Using Text Data Mining & Natural Language Processing for Psychological Research

Yuyang Zhong
me@yuyangzhong.com

SIPS 2022 Workshop | Victoria, BC, Canada
June 28, 2022

About Me

Berkeley
UNIVERSITY OF CALIFORNIA

UNIVERSITY OF MICHIGAN

coding it forward >

CIVIC DIGITAL FELLOWSHIP

NIH
Office of AIDS Research

Today's Agenda

Why text data?

Text data in the industry vs. in psychological research

Working with text data

- Scraping text data
- Cleaning text data
- Analyzing text data

Preface

- There’s a lot of content!
- Demos at more surface level examples
- Maybe a future, more in-depth workshop/hackathon?
- (Hopefully) time at the end to explore project ideas

Workshop Setup

LOCAL SETUP
JUPYTER NOTEBOOK

CLOUD SETUP
GOOGLE COLLABORATORY

Running Notebooks

- Use `SHIFT+Enter` to run the current cell and advance forward
- Use `CTRL/CMD+Enter` to run the current cell and remain there

Note:

Please make sure your notebook is open + you've run the top cells to install & import all required packages if you haven't already.

Why text data?

Why text data?

- Text data is everywhere
- Likert scales don’t tell the full story
- Avoids researcher confirmation bias
- Methods scalable and reusable

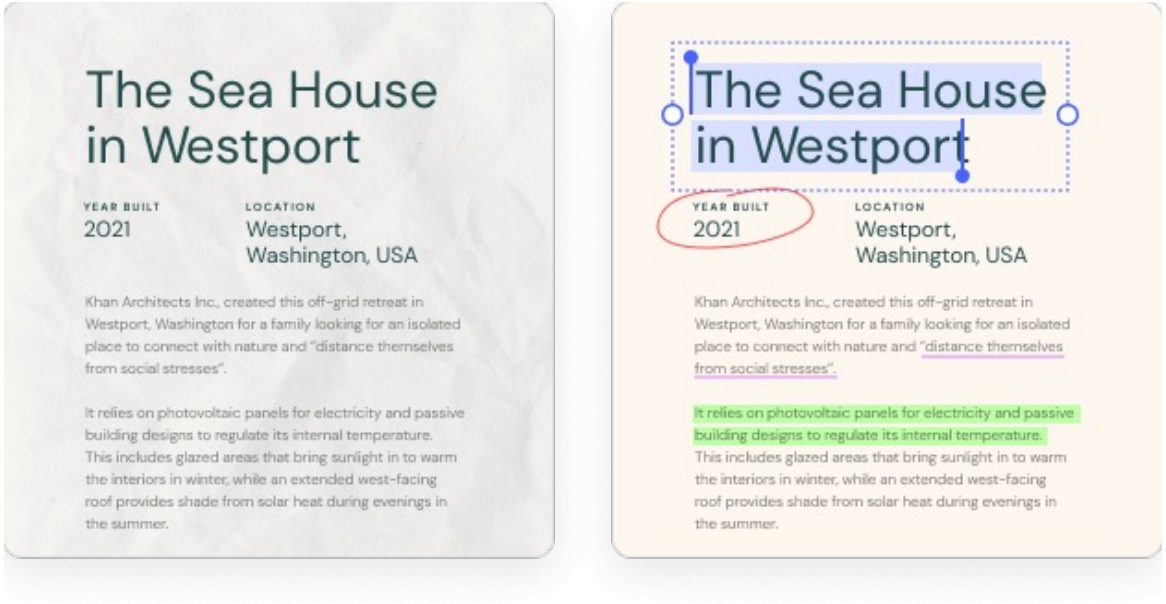
Why haven't we used text data?

- Unstructured in nature (i.e., not numbers)
- Less common practice in psychology
- Limited training for natural language processing

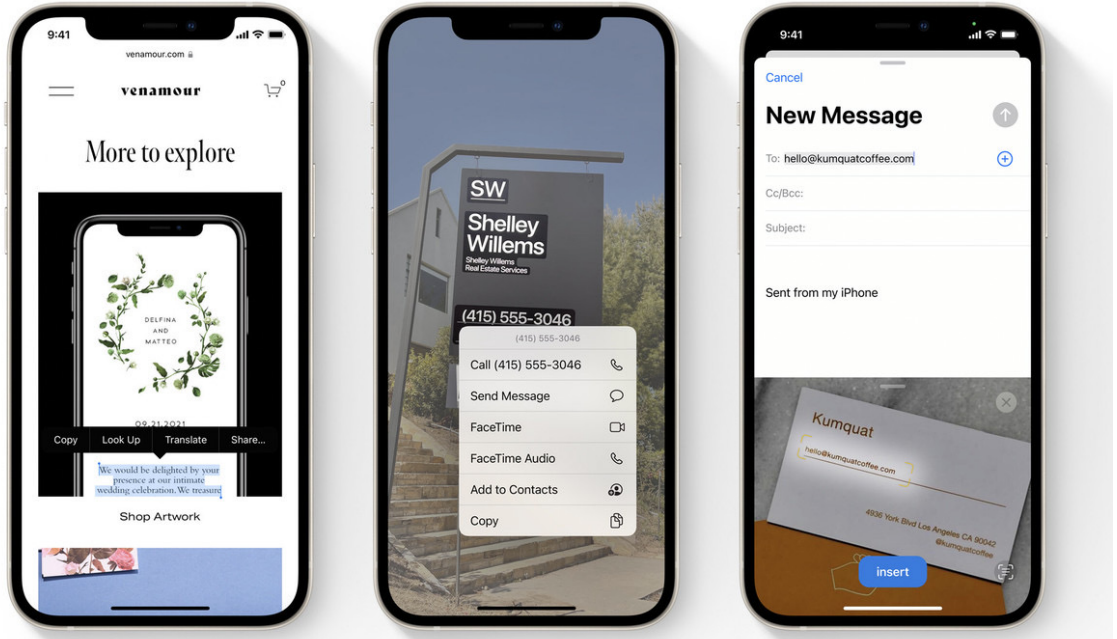
Text Data in the Industry

How does the tech industry use text data?

Optical Character Recognition (OCR)



PDFelement: Converting Documents



Apple: Live Text from Photos

Categorizing Transactions

- How does tools like Intuit Mint/Plaid know how to extract information and categorize your transactions?

MCDONALD'S 11016 LA CANADA CA

UNIV OF MICHIGAN DIR DEP PPD ID: 3823982012

WHOLDFDS STV #10571

AplPay BOBA GUYS ROC Oakland CA

Categorizing Transactions

- How does tools like Intuit Mint/Plaid know how to extract information and categorize your transactions?

MCDONALD'S 11016 LA CANADA CA

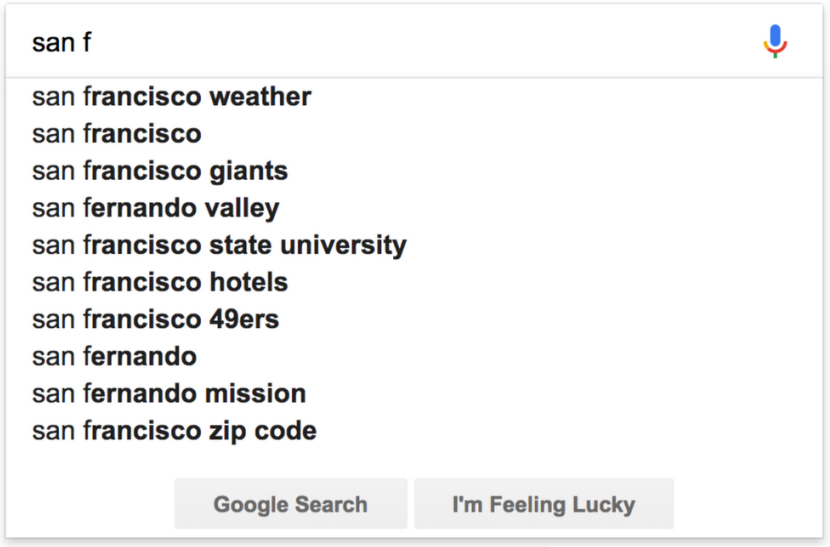
UNIV OF MICHIGAN DIR DEP PPD ID: 3823982012

WHOLDFDS STV #10571

AplPay BOBA GUYS ROC Oakland CA

Text Completion (Auto-complete)

Google Searches



GitHub Co-pilot

https://github.com/features/copilot/

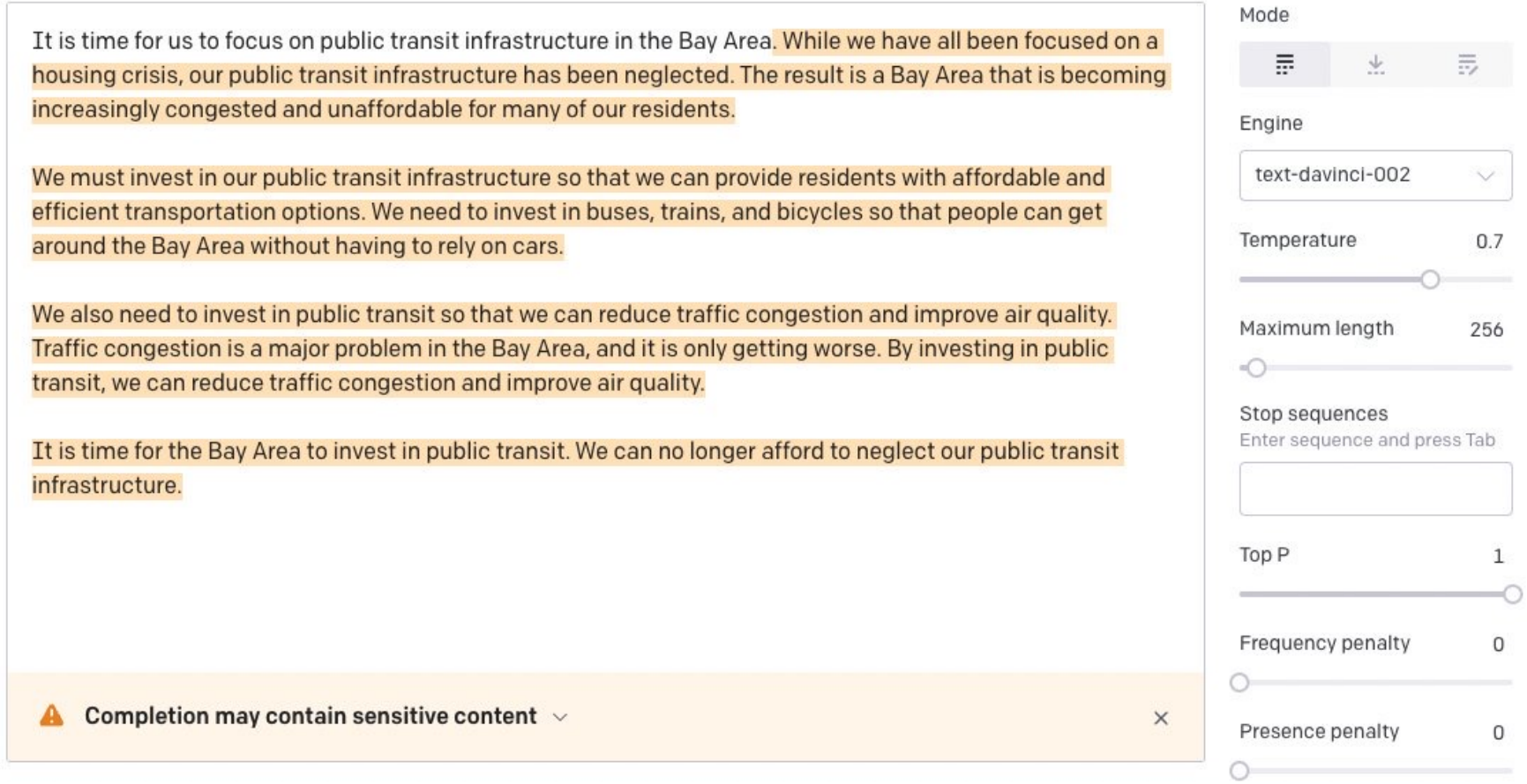
```ts
#!/usr/bin/env ts-node

import { fetch } from "fetch-h2";

// Determine whether the sentiment of text is positive
// Use a web service
async function isPositive(text: string): Promise<boolean> {
  const response = await fetch(`http://text-processing.com/api/sentiment/`, {
    method: "POST",
    body: `text=${text}`,
    headers: {
      "Content-Type": "application/x-www-form-urlencoded",
    },
  });
  const json = await response.json();
  return json.label === "pos";
```

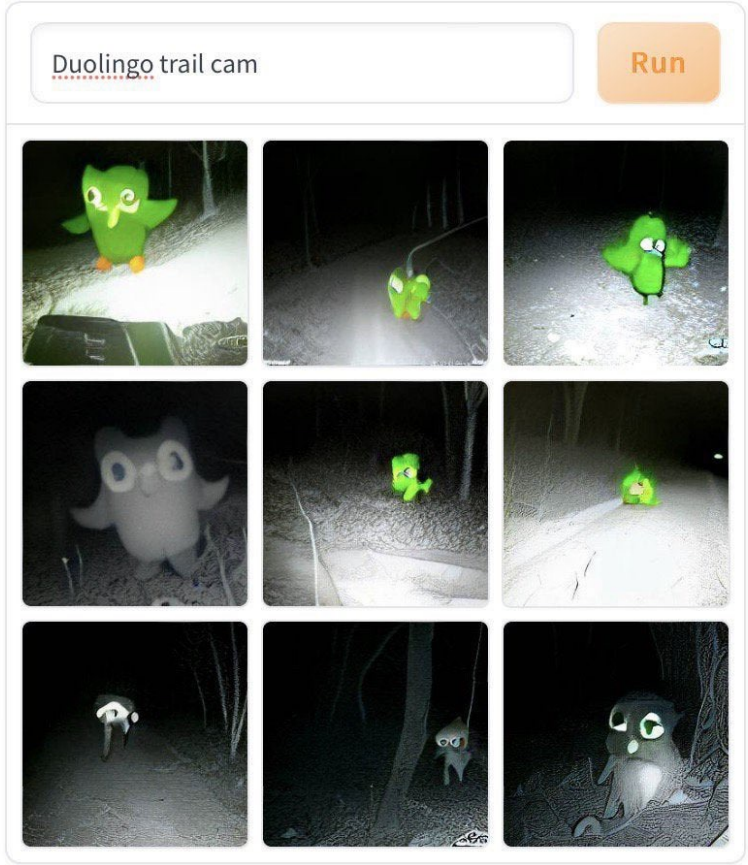
Text Completion (writing essay!)

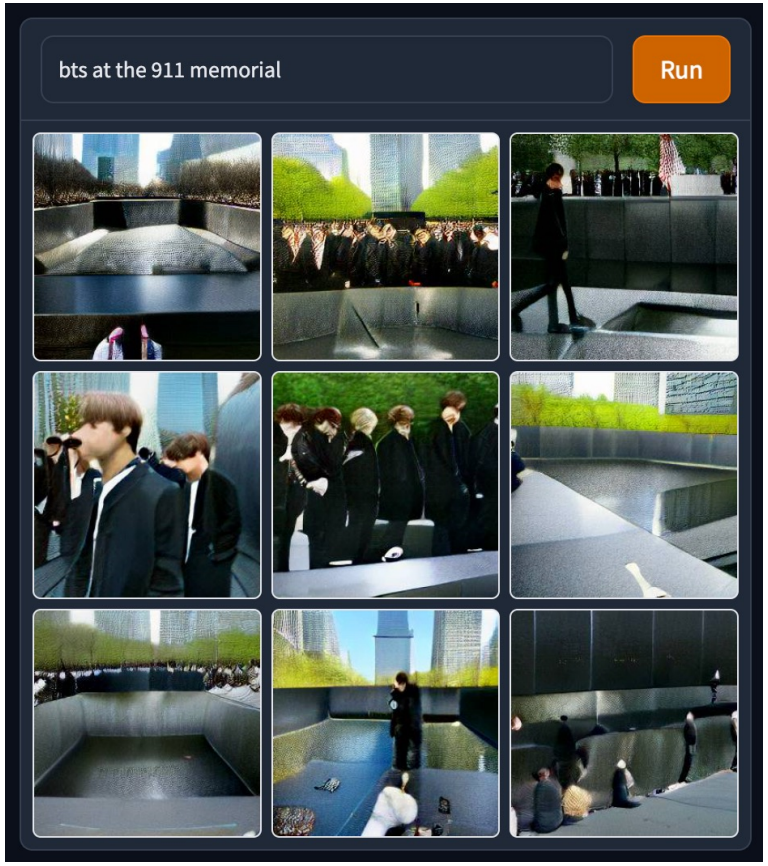
https://beta.openai.com/docs/guides/completion



Text-to-Image Generation

https://huggingface.co/spaces/dalle-mini/dalle-mini





Beyond Text Data -> Natural Languages

Speech recognition

Text-to-speech

Text Data in (Psychological) Research

How has previous research worked with text data?

Linguistic Features of Stigmatized Groups

Riddle et al. 2015

Identify linguistic features of race & gender in value affirmation essays

Topic models reveal differences in values in relations to identities

Sentiment Analysis & Polarization

Cook, Huang, & Xie 2021

- 297 million tweets
  (up to June 2020)
- Awareness of COVID-19
- Anti-China attitudes

Charter School Websites

Haber 2020; Haber, Camp, & Kim (in progress)

Influences of Parental Judgements

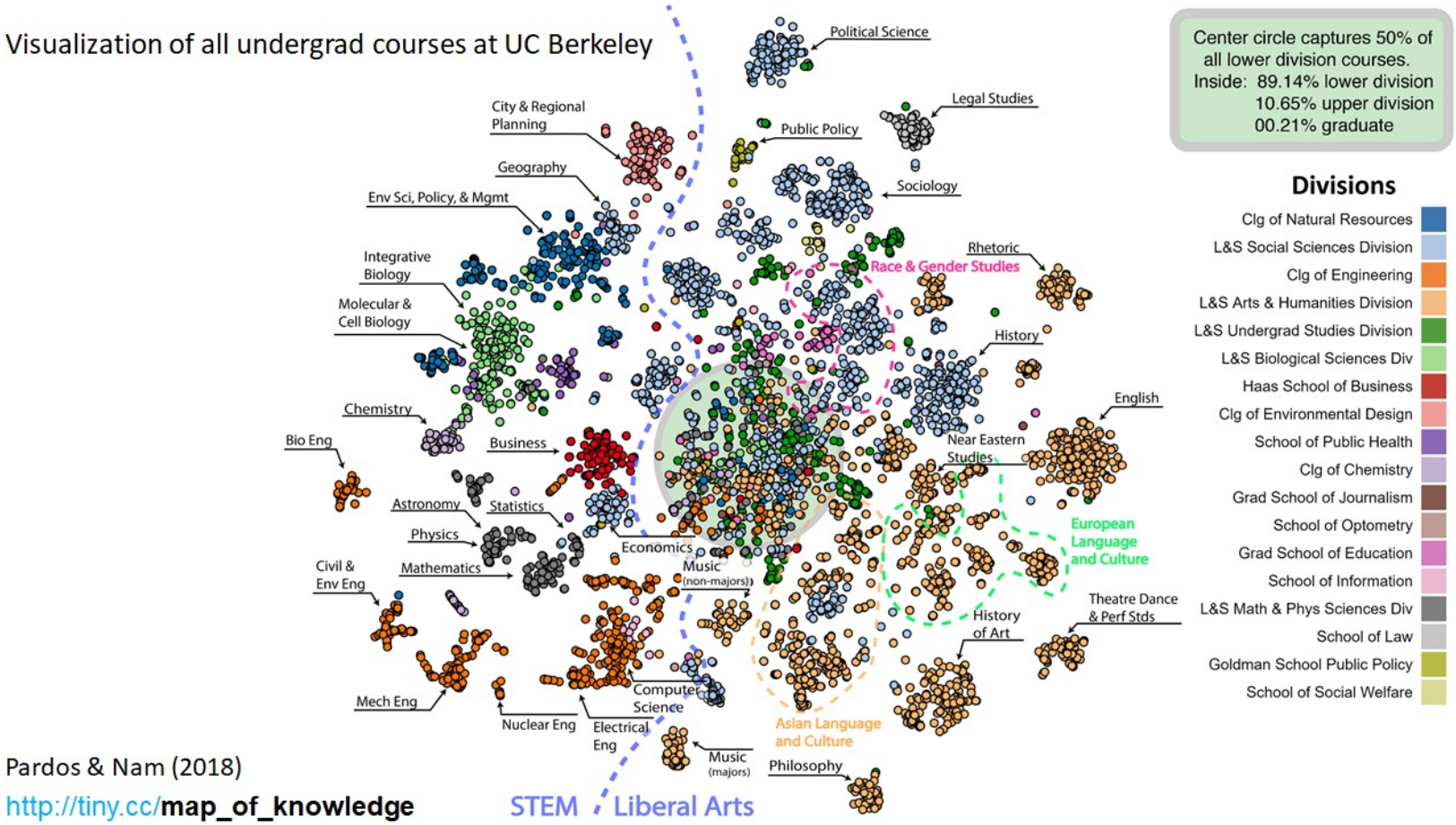
- Racial composition
- School quality

Consequences

- Encourages segregation
- Contradict claims of equity & access to education

Recommending Undergraduate Courses

Pardos & Nam 2018



Pardos & Nam (2018)
http://tiny.cc/**map_of_knowledge**

- Student Transcripts
- Course catalog descriptions
- Doc2Vec model, comparing similarities

Beyond Text Data -> Natural Language

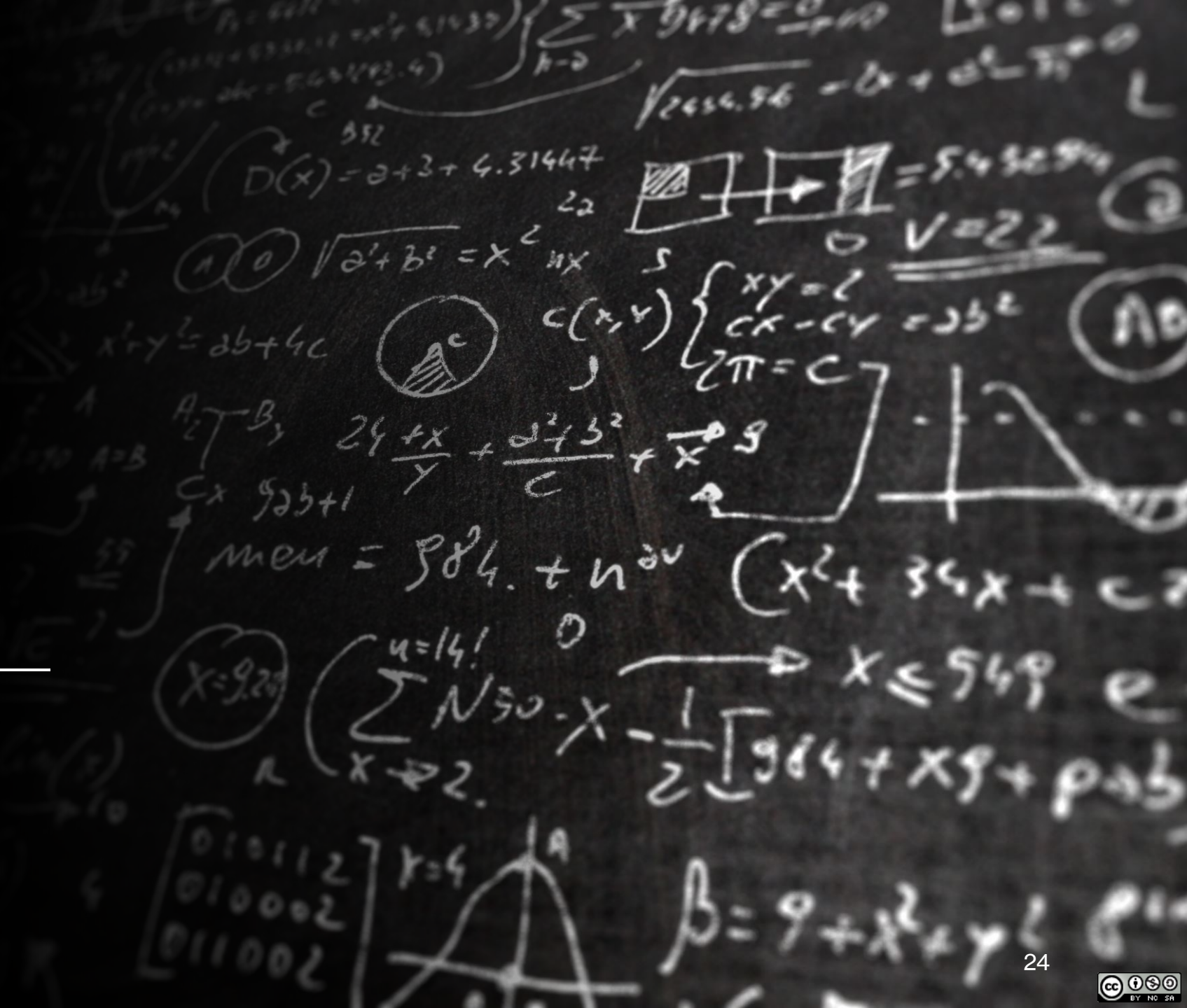
Voigt, et al. 2017; Camp et al. 2020; Rho et al. (in prep.)

Analyzing Transcripts of police officer in traffic stops

Prosody (tone of voice) of police officer in traffic stops

Working with Text Data

Preface:
Working with text in English



Typical Data Sources

Open-ended survey

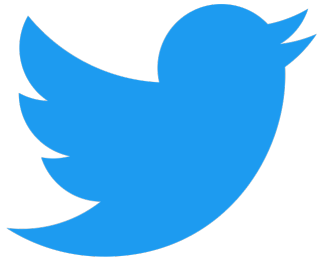
Web sources

Websites

Social Media

Scraping Social Media

- Social media platforms offer Application Programming Interfaces (APIs)
  - Request permission for academic research typically gives free and extended access



Scraping Online Text Data

ProQuest TDM Studio

Requests: pull full HTML code

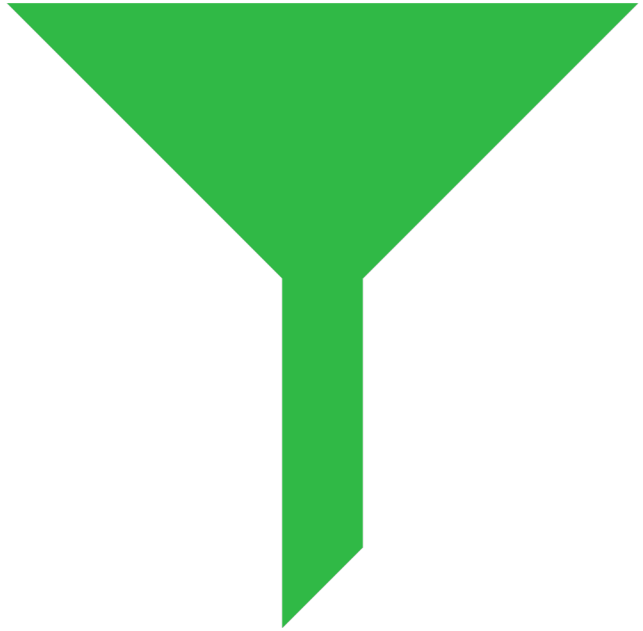
Beautiful Soup: parse through HTML code

(Demo)

Text Data Cleaning

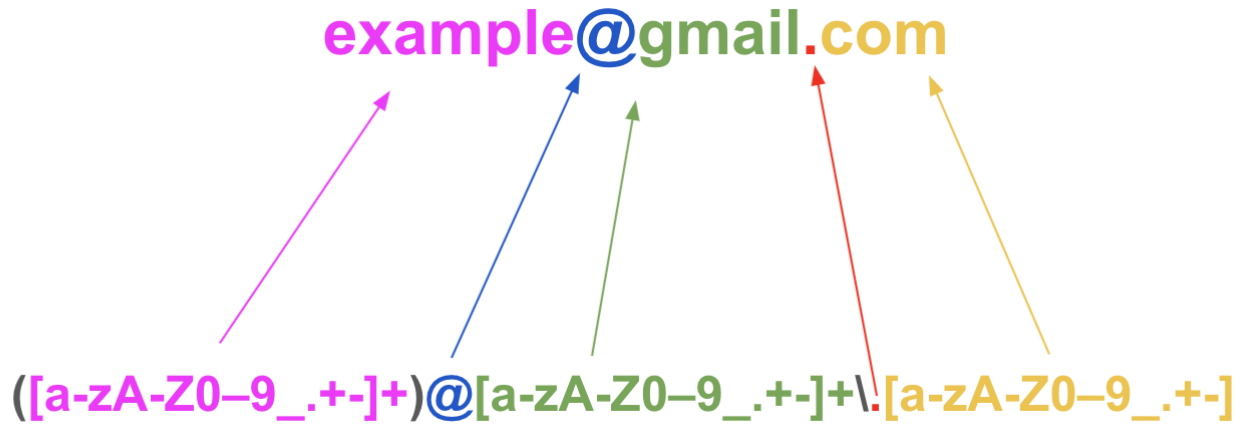
Typical data cleaning methods:

- Lemmatization
- Filter Stopwords
- Tokenization
  - Convert to lower case
  - Remove symbols/punctuations
- Tagging Parts of Speech
- Recognizing Named Entities



Pattern Matching

- Use Regular Expression (RegEx) to match symbols/words/patterns
  - Extract
  - Replace



Cleaning: Lemmatization

- Multiple words can stem from 1 root
  - Reducing duplicates/overlaps in analysis
  - ”Return” each word back to a basic form

Example:

epithelia -> epithelium (quantity)

rose (v.) -> rise (verb tenses)

Cleaning: Filter Stopwords

- Filtering out stopwords that:
  - Might not contribute to context/meaning
  - May throw off frequency analysis

Examples: The, a, me, I, he, her, them, doing, and, with, by, should

More Cleaning Techniques

Tokenization

- Segmenting sentences and words so they are treated as different parts of text
- Helps speeding up other cleaning steps when words are recognized as tokens

Tagging Parts of Speech (POS)

- Recognizing different parts of the sentence to better inform structure/meaning

Recognizing Named Entity

- Special case of POS: preserving certain proper nouns/named entities

(Demo)

Analyzing Text Data

Sentiment Analysis

Frequency Analysis

Topic Modeling

Sentiment Analysis

- Learning & classifying affect/polarity of written text
- Aggregate across words to identify overall sentiment
- 2 Approaches
  - Pre-trained models
  - Train your own models

Examples: VADAR, TextBlob

- VADAR: Valence Aware Dictionary for Sentiment Reasoning
  - Pre-trained polarity dictionary
- TextBlob
  - Pre-trained polarity & subjectivity dictionary

(Demo)

Frequency Analysis

- Can the frequency of certain words reveal information of a corpus?
  - Yes!
- Other uses
  - Document indexing/importance by keywords

Example: Bag of Words

- Keeps track of term occurrences in documents
- Does not consider positions or contexts

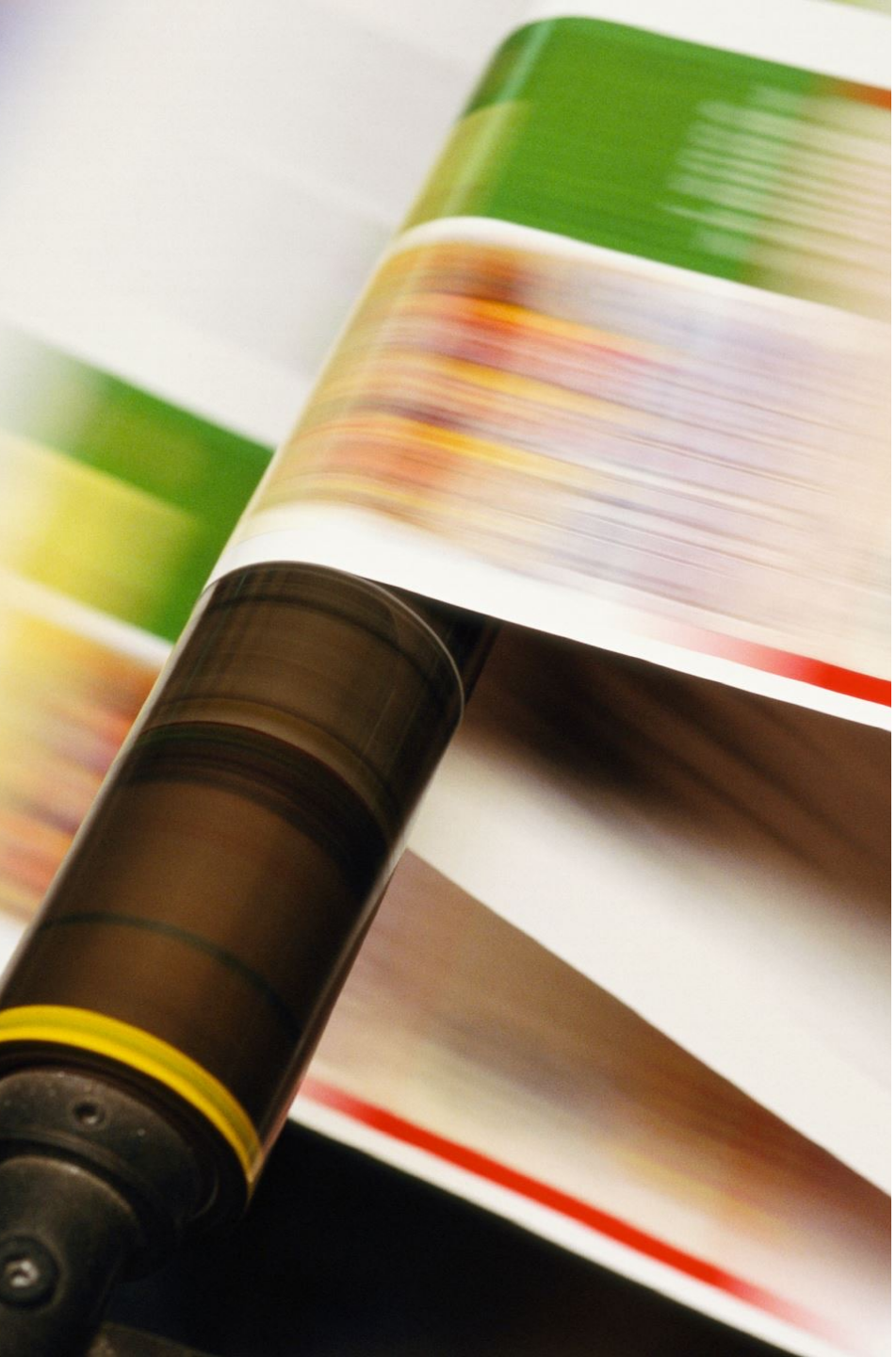
Document	the	cat	sat	in	hat	with
the cat sat	1	1	1	0	0	0
the cat sat in the hat	2	1	1	1	1	0
the cat with the hat	2	1	0	0	1	1

Example: TF-IDF

- Term Frequency-Inverse Document Frequency
  - IDF: takes in account common words across all documents (stopword occurrences)
- Calculate relative importance of a term in corpus of documents
- Signals similarities of documents (with similar term occurrences)

$$w_{i,j} = tf_{i,j} \times log(\frac{N}{df_i})$$

(Demo)



Topic Modeling

- Capturing topics and themes from text
- Can utilize frequency or context of words
- Can integrate with unsupervised ML (clustering) to identify topic clusters
- Can be used to predict context based on words, or use context to generate words

Frequency: Latent Dirichlet Allocation

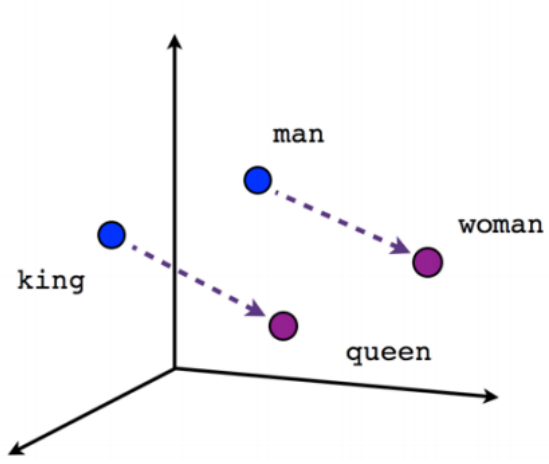
Builds topic per document

Modeled with Dirichlet distribution

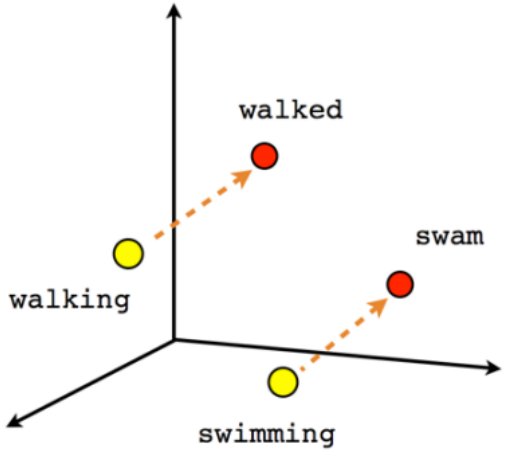
Visualize frequency of topics across documents

(Demo)

Context: Word2Vec



Male-Female



Verb tense

Context: Vector Relationships

Mikolov, Yih, & Zweig 2013

Table 8: *Examples of the word pair relationships, using the best word vectors from Table 4 (Skip-gram model trained on 783M words with 300 dimensionality).*

Relationship	Example 1	Example 2	Example 3
France - Paris	Italy: Rome	Japan: Tokyo	Florida: Tallahassee
big - bigger	small: larger	cold: colder	quick: quicker
Miami - Florida	Baltimore: Maryland	Dallas: Texas	Kona: Hawaii
Einstein - scientist	Messi: midfielder	Mozart: violinist	Picasso: painter
Sarkozy - France	Berlusconi: Italy	Merkel: Germany	Koizumi: Japan
copper - Cu	zinc: Zn	gold: Au	uranium: plutonium
Berlusconi - Silvio	Sarkozy: Nicolas	Putin: Medvedev	Obama: Barack
Microsoft - Windows	Google: Android	IBM: Linux	Apple: iPhone
Microsoft - Ballmer	Google: Yahoo	IBM: McNealy	Apple: Jobs
Japan - sushi	Germany: bratwurst	France: tapas	USA: pizza

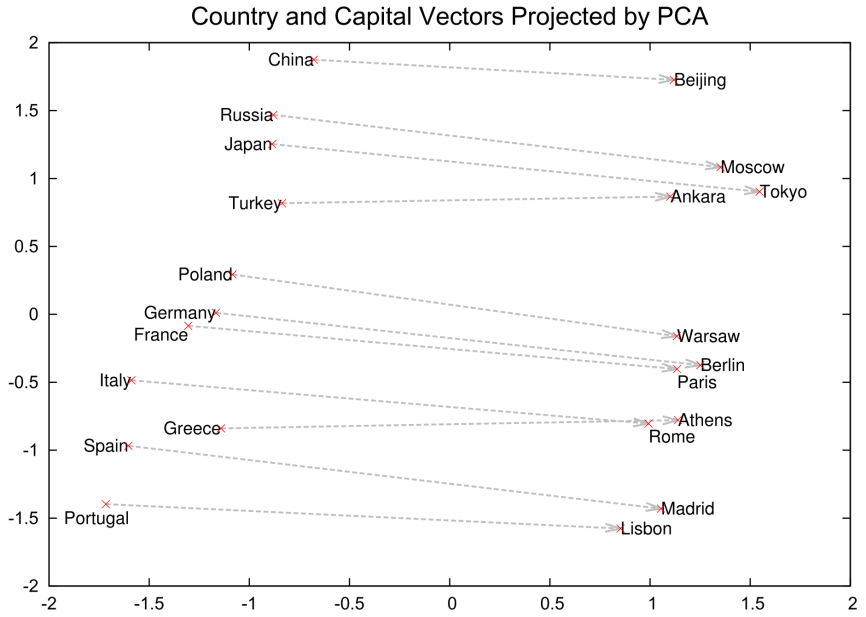


Figure 2: Two-dimensional PCA projection of the 1000-dimensional Skip-gram vectors of countries and their capital cities. The figure illustrates ability of the model to automatically organize concepts and learn implicitly the relationships between them, as during the training we did not provide any supervised information about what a capital city means.

(Demo)

Comparing Methods

	Advantages	Disadvantages
Sentiment Analysis	Good for identifying discrete affects and polarity in text; Easy to scale up for social media data	Does not take into account satire/sarcasm (limited understanding of contexts)
Frequency Analysis	Good for identify important terms across and within documents	Does not take into account words with multiple meanings; no consideration for contexts
Topic Modeling	Good for in-depth identification of topics across a large corpus of documents	Requires a large amount of text data, not reliable with small corpus (unless using pre-trained model)

Resources for Learning/Using TDM/NLP

- nltk: natural language toolkit
- gensim: topic modeling
- scikit-learn: data processing, machine learning
- Towards Data Science (Medium blog)

Thank You!

Questions? Email: me@yuyangzhong.com

Credit:

- Zachary Pardos, Ph.D.
- Towards Data Science, Medium Blog
- UC Berkeley Library
- University of Sydney Library