Similarity Search Tool For The U.S. Census Bureau's Working Papers

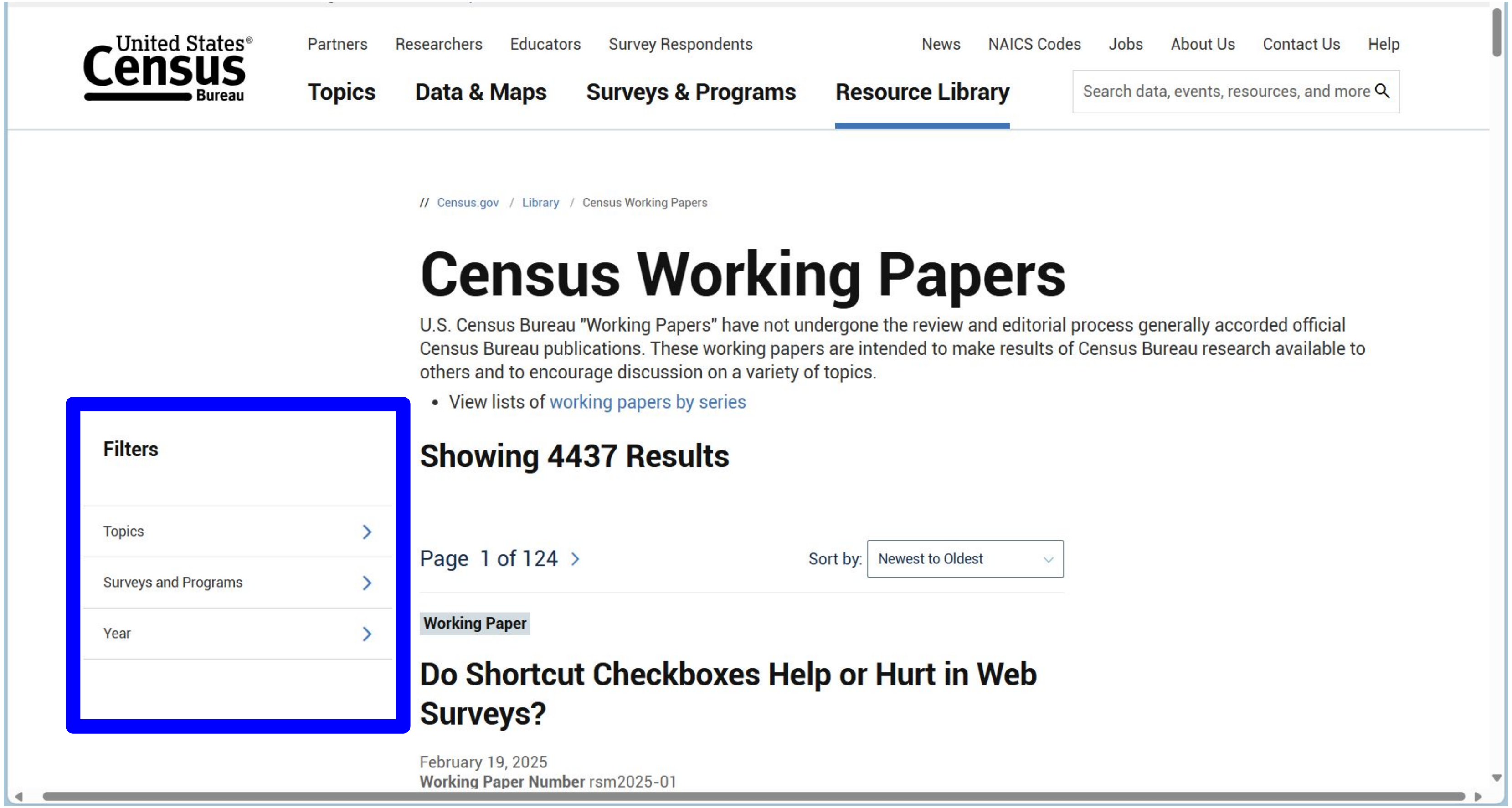
Paper Chaser - Ethan Crouse, Eric Zou, and Rahul Ramakrishnan

Acknowledgements

Throughout the semester, we have received assistance from and would like to thank:

Dr. Rae Ellis, Maxwell Hope, and Ian Le from the U.S. Census Bureau: For taking time to meet with us weekly and providing insightful feedback

Hayden Ringer: For guidance throughout the semester and helping us brainstorm ideas and crucial questions



Census Working Papers

U.S. Census Bureau "Working Papers" have not undergone the review and editorial process generally accorded official Census Bureau publications. These working papers are intended to make results of Census Bureau research available to others and to encourage discussion on a variety of topics.

- View lists of working papers by series

Showing 4437 Results

Do Shortcut Checkboxes Help or Hurt in Web Surveys?

February 19, 2025
Working Paper Number rsm2025-01

Problem Statement

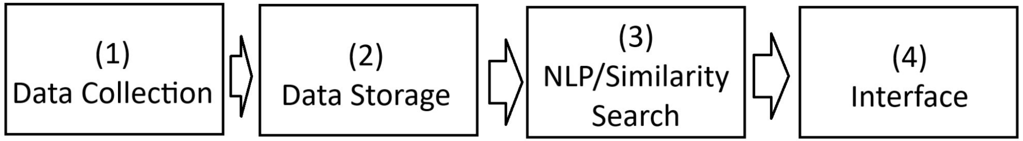
Problem:

- 4000+ working papers that are open to the public, but limited searchability
- Optimizing searchability can lead to advances in research that aren't possible with current limitations
- Building a tool to enhance searchability using machine learning to provide faster, easier, and more accurate searches.

Scope:

- Custom searches on U.S. Census Bureau working papers that have abstracts and downloadable pdfs

Components & Criteria



Data Collection

- Download all 4000+ working papers and their metadata

Data Storage

- Store the data to be easily access by our natural language processing tool

Natural language processing

- Effectively finds similar articles with one search query

User Interface

- Clean, intuitive design with easy navigation

Literature review

- **Gahman et al.(2023) A Comparison of Document Similarities Algorithms**
  - The main focus of our literature review was finding suitable NLP algorithms
  - Five NLP-based document similarity algorithms and their accuracies
  - MT, SEMB, WMD, SNK, LSTM
  - Caused us to look into MT-DNN
- **Xiaodong et al. (2020) The Microsoft Toolkit of Multi-Task Deep Neural Networks for Natural Language Understanding**
  - Training models to multiple tasks
  - We decided not to directly use it, but it brought us to using BERT
- **Reimers et al. (2019) Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks**
  - Bidirectional Encoder Representations from Transformers
  - Best balance of accuracy and speed for us. We don't need the multi-task part
  - Generates storable embeddings allowing for processing to be done before searches
  - **Takeaway**: we want to use SBERT or a similar model

Survey of Methods - NLP

Term Frequency-Inverse Document Frequency (TF-IDF)	Fast and requires little storage, but low accuracy
Word2Vec	Better than TF-IDF but still no context awareness
Bidirectional Encoder Representations from Transformers	Contextually aware and gives better results, but is computationally expensive
Ada	Most accurate but most computationally expensive and costs money (OpenAI). Overkill for us

Solution Approach

Natural Language Processing

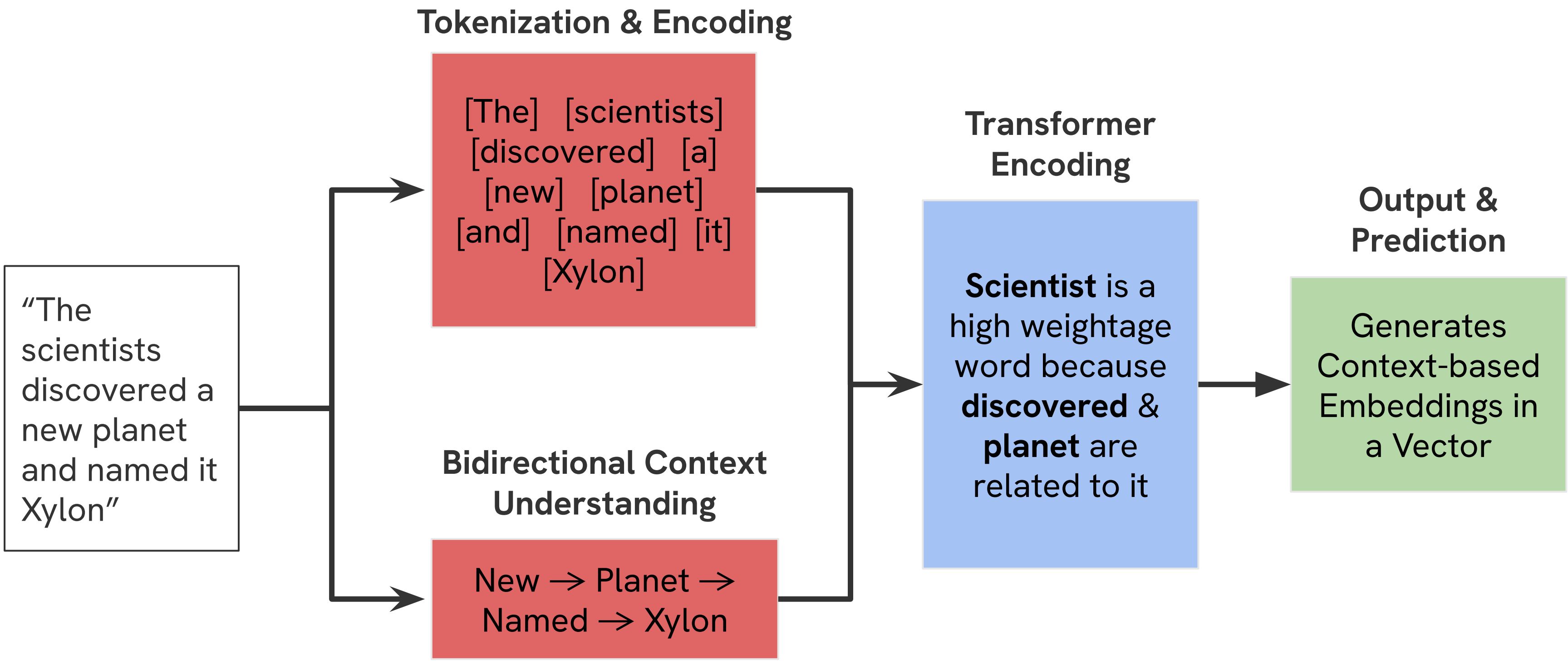
We decided to go with an embedding based model for better search results. From our research the best free architecture to use would be a BERT based model, specifically all-MiniLM-L6-v2

Storage

Because we chose an embedding based model we will need a vectorstore. We decided to use FAISS as it was the simplest and easily scalable.

User Interface

For modularity and simplicity of the overall product, we will use Flask and React.js for the front-end


“The scientists discovered a new planet and named it Xylon”
Tokenization & Encoding
[The] [scientists] [discovered] [a] [new] [planet] [and] [named] [it] [Xylon]
Bidirectional Context Understanding
New → Planet → Named → Xylon
Transformer Encoding
Scientist is a high weightage word because **discovered** & **planet** are related to it
Output & Prediction
Generates Context-based Embeddings in a Vector

Results

Results | Data Collection

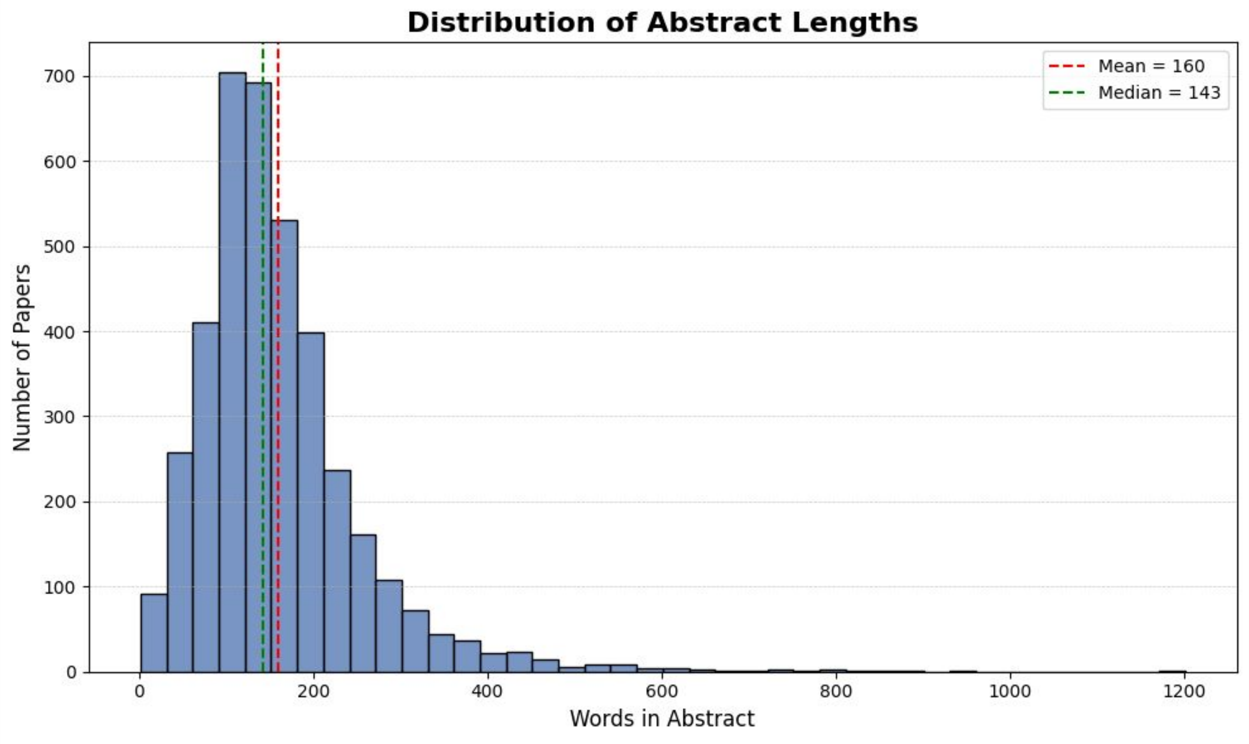
Data Collected

- 3996 Useable papers out of 4447 total
- Python based scraper using Requests, Selenium, and Beautifulsoup
- 17 hours to run if following guidelines, 30 minutes if not.
- Key attributes (title, link to paper, dates, authors, abstract, files)

Database Management

- Currently stored in a local CSV file
- Embeddings will be stored in FAISS index
- Stores Attributes in Table Format

Results | Data Collection



Results | NLP

Algorithms & Schematics

- High-dimensional vectorization using SentenceTransformer ('all-MiniLM-L6-v2') stored in a FAISS index
- all-MiniLM-L6-v2 was selected for its balance of high accuracy and extremely fast inference speed, making it ideal for large-scale search applications

Results & Analysis

- Search returns top-matching research papers based on semantic similarity to the query.
- Each result includes the title, abstract snippet, PDF link (if available), and similarity score indicating relevance.

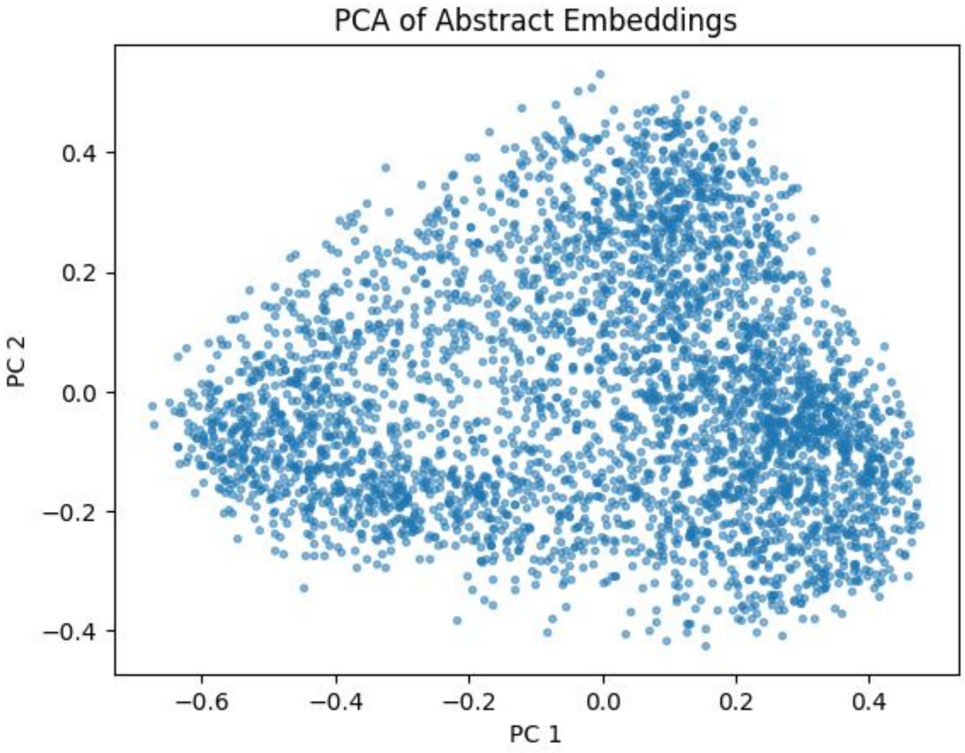
Precision@5 Scoring Metric:

	Income Disparity	Small Business	Administrative Records
TF-IDF	0.20	0.40	0.40
Word2Ve c	0.60	0.0	0.0
Doc2Vec	0.40	0.0	0.20
SBERT	0.80	0.60	0.80

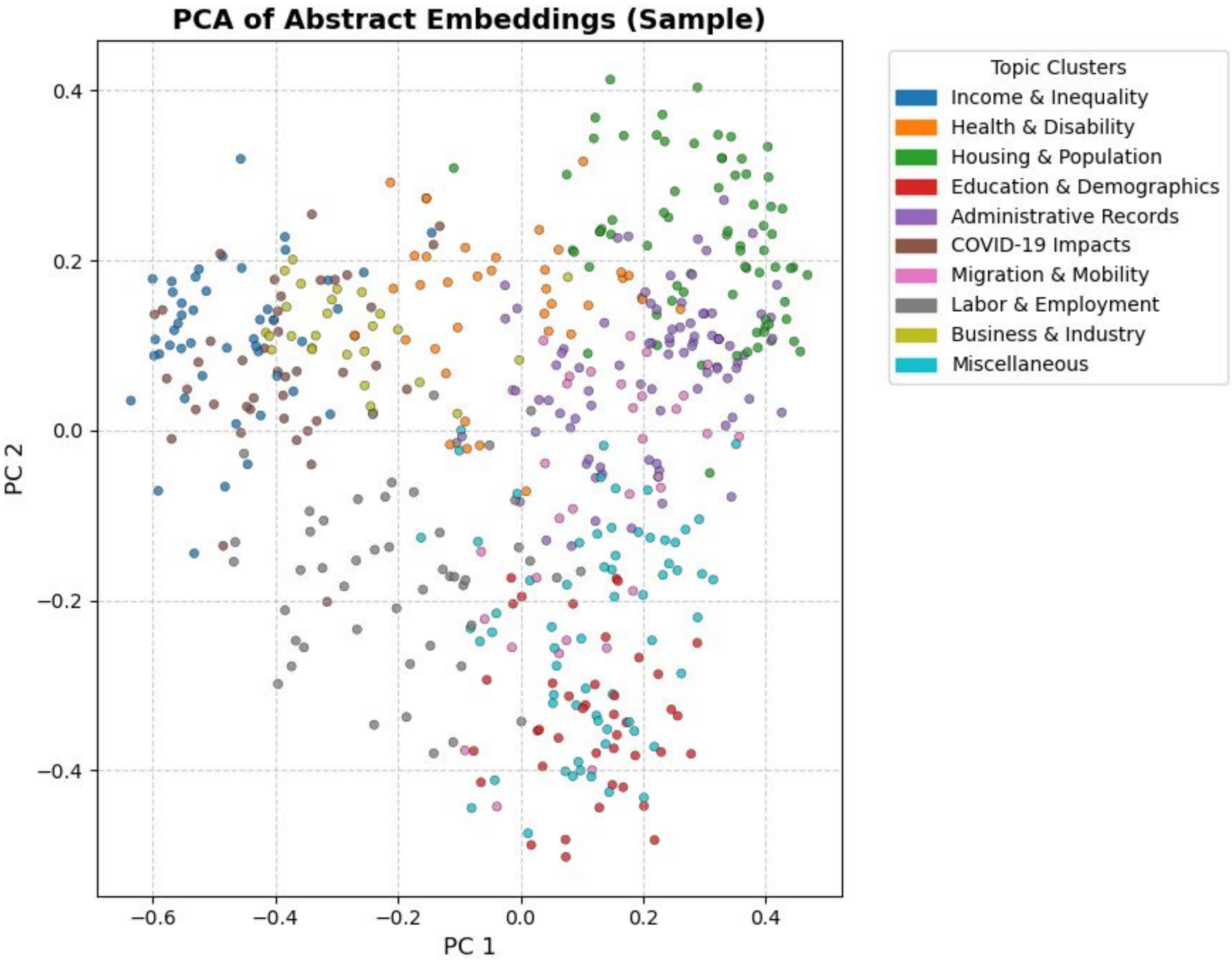
Models and Times to Embed Vectors on Google Colab

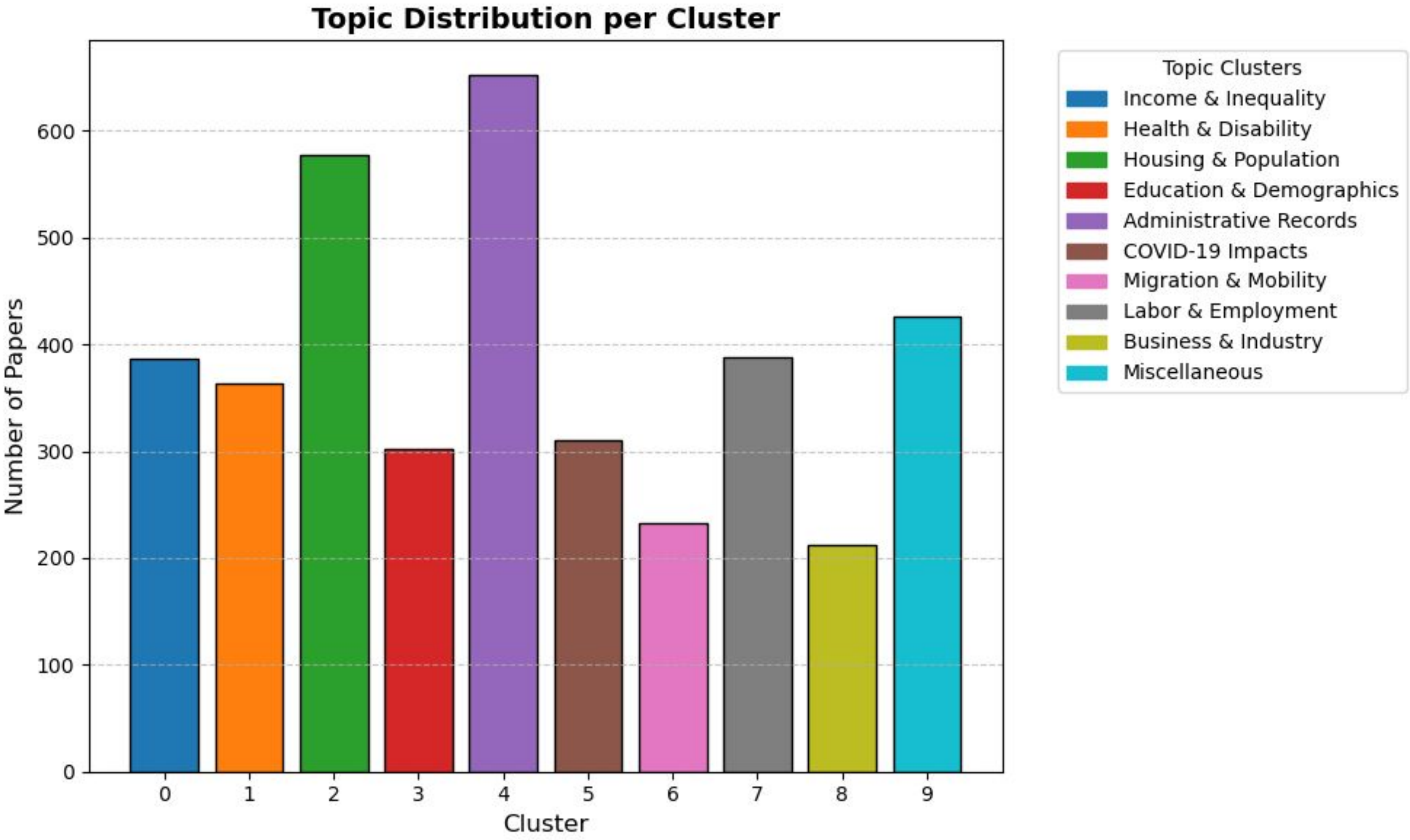
Model	**Time to Embed (seconds)**
Word2Vec	1.42
Doc2Vec	26.17
SBERT	370.47

Results | NLP

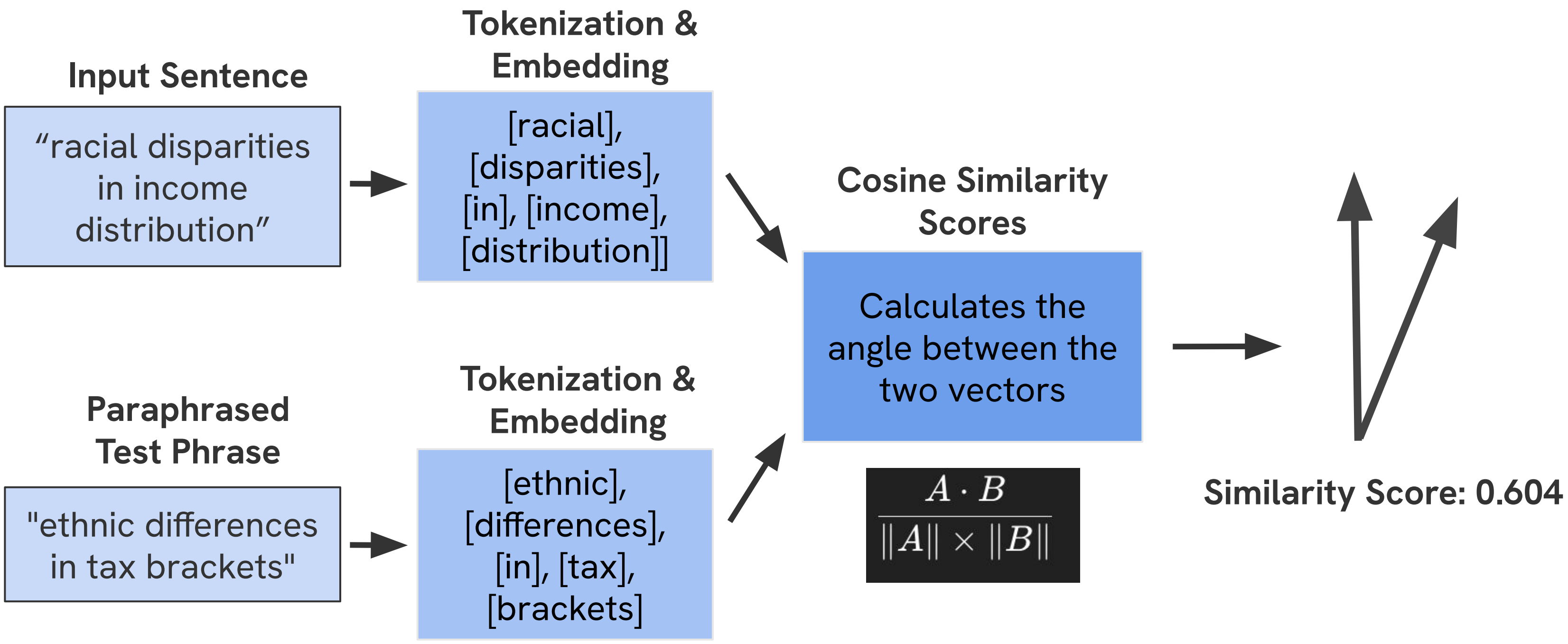


- PCA visualization showed clear topic clustering in the SBERT embeddings.
- Cosine distance flagged outlier papers outside typical Census themes.
- Outliers included niche topics like autocovariance estimation and statistical interviews.


PCA of Abstract Embeddings (Sample)
0.4
0.2
0.0
−0.2
−0.4
PC 2
−0.6
−0.4
−0.2
0.0
0.2
0.4
PC 1
Topic Clusters
Income & Inequality
Health & Disability
Housing & Population
Education & Demographics
Administrative Records
COVID-19 Impacts
Migration & Mobility
Labor & Employment
Business & Industry
Miscellaneous


Topic Distribution per Cluster
Number of Papers
0
100
200
300
400
500
600
Cluster
0
1
2
3
4
5
6
7
8
9
Topic Clusters
🟦 Income & Inequality
🟧 Health & Disability
🟩 Housing & Population
🟥 Education & Demographics
🟪 Administrative Records
🟫 COVID-19 Impacts
Migration & Mobility
Labor & Employment
Business & Industry
Miscellaneous

Results | NLP



Input Sentence

“racial disparities in income distribution”

Tokenization & Embedding

[racial], [disparities], [in], [income], [distribution]]

Paraphrased Test Phrase

"ethnic differences in tax brackets"

Tokenization & Embedding

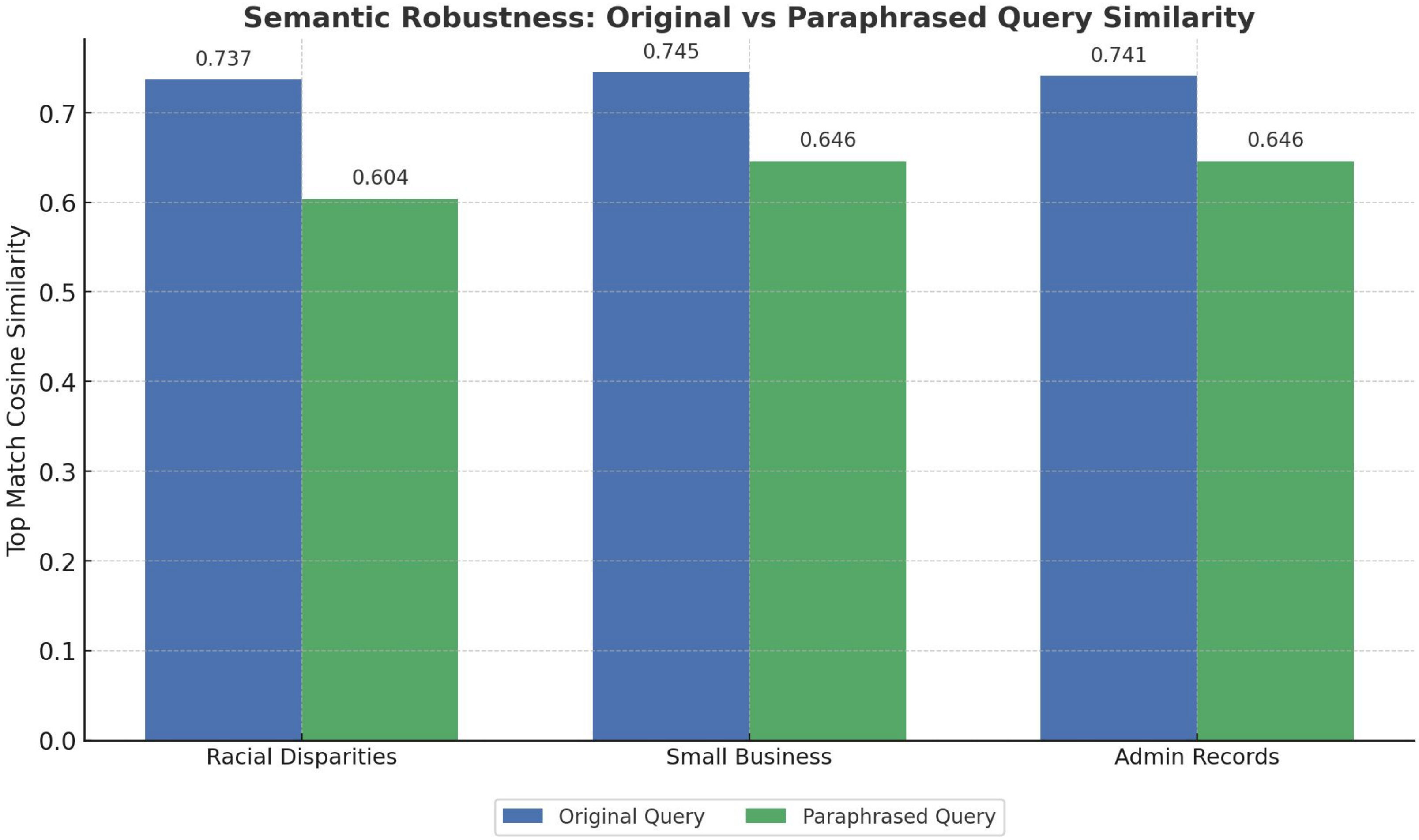
[ethnic], [differences], [in], [tax], [brackets]

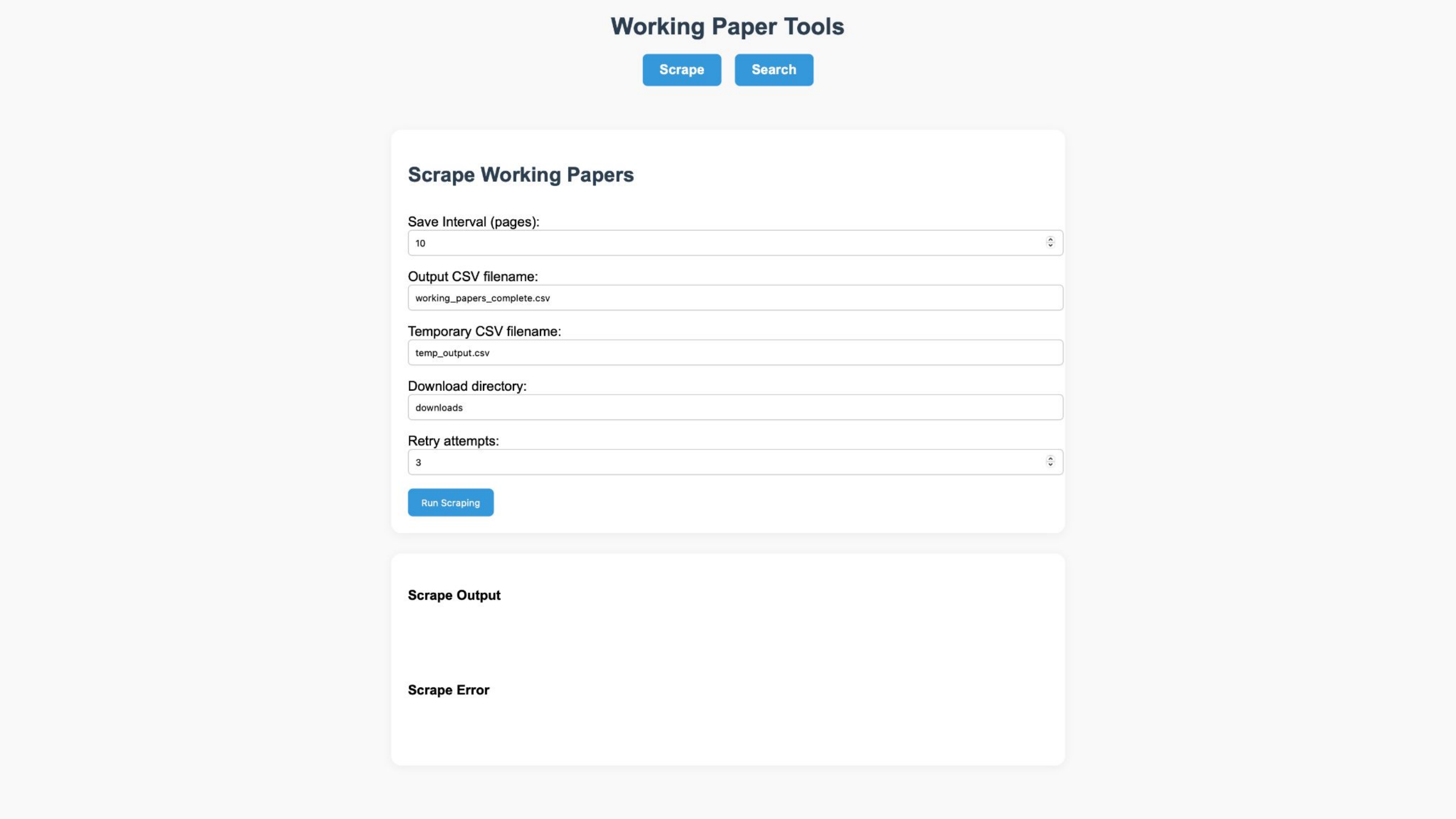
Cosine Similarity Scores

Calculates the angle between the two vectors

$$\frac{A \cdot B}{\|A\| \times \|B\|}$$

Similarity Score: 0.604


Semantic Robustness: Original vs Paraphrased Query Similarity
Top Match Cosine Similarity
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.737
0.604
0.745
0.646
0.741
0.646
Racial Disparities
Small Business
Admin Records
Original Query
Paraphrased Query



Working Paper Tools

Scrape Working Papers

Save Interval (pages):
10

Output CSV filename:
working_papers_complete.csv

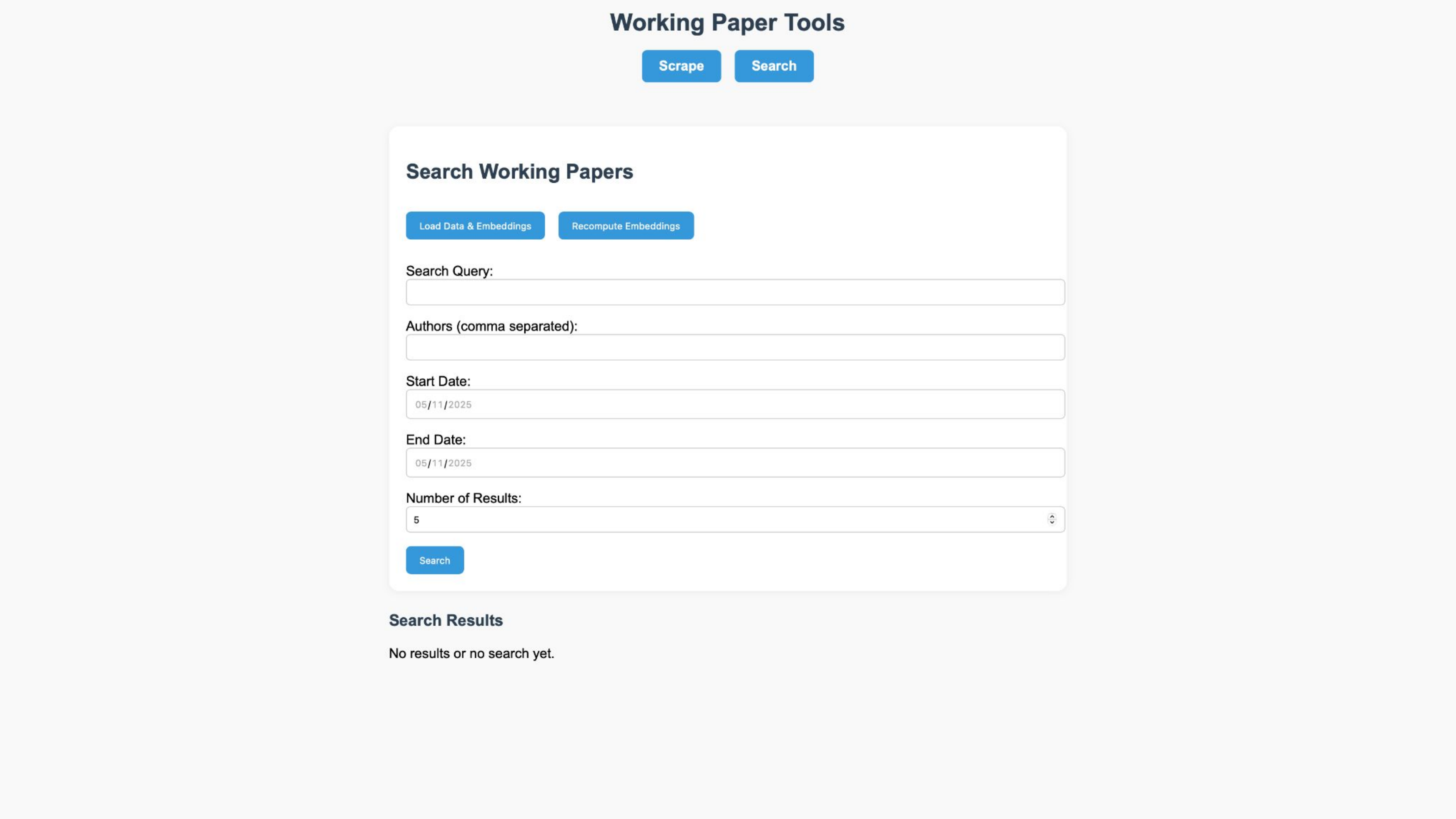
Temporary CSV filename:
temp_output.csv

Download directory:
downloads

Retry attempts:
3

Scrape Output

Scrape Error



Working Paper Tools

Search Working Papers

Search Query:

Authors (comma separated):

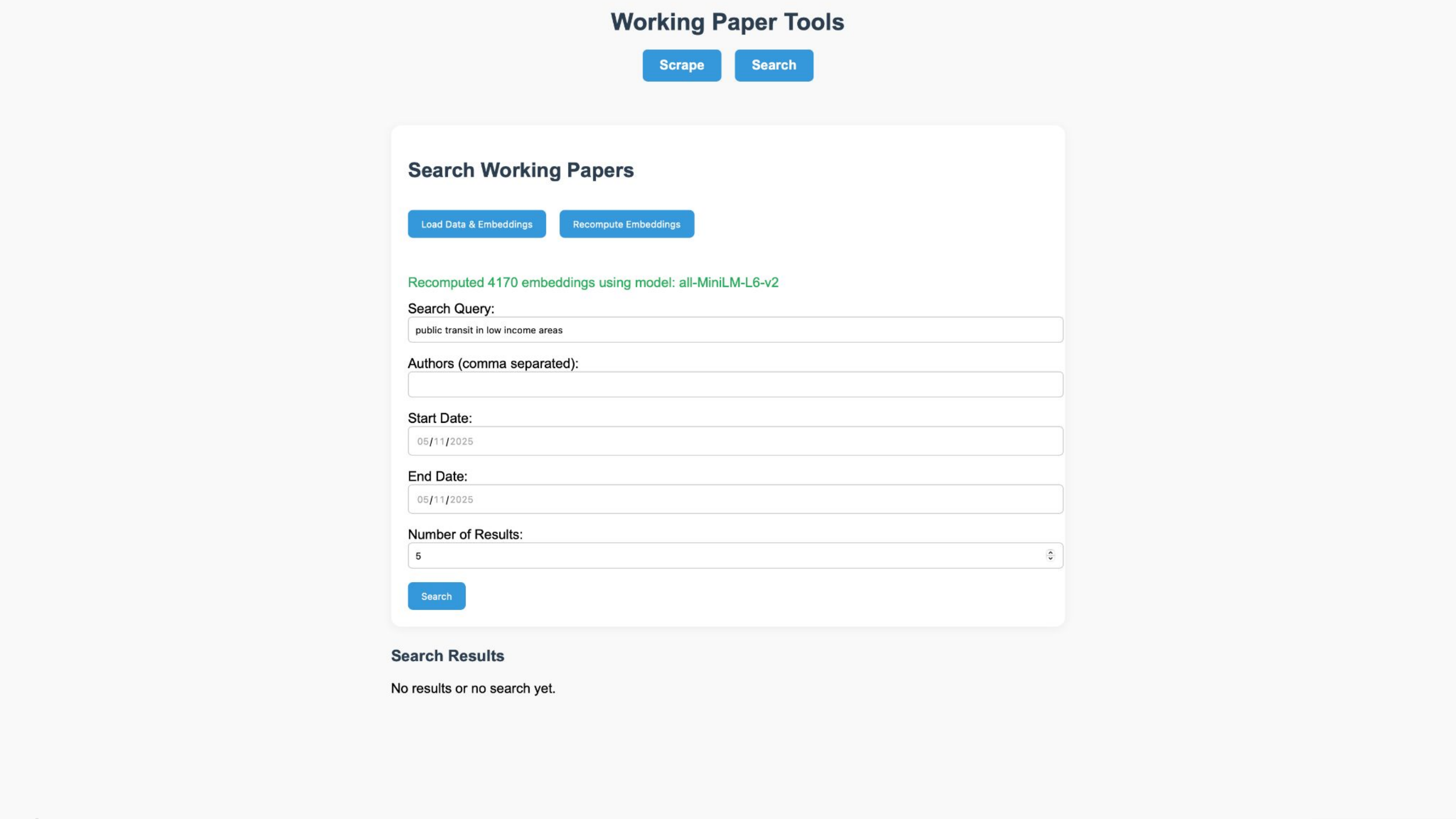
Start Date: 05/11/2025

End Date: 05/11/2025

Number of Results: 5

Search Results

No results or no search yet.



Working Paper Tools

Search Working Papers

Recomputed 4170 embeddings using model: all-MiniLM-L6-v2

Search Query: public transit in low income areas

Authors (comma separated):

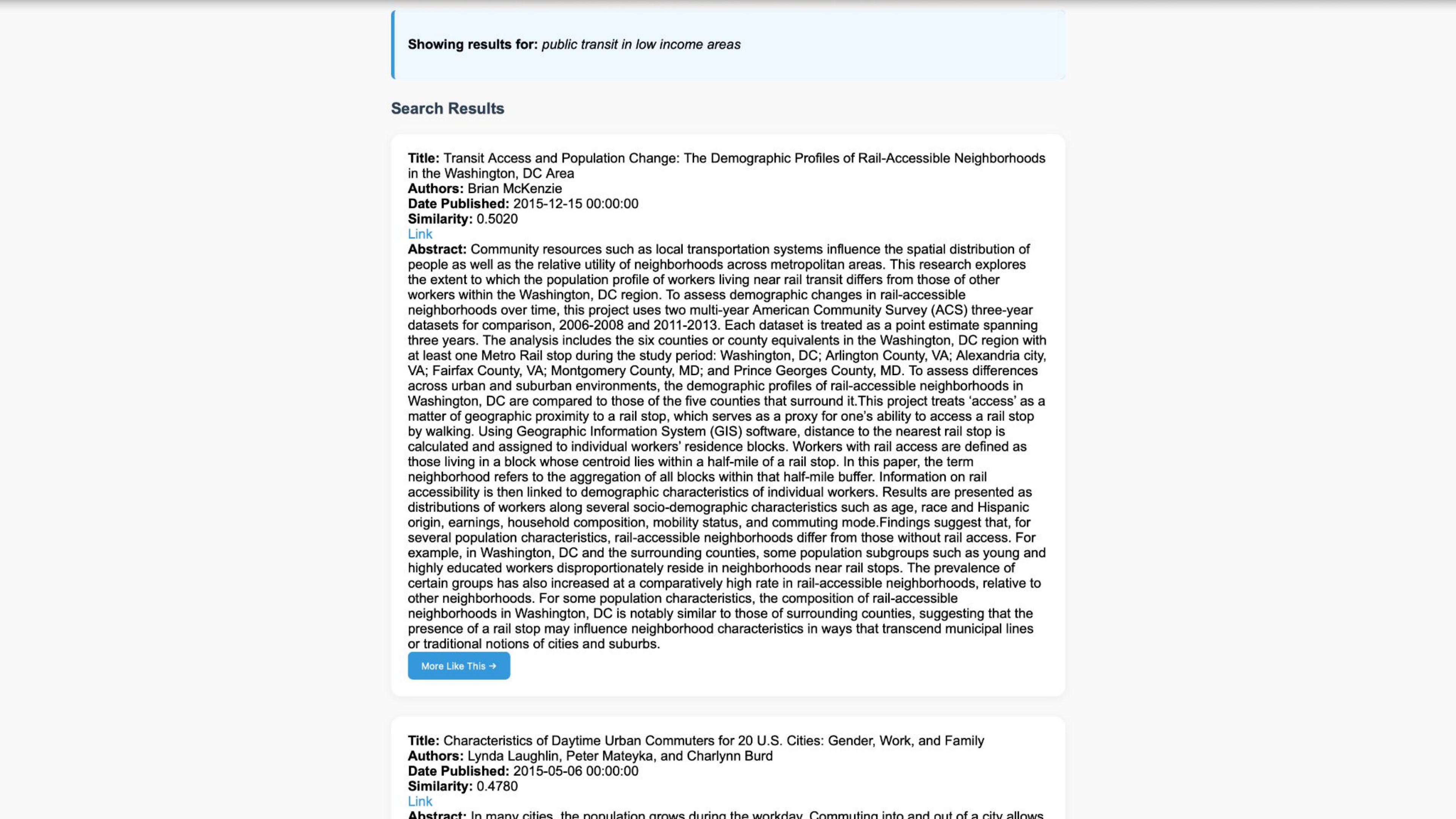
Start Date: 05/11/2025

End Date: 05/11/2025

Number of Results: 5

Search Results

No results or no search yet.

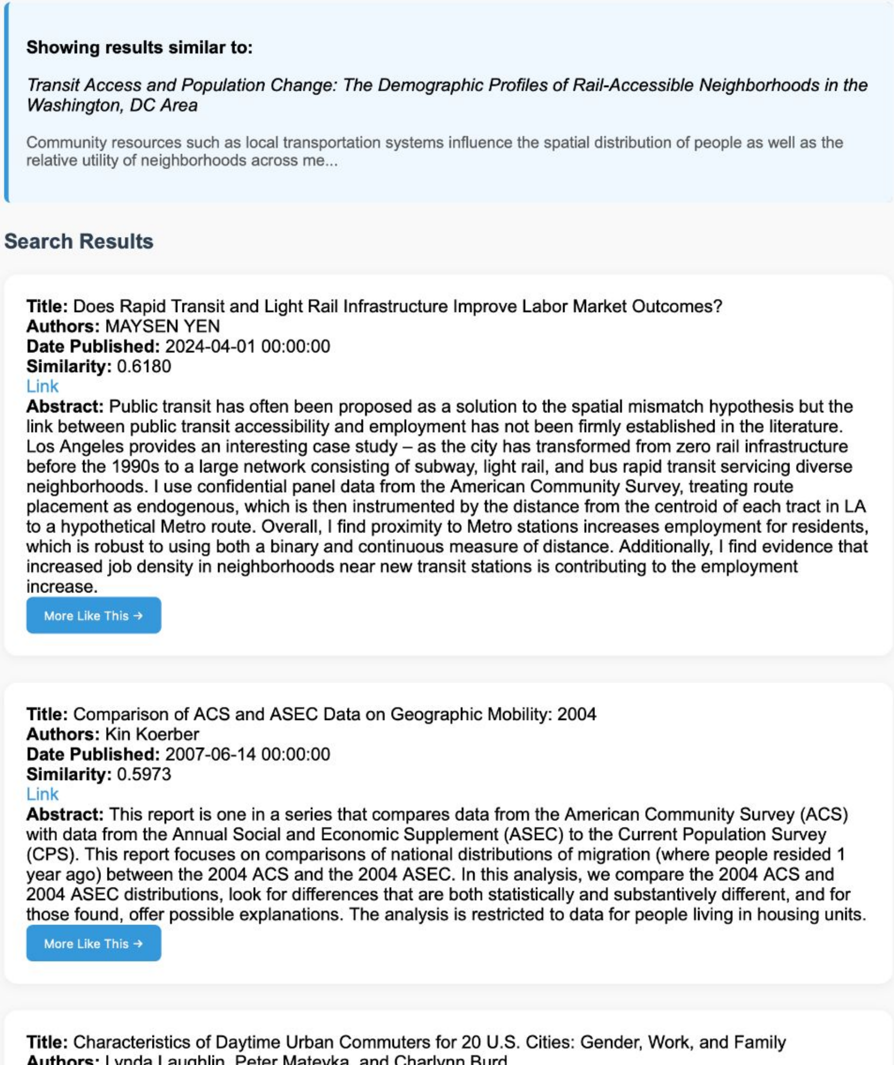


Showing results for: *public transit in low income areas*

Search Results

Title: Transit Access and Population Change: The Demographic Profiles of Rail-Accessible Neighborhoods in the Washington, DC Area
Authors: Brian McKenzie
Date Published: 2015-12-15 00:00:00
Similarity: 0.5020
Link
Abstract: Community resources such as local transportation systems influence the spatial distribution of people as well as the relative utility of neighborhoods across metropolitan areas. This research explores the extent to which the population profile of workers living near rail transit differs from those of other workers within the Washington, DC region. To assess demographic changes in rail-accessible neighborhoods over time, this project uses two multi-year American Community Survey (ACS) three-year datasets for comparison, 2006-2008 and 2011-2013. Each dataset is treated as a point estimate spanning three years. The analysis includes the six counties or county equivalents in the Washington, DC region with at least one Metro Rail stop during the study period: Washington, DC; Arlington County, VA; Alexandria city, VA; Fairfax County, VA; Montgomery County, MD; and Prince Georges County, MD. To assess differences across urban and suburban environments, the demographic profiles of rail-accessible neighborhoods in Washington, DC are compared to those of the five counties that surround it.This project treats ‘access’ as a matter of geographic proximity to a rail stop, which serves as a proxy for one’s ability to access a rail stop by walking. Using Geographic Information System (GIS) software, distance to the nearest rail stop is calculated and assigned to individual workers’ residence blocks. Workers with rail access are defined as those living in a block whose centroid lies within a half-mile of a rail stop. In this paper, the term neighborhood refers to the aggregation of all blocks within that half-mile buffer. Information on rail accessibility is then linked to demographic characteristics of individual workers. Results are presented as distributions of workers along several socio-demographic characteristics such as age, race and Hispanic origin, earnings, household composition, mobility status, and commuting mode.Findings suggest that, for several population characteristics, rail-accessible neighborhoods differ from those without rail access. For example, in Washington, DC and the surrounding counties, some population subgroups such as young and highly educated workers disproportionately reside in neighborhoods near rail stops. The prevalence of certain groups has also increased at a comparatively high rate in rail-accessible neighborhoods, relative to other neighborhoods. For some population characteristics, the composition of rail-accessible neighborhoods in Washington, DC is notably similar to those of surrounding counties, suggesting that the presence of a rail stop may influence neighborhood characteristics in ways that transcend municipal lines or traditional notions of cities and suburbs.

Title: Characteristics of Daytime Urban Commuters for 20 U.S. Cities: Gender, Work, and Family
Authors: Lynda Laughlin, Peter Mateyka, and Charlynn Burd
Date Published: 2015-05-06 00:00:00
Similarity: 0.4780
Link
Abstract: In many cities, the population grows during the workday. Commuting into and out of a city allows



Showing results similar to:

Transit Access and Population Change: The Demographic Profiles of Rail-Accessible Neighborhoods in the Washington, DC Area

Community resources such as local transportation systems influence the spatial distribution of people as well as the relative utility of neighborhoods across me...

Search Results

Title: Does Rapid Transit and Light Rail Infrastructure Improve Labor Market Outcomes?
Authors: MAYSEN YEN
Date Published: 2024-04-01 00:00:00
Similarity: 0.6180
Link
Abstract: Public transit has often been proposed as a solution to the spatial mismatch hypothesis but the link between public transit accessibility and employment has not been firmly established in the literature. Los Angeles provides an interesting case study – as the city has transformed from zero rail infrastructure before the 1990s to a large network consisting of subway, light rail, and bus rapid transit servicing diverse neighborhoods. I use confidential panel data from the American Community Survey, treating route placement as endogenous, which is then instrumented by the distance from the centroid of each tract in LA to a hypothetical Metro route. Overall, I find proximity to Metro stations increases employment for residents, which is robust to using both a binary and continuous measure of distance. Additionally, I find evidence that increased job density in neighborhoods near new transit stations is contributing to the employment increase.

Title: Comparison of ACS and ASEC Data on Geographic Mobility: 2004
Authors: Kin Koerber
Date Published: 2007-06-14 00:00:00
Similarity: 0.5973
Link
Abstract: This report is one in a series that compares data from the American Community Survey (ACS) with data from the Annual Social and Economic Supplement (ASEC) to the Current Population Survey (CPS). This report focuses on comparisons of national distributions of migration (where people resided 1 year ago) between the 2004 ACS and the 2004 ASEC. In this analysis, we compare the 2004 ACS and 2004 ASEC distributions, look for differences that are both statistically and substantively different, and for those found, offer possible explanations. The analysis is restricted to data for people living in housing units.

Title: Characteristics of Daytime Urban Commuters for 20 U.S. Cities: Gender, Work, and Family
Authors: Lynda Laughlin, Peter Mateyka, and Charlynn Burd

Conclusions & Interpretation

Interpretation of Results

Takeaways

- SBERT is a suitable engine for semantic search on the Census Working Papers
- all-MiniLM-L6-v2 is a good starting point as a model
- For embedding based models embedding time varies heavily, but inference time will almost always be fast enough

Limitations and Potential Next Steps

Limitations

- Data scraping script takes hours
- Added computational overhead
  - Maintaining FAISS index
- Only searches on abstracts
- Potential abstract length bias
- Some older papers don’t have all the data needed

Potential Next Steps

- Switch to a vector store more aligned with what's in use currently
- Combine with in-document analytics

Conclusions

- Improved document retrieval by integrating machine learning and NLP techniques
- Compared to TF-IDF, Word2Vec and Doc2Vec, SBERT gives the best searches according to our testing
- Made a Flask based GUI that allows for searching and similar article recommendations
- Next steps: Adapt to current Census website infrastructure, combine with in-document analytics

Bibliographical References

Gahman, Nicholas, and Vinayak Elangovan. "A Comparison of Document Similarity Algorithms." *arXiv preprint arXiv:2304.01330* (2023).

Liu, Xiaodong, Yu Wang, Jianshu Ji, Hao Cheng, Xueyun Zhu, Emmanuel Awa, Pengcheng He et al. "The microsoft toolkit of multi-task deep neural networks for natural language understanding." *arXiv preprint arXiv:2002.07972* (2020).

Reimers, Nils, and Iryna Gurevych. "Sentence-bert: Sentence embeddings using siamese bert-networks." *arXiv preprint arXiv:1908.10084* (2019).

Kang, Y., Cai, Z., Tan, C. W., Huang, Q., & Liu, H. (2020). Natural language processing (NLP) in management research: A literature review. Journal of Management Analytics, 7(2), 139–172. https://doi.org/10.1080/23270012.2020.1756939