# Uncovering Social Science Literatures: Harnessing the Power of Google Books Data

Violet Huang, Ryan Liang, April Wang

# The Problem at Hand

The task of scraping and analyzing social science related books using the Google Books API carries social significance as it presents valuable opportunities for information mining. By accessing and analyzing a vast collection of books using large scale method, we can explore valuable insights and patterns in social science topics.

# Part 1: Collect Data Using Google Books API

Parallelization Strategy:

- Collect all social science-related books → Collect books from a single category
- Collect all books from a single category → Collect a subset of 40 books from the category
- Parallelize with customized batch, Lambda function, and step function – search term and startIndex

# Part 1: Collect Data Using Google Books API

Limitation:

- Scraping books using a uniform max start index may lead to limitations in terms of representativeness
- To address this, try to parallelize the retrieval of specific max startIndex for each category using lambda function
- Distributing the requests through lambda functions, consistently encountered the 429 status code. This rate limitation imposed constraints on the data collection process.

# Part 1: Collect Data Using Google Books API

```
OrderedDict([('book_id', 'uy5lAAAACAAJ'),

('book_info', '{"book_id": "uy5lAAAACAAJ", "title": "Sociology", "subtitle": "A Global Introduction",

"authors": "John J. Macionis", "publisher": "Prentice Hall", "published_date": "1997-01-17",

"description": "An introductory text covering the foundations of sociology and research strategies, the ideas of key thinkers such as Karl Marx
and Max Weber, social inequality and stratification, institutions, and global social change. Features color photos, topic boxes, chapter- opening
vignettes, sociological maps, questions, and summaries. This fifth edition includes new US maps, a chapter on the natural environment, and
expanded discussion on topics such as suicide, Asian Americans, and feminist research methods. Annotation copyright by Book News, Inc.,
Portland, OR",

"Categories": “sociology”,

"imageLinks": {"smallThumbnail":
"http://books.google.com/books/content?id=uy5lAAAACAAJ&printsec=frontcover&img=1&zoom=5&source=gbs_api", "thumbnail":
"http://books.google.com/books/content?id=uy5lAAAACAAJ&printsec=frontcover&img=1&zoom=1&source=gbs_api"}}')])
```

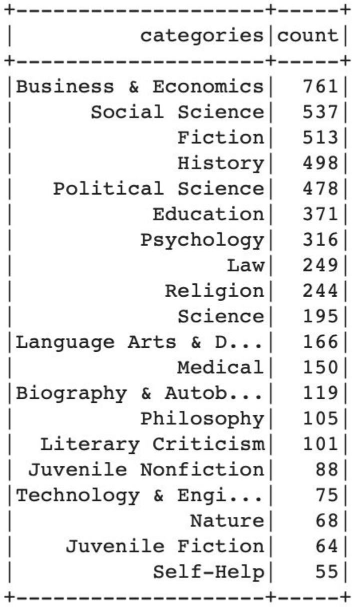
# Part 2: Supervised Prediction Task

Parallelization Strategy using Spark:

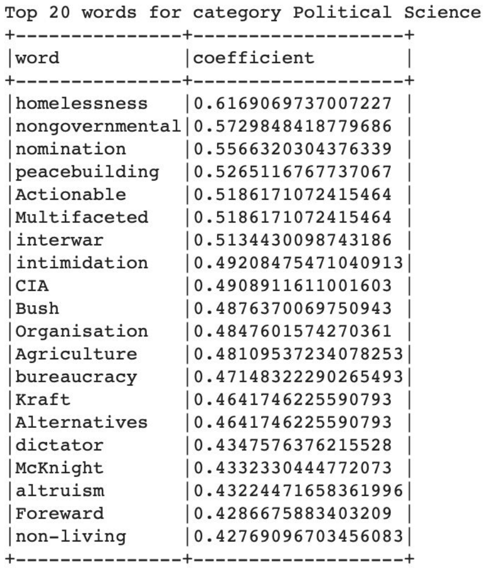
- Data Reading
- Data Preprocessing
- Sampling
- Model Training
- Hyperparameter Tuning
- Model Evaluation

# Part 2: Supervised Prediction Task

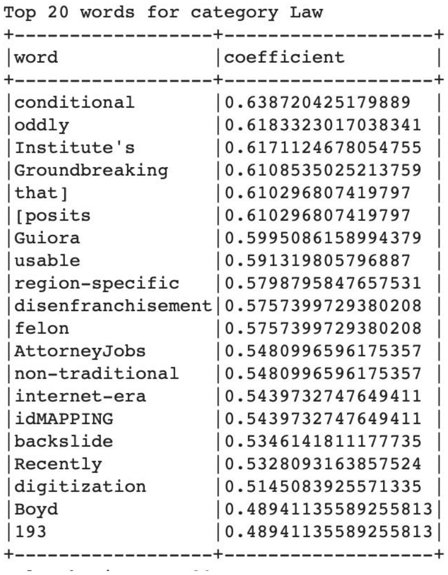
Predicting the category of books based on their descriptions and subsequently discern the most predictive words for each category



```
+--------------------+-----+
|          categories|count|
+--------------------+-----+
|Business & Economics|  761|
|      Social Science|  537|
|             Fiction|  513|
|             History|  498|
|   Political Science|  478|
|           Education|  371|
|          Psychology|  316|
|                 Law|  249|
|            Religion|  244|
|             Science|  195|
|Language Arts & D...|  166|
|             Medical|  150|
|Biography & Autob...|  119|
|          Philosophy|  105|
|  Literary Criticism|  101|
| Juvenile Nonfiction|   88|
|Technology & Engi...|   75|
|              Nature|   68|
|    Juvenile Fiction|   64|
|           Self-Help|   55|
+--------------------+-----+
```



```
Top 20 words for category Political Science
+---------------+-------------------+
|word           |coefficient        |
+---------------+-------------------+
|homelessness   |0.6169069737007227 |
|nongovernmental|0.5729848418779686 |
|nomination     |0.5566320304376339 |
|peacebuilding  |0.5265116767737067 |
|Actionable     |0.5186171072415464 |
|Multifaceted   |0.5186171072415464 |
|interwar       |0.5134430098743186 |
|intimidation   |0.49208475471040913|
|CIA            |0.4908911611001603 |
|Bush           |0.4876370069750943 |
|Organisation   |0.4847601574270361 |
|Agriculture    |0.48109537234078253|
|bureaucracy    |0.47148322290265493|
|Kraft          |0.4641746225590793 |
|Alternatives   |0.4641746225590793 |
|dictator       |0.4347576376215528 |
|McKnight       |0.4332330444772073 |
|altruism       |0.43224471658361996|
|Foreward       |0.4286675883403209 |
|non-living     |0.42769096703456083|
+---------------+-------------------+
```



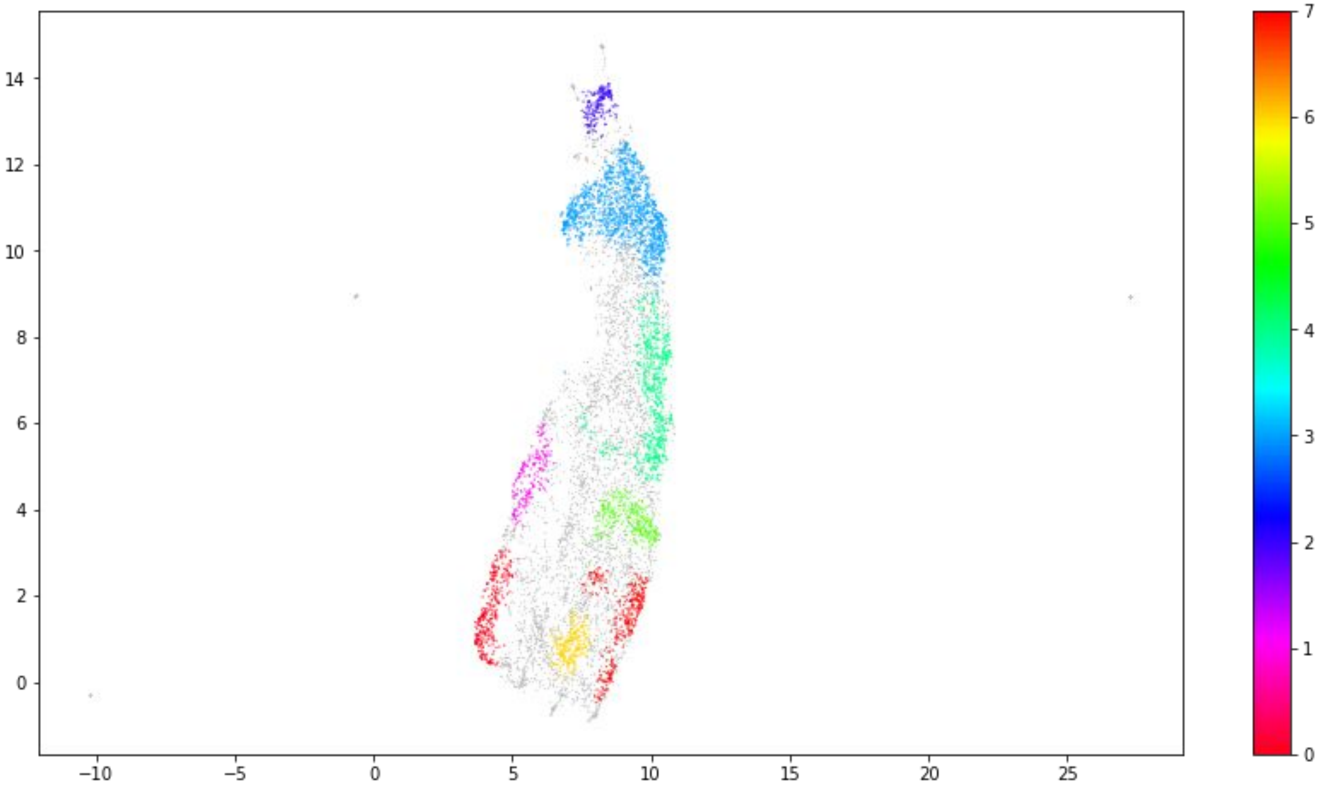
```
Top 20 words for category Law
+------------------+-------------------+
|word              |coefficient        |
+------------------+-------------------+
|conditional       |0.638720425179889  |
|oddly             |0.6183323017038341 |
|Institute's       |0.6171124678054755 |
|Groundbreaking    |0.6108535025213759 |
|that]             |0.610296807419797  |
|[posits           |0.610296807419797  |
|Guiora            |0.5995086158994379 |
|usable            |0.591319805796887  |
|region-specific   |0.5798795847657531 |
|disenfranchisement|0.5757399729380208 |
|felon             |0.5757399729380208 |
|AttorneyJobs      |0.5480996596175357 |
|non-traditional   |0.5480996596175357 |
|internet-era      |0.5439732747649411 |
|idMAPPING         |0.5439732747649411 |
|backslide         |0.5346141811177735 |
|Recently          |0.5328093163857524 |
|digitization      |0.5145083925571335 |
|Boyd              |0.48941135589255813|
|193               |0.48941135589255813|
+------------------+-------------------+
```

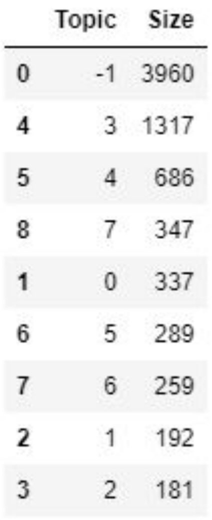
# Part 3: Natural Language Processing with BERTopic

Parallelization Strategy:

- Transfer Learning with RoBerta → Spark NLP
- Dimension Reduction with UMAP → cuML with GPU
- Hierarchical Clustering with HDBSCAN → cuML with GPU

# Part 3: Natural Language Processing with BERTopic





| | Topic | Size |
|---|---|---|
| 0 | -1 | 3960 |
| 4 | 3 | 1317 |
| 5 | 4 | 686 |
| 8 | 7 | 347 |
| 1 | 0 | 337 |
| 6 | 5 | 289 |
| 7 | 6 | 259 |
| 2 | 1 | 192 |
| 3 | 2 | 181 |

# Part 3: Natural Language Processing with BERTopic

Salient Keywords of Interesting BERTopic Clusters:

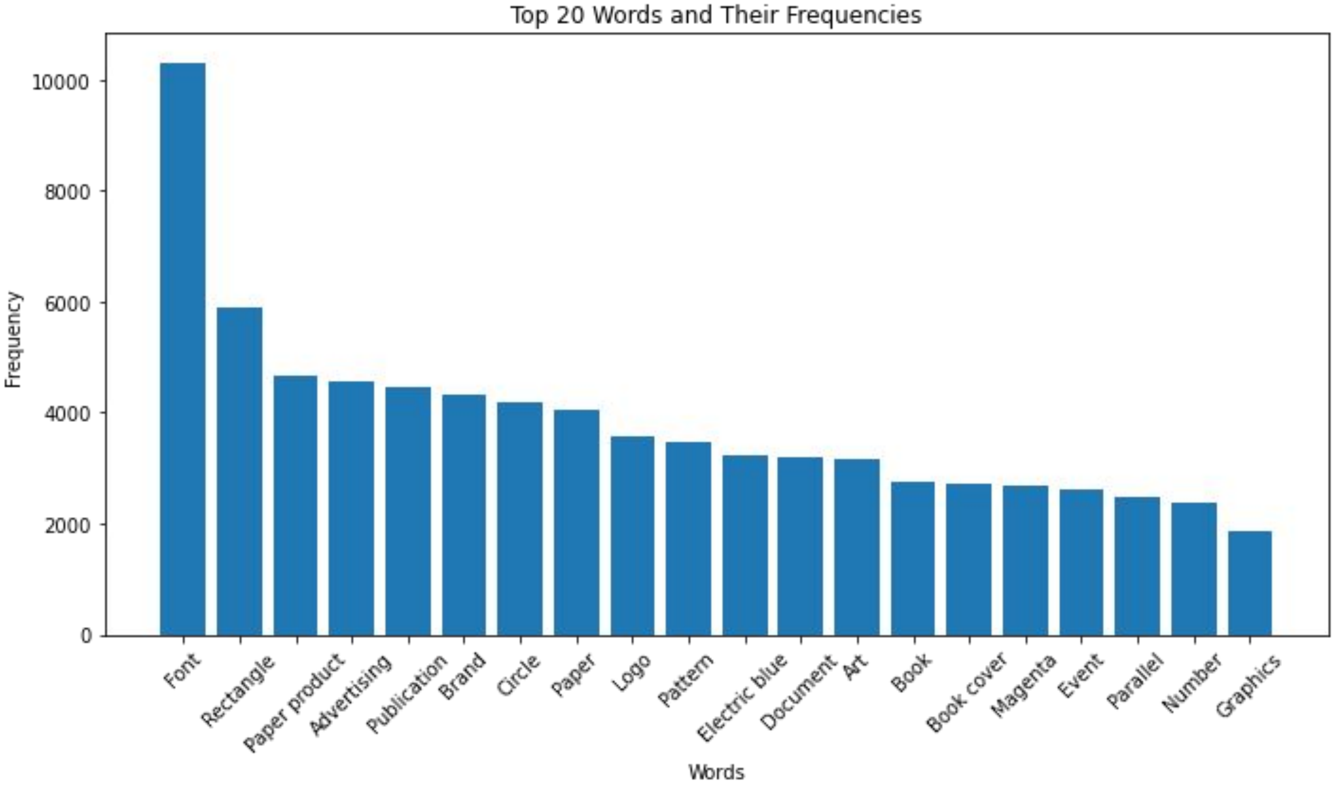
- York, novel, stori, bestsel, murder
- Novel, murder, stori, one, man, young, friend, woman
- Examin, explor, present, collect, provid
- Polici, research, field, commun, environment, health
- Social, polit, polici, law, legal, examin
- Polit, american, argu, law, race, cultur

# Part 4: Computer Vision with Google Cloud API

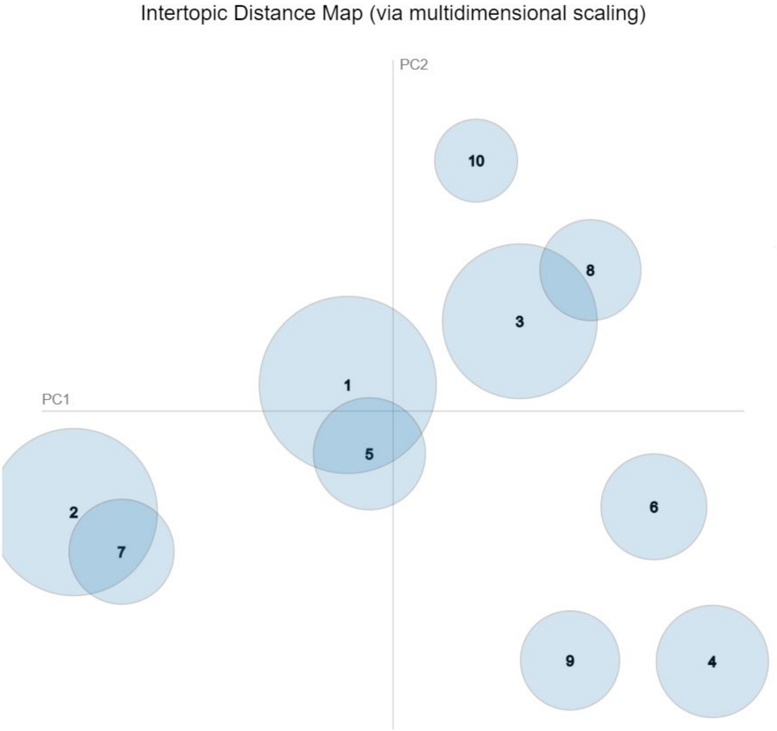
Parallelization Strategy:

- Google Cloud Vision Image Object Recognition API → AWS Lambda
- Result Analysis → Dask
- Pattern Analysis of Cover Design → Multicore LDA with Gensim

# Part 4: Computer Vision with Google Cloud API



# Part 4: Computer Vision with Google Cloud API



Interactive version can be found at Google Vision API Cover Analysis.html

# Part 4: Computer Vision with Google Cloud API

Interesting LDA Cluster of Cover Design Element Patterns:

- Grass, Plant, Landscape, Soil, Tree
- Photo caption, Happy, Gesture, Fun, Formal Wear
- Art, History, Facade, Building, City, Landscape
- Fashion Accessory, Liquid, Drink, Eyelash, Liquer

# Thank You!