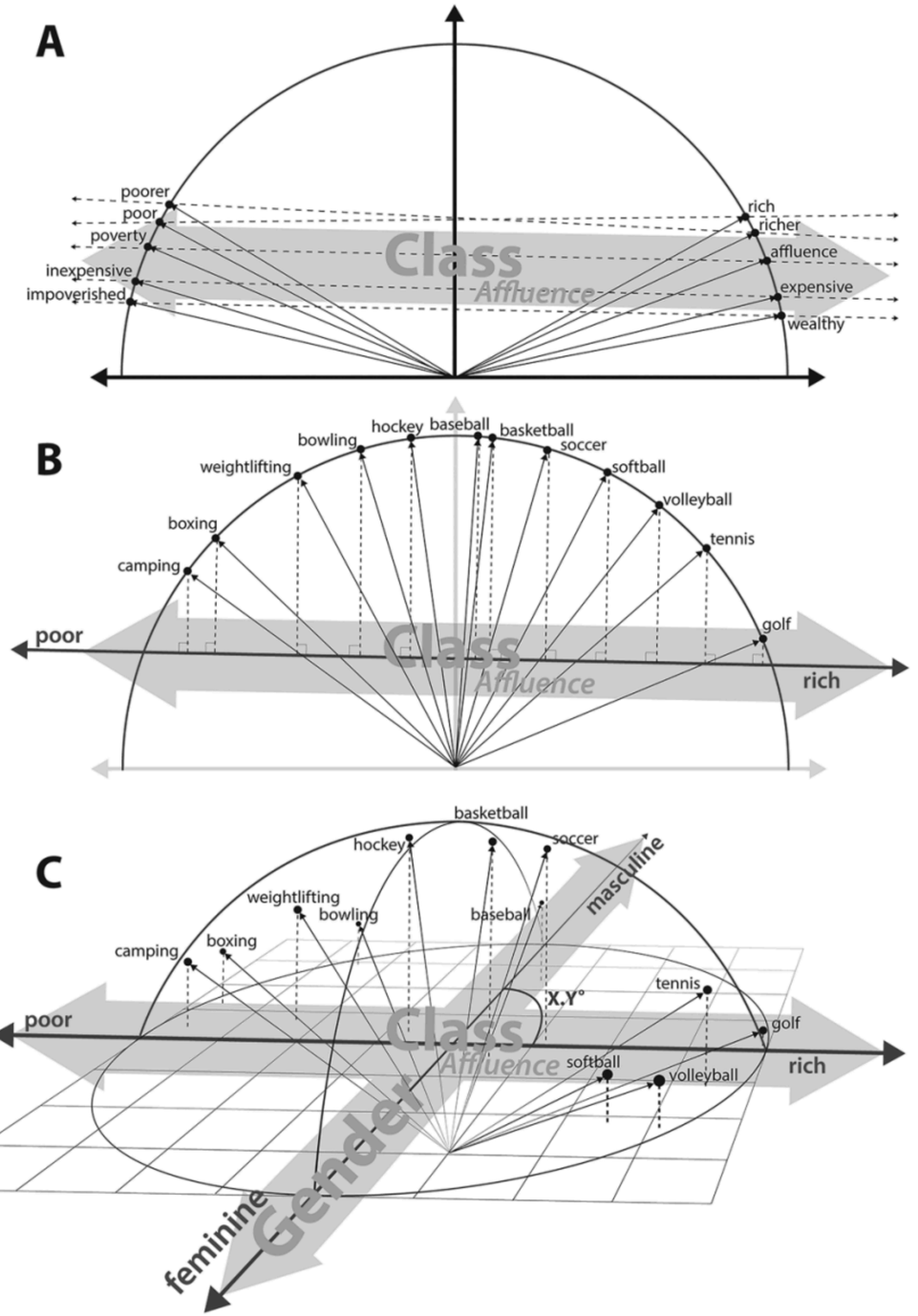


Figure 3.2 Word Frequency

Culture

Kozlowsky (2019) The geometry of culture

- Data: Millions of books published over 100 years
- Methods: ***Word embeddings***
- Findings:
  - Culture dimensions of class remain stable
  - Education became tightly linked with affluence



What is text mining?

What text mining is NOT

Some common misperceptions

Wrong	RIGHT
It's science.	It's science + engineering + art
It's quantitative.	It's quantitative + qualitative.
It's language-specific.	The rationale behind methods for different languages are mostly the same.
It's just about web scraping.	That's just one way to *collect* text data for mining
It's just about word clouds.	It's just one of many text data visualization methods.
It's just about topic modeling.	It's just one out of many text mining models

Principles of text mining

Grimmer & Stewart (2013)

- All quantitative models of language are wrong, but some are useful
- Quantitative methods augment humans, not replace them
- There is no globally best method for automated text analysis
- Validate, validate, validate

Text mining goes beyond “mining”

Causal identification it can do! (Egami et al. 2018)

- Text mining for causal identification
  - Correlations are not causations! (See this year’s Nobel Prize in Econ?)
  - *Analyst induced SUTVA violation*
    - You as researchers read texts with bias: “you see what you want to see”
    - **Solution:** “Cover one of your eyes”
      - Analyze a random sample of your text data
      - Estimate the causal relations of interest in the rest.

Text mining and Natural Language Processing

NLP. A larger toolkit designed for different objectives

- You have heard about the term Natural Language Processing, right?
  - Check out the Wikipedia page of this term
- The text mining methods I am familiar with and you will learn about:
  - Rooted in computer scientists' work on Natural Language Processing
  - BUT we redirect them to answer social sciences questions
  - Application, adaptation, and extension

Computational tools empowering text mining

- General-purpose tools applied to text mining
  - EDA and visualization: Variants of what we learned
  - Stats/Machine learning models (e.g., regressions etc.): Some we will learn
  - Deep learning (a special type of machine learning at the frontier)
    - Feature: COMPLEX models, HUGE data, GPU taxing, GREAT performance, and NOBODY knows why (“alchemy” ref)
  - “AI”

Text Mining Workflow Elaborated

Text Mining Workflow

Part 1: Retrieval

- **Indexing**: List gateways to documents you need
- **Retrieval**: Retrieve the raw documents
- **Parsing**: Obtain content you want
- **Organization**: Put documents into tables

Text Mining Workflow

Part 2: Analysis

- **Clean** the text data
- **Tokenize** texts into smaller unit of analysis
- **Wrangle** tokenized text to facilitate analysis
- **Explore** text data to discover patterns
- **Extract** information you need

RETRIEVAL

Index. Retrieve. Parse. Organize

Retrieval Method 1: API

- API (application programming interface)
  - As oppose to a user interface
  - Protocols that facilitates programs “talking with each other”
- Friendly to automated retrieval of large text data


```
{
  "created_at": "Thu Apr 06 15:24:15 +0000 2017",
  "id_str": "850006245121695744",
  "text": "1\/ Today we\u2019re sharing our vision for the future of the Twitter API platform!\n
  "user": {
    "id": 2244994945,
    "name": "Twitter Dev",
    "screen_name": "TwitterDev",
    "location": "Internet",
    "url": "https:\/\/dev.twitter.com\/",
    "description": "Your official source for Twitter Platform news, updates & events. Need techn
  },
```

Retrieval Method 1: API

- Pros: Low technical barrier as long as you get access
- Cons: May be costly and less accessible
- Examples: New York Times, Twitter

Retrieval Method 2: Web Scraping

Gather links to contents of interest I: Web Scraping

- Web scraping
  - Retrieve information from webpages (usually with automation)
  - Parse webpages to get specific information needed
- Pros: Flexible
- Cons: Higher technical complexity.

Index

- Put together a collection of identifiers of documents to collect. E.g.,
  - URLs
  - User identifiers: names, IDs
  - Document identifiers: (keywords, dates of publication)
- May be a iterative process: You don’t always know what you want/ can get
  E.g., keyword expansion

Retrieve

Retrieve raw documents that contain text

- Web-based text documents are formatted in a variety of formats
- Most popular formats: *JSON*, *XML*, *HTML*, *PDF*
- Retrieving them as a crucial second step
- [See examples]

Parse

Obtain text from raw text document

- Use tools to get the text data of interest from the raw documents
- Useful tools
  - JSON: package “jsonlite”
  - XML and HTML: package “xml” and “rvest”
  - PDF: pdftools
- [Demonstration in R]

Organize

Put many text data together for analysis

- Text are analyzed along with other data
- E.g., social media posts' timestamp, popularity
- Typical format: A table where each row is a document
- Challenges: Loading and saving
  - Size: If too large, break into chunks, put in databases designed for large data
  - Encoding: If multiple languages, properly convert encodings and delete special characters that can cause problems

ANALYSIS

Clean. Tokenize. Wrangle. Explore. Extract.

Basic String Operations

- In many cases, you only need to systematically do some simple operations
- Example
  - Find documents containing some keywords
  - Extract paragraphs/ sections in certain forms/ shape
  - Or remove contents fitting certain patterns
- What you need: Basic string operations

Tokenization

Cut text into small units of analysis

- Objective: Find documents' common ingredients —> Comparing 🍎 to 🍎
- "Tokenization:" Separate documents into smaller units of analysis
  - Paragraphs
  - Sentences
  - **WORDS (1 word, 2 words, 3 words…)**
  - Phrases (e.g., Hong_Kong, New_York_Times, Donald_Trump)
  - Letter / Stroke (笔画)!

Wrangling

- Removing redundant words
  - Removing stop words
  - Removing other selected words (e.g., procedural words)
- “Harmonize”
  - Change all to lower cases
  - Remove numbers, special characters, etc.
  - Stemming

Exploratory Data Analysis

- Exploratory data analysis: Word frequency
  - Check word frequency
  - Plot word clouds
  - Compare word frequencies between two sets of documents

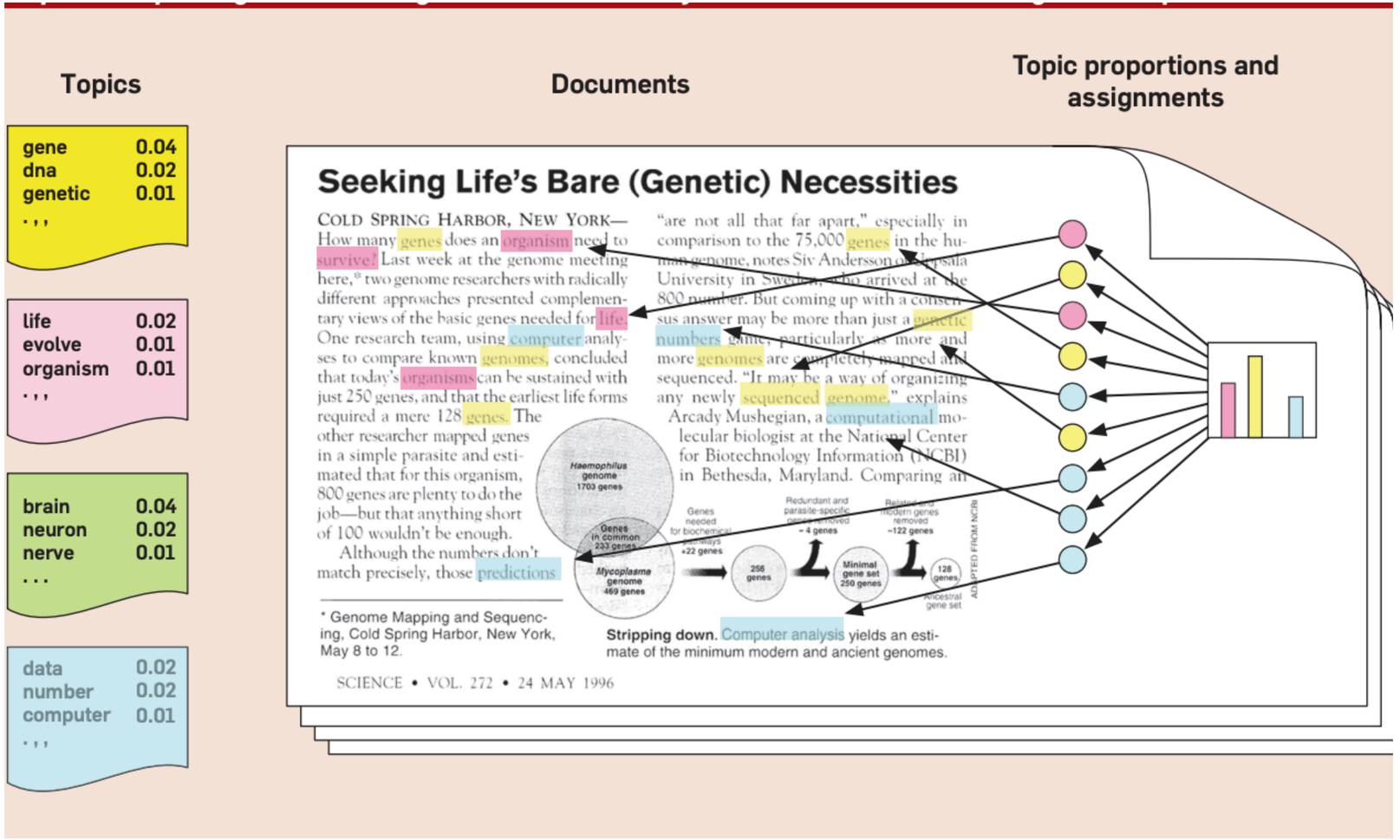
Extract

Sentiment Analysis

- Detect sentiment and emotions in expressed in text
- Different approaches
  - Basic: Keyword search based on a sentiment/emotion dictionary
  - Advanced: Machine learning classifier based on labeled data
- Caution: Sentiments, emotions, and stances are not the same thing. Expressions across different types of documents differ

Extract

Topic Modeling

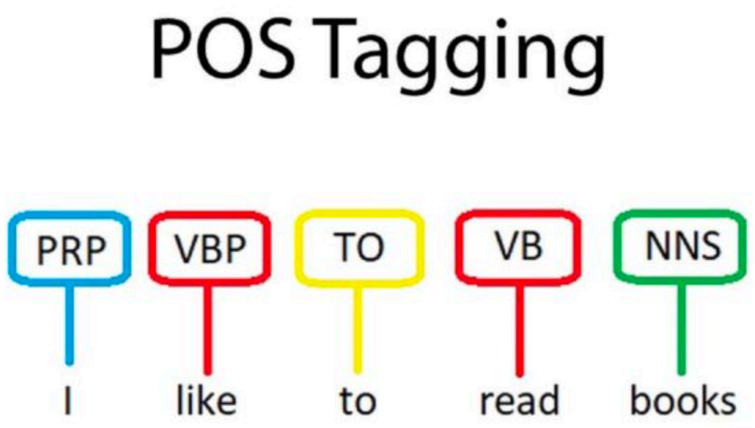


Source: Blei (2012) http://www.cs.columbia.edu/~blei/papers/Blei2012.pdf

Extract (intermediate II)

Standardized target (externally defined)

- Extract standardized targets
  - Part-of-speech tagging (POS tagging): noun, verb, preposition, etc.
  - Named entity recognition (NER): location, organization names, crime, etc.)



Sources: https://byteiota.com/pos-tagging/

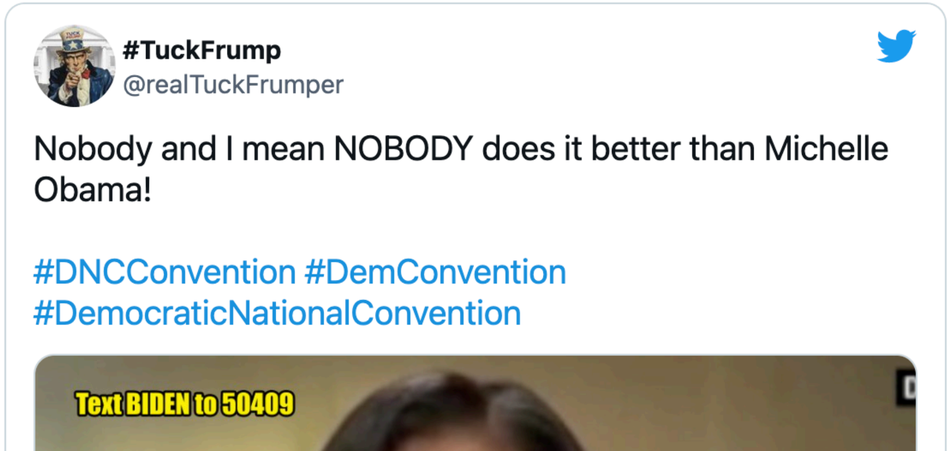
Sources: https://www.aiimi.com/insights/aiimi-labs-on-named-entity-recognition

Extract

When what you want to extract contains complex concept

- Extract complex defined targets: Supervised machine learning models
  - Manually label a portion of the text
  - Use the data to “train” the machine
  - Machine label of the remainder to data

Nobody and I mean NOBODY does it better than Michelle Obama! #DNCConvention #DemConvention #DemocraticNationalConvention https://t.co/y06hyrIodV



DEMOCRATIC Aversion/Affection

😀 Affection

DEMOCRATIC target

Others

REPUBLICAN Aversion/Affection

REPUBLICAN target

Othering

Moralization

Reason

Behavior

Media

Video

Relevance

Large Language Models as (better) alternatives?

Large Language Models

For social text data mining

- Can help with with each step of the tasks
- But you still need to figure what you want
- For more information: https://github.com/haohanchen/Lecture_ChatGPT