Contents

Course and Team Information

- **Course:** CIT5900-002
- **Semester:** Spring 2025
- **Instructor:** Kaihua Ding
- **Group 7 Team Members:**
  - Yuqiao Xue (joexue@seas.upenn.edu)
  - Ziad Albitty (zalbitty@seas.upenn.edu)
  - Jichang Wen (wenjc@seas.upenn.edu)
  - Chenxu Li (chenxuli@seas.upenn.edu)
  - Nii Armah (na554@seas.upenn.edu)

1 Project Overview

This project represents the final stage of our FSRDC research output pipeline, building on data gathered and cleaned in earlier phases. In Project 3, we integrated deliverables from all student groups into a single validated dataset, applied advanced data processing techniques, and conducted a range of exploratory analyses.

Our goals were multifaceted: to consolidate and enrich the research outputs, uncover latent themes, group similar documents, analyze sentiment tones, visualize clustering patterns, and investigate keyword trends over time. These tasks provided a comprehensive view of the FSRDC publication landscape, revealing how researchers interact with restricted-use data and how academic interests evolve across time and topics.

Through a combination of API-driven enrichment, unsupervised learning (clustering, topic modeling), and data visualization, we aimed to produce insights that are both technically rigorous and thematically meaningful. The outcomes were presented via an interactive GitHub Pages site to ensure accessibility and reproducibility for stakeholders.

2 Input Processing

The goal of thai part is to construct a robust and scalable data pipeline that consolidates, cleans, and enriches research output data submitted by eight student groups. These groups independently analyzed FSRDC-related publications based on questions related to prolific authors, institutions, and topics. Because their methodologies and source data varied, the submitted files contained inconsistent formats, overlapping entries, and incomplete metadata. This pipeline ensures all teams start subsequent analysis from a unified, validated dataset.

2.1 Objective

The pipeline's objective is to ingest and standardize disparate research output files, remove duplicates, validate FSRDC relevance, and enrich missing fields using both internal and external metadata sources. The final output must conform strictly to the standard column schema provided in ResearchOutputs.xlsx, with each field clearly defined and reliably populated. Additionally, the pipeline applies real-world data science practices, such as fuzzy matching, TF-IDF-based similarity scoring, API querying, and conservative data merging strategies.

2.2 APIs Used

Two primary APIs were employed to enrich incomplete entries:

- **CrossRef API**: Used to retrieve bibliographic metadata including publication venue, citation string, document type, volume, number, and pagination.
- **OpenAlex API**: Served as a fallback API, providing similar metadata. It was especially useful for preprint repositories and working papers not indexed in CrossRef.

Both APIs were queried using normalized OutputTitle fields, and results were used only to fill missing values in select columns, ensuring that existing correct data was not overwritten.

2.3 Methodology

The data processing pipeline was structured into modular, testable steps:

2.3.1 Cleaning and Normalization

Each group file (`group1.csv` to `group8.csv`) was loaded and processed to normalize column headers and values. A column-mapping dictionary translated inconsistent field names to a unified schema. Text fields were stripped of punctuation, HTML tags, and redundant whitespace using BeautifulSoup and regex-based normalization.

2.3.2 Consolidation

All cleaned files were merged into a single CSV, and each row was tagged with its source filename for traceability. This created a master dataset containing over 39,000 rows.

2.3.3 Deduplication

Duplicate records were removed using a two-tier approach: first by `OutputTitle`, then by `DOI`. Rows missing both `DOI` and `URL` were discarded.

2.3.4 Metadata Enrichment

Using `ProjectsAllMetadata.xlsx`, missing project-level metadata such as `ProjID`, `ProjectStatus`, and `ProjectPI` was filled using unambiguous PI matches.

2.3.5 Validation with Fuzzy Matching and Abstract Similarity

- Fuzzy matching on `ProjectPI` (threshold ≥ 90)
- TF-IDF cosine similarity on Abstracts (threshold ≥ 0.6)

Entries that satisfied either criterion were retained. This step helped eliminate false positives while preserving relevant records that may have minor textual variations.

2.3.6 DOI and Year Fixes

Numerical fields like ProjID and OutputYear were cleaned to remove extraneous characters or formatting issues (e.g., trailing zeros or embedded spaces). DOIs were also reconstructed from malformed strings, restoring them to a valid URL format.

2.3.7 External Metadata Enrichment via API

For entries missing key bibliographic fields, the pipeline queried CrossRef and OpenAlex APIs. Matching results were used to fill in only those fields that were empty in the original data, such as OutputBiblio, OutputType, OutputVenue, OutputYear, OutputPages, and OutputStatus. To prevent rate-limiting, queries were throttled with a delay between requests.

2.3.8 Final Deduplication and Merge

The enriched dataset was merged with the previously curated `ResearchOutputs.xlsx`. Normalized OutputTitle fields were used to identify and remove duplicates. Preference was given to entries enriched via pipeline over those in the static Excel sheet. Remaining fallback data from secondary columns (e.g., ProjectStartYear) was mapped to primary fields, ensuring completeness.

2.4 Seed Input

- `group1.csv` to `group8.csv`: Raw input from eight student groups.
- `ProjectsAllMetadata.xlsx`: Official metadata file containing project-level information.
- `ResearchOutputs.xlsx`:Reference dataset for final merge and validation.

2.5 Pipeline Output

The final deliverable is `FinalCleanedMergedResearchOutputs.csv`, or `ResearchOutputs_Group7.csv` — a fully deduplicated, validated, and enriched dataset. It contains the following columns:

Table 1: Final Output Columns

Column Name
ProjID
ProjectStatus
ProjectTitle
ProjectRDC
ProjectYearStarted
ProjectYearEnded
ProjectPI
OutputTitle
OutputBiblio
OutputType
OutputStatus
OutputVenue
OutputYear
OutputMonth
OutputVolume
OutputNumber
OutputPages

3 EDA Analysis

3.1 Top 10 RDCs by Research Output Count

Table 2: Top 10 RDCs by Number of Outputs

Rank	**RDC**	**# Outputs**
1	Washington	89
2	Boston	83
3	Triangle	76
4	Chicago	69
5	Michigan	60
6	Berkeley	47
7	Baruch	45
8	Cornell	38
9	Atlanta	32
10	Penn State	28

The Washington RDC leads the list, followed by Boston and Triangle. Major Census RDCs also feature in the top tier reflecting long standing program and research activity. Smaller centers (e.g., Atlanta, Penn State) produce fewer outputs but still make the top 10.

3.2 Publication per Year

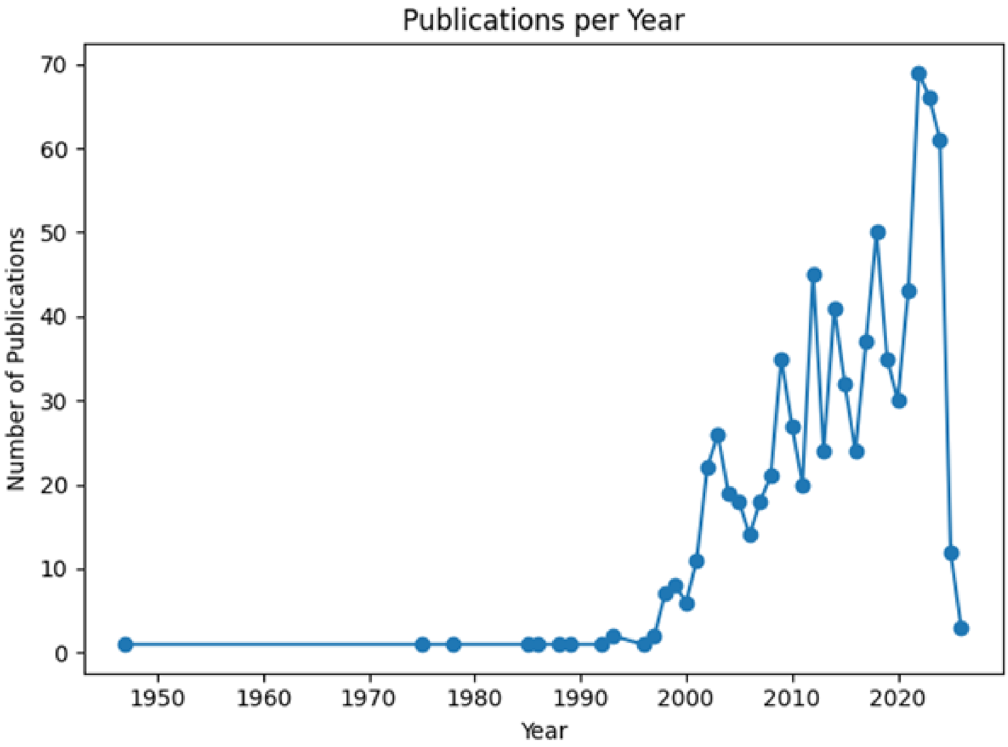

Figure 1: Number of FSRDC Publications Per Year

From 1945 through the early 1990s, very few FSRDC-based outputs appeared. Starting around 2000, output volume climbed sharply—reaching roughly 35–50 publications annually by the late 2000s. The peak years (2020–2022) saw 60–70 outputs.

3.3 Top 10 Most Prolific Authors

Table 3: Top 10 Most Prolific Authors

Rank	Author	# Publications
1	Javier Miranda	14
1	Scott H. Nolan	14
3	Jérôme P. Reiter	11
4	Eric P. Baumer	10
4	Qingfang Wang	10
6	Jason Fletcher	8
6	Emin M. Dinlersoz	8
8	John R. Logan	7
8	Amy B. O'Hara	7
8	Kyle E. Walker	7

Two leading scholars—Javier Miranda and Scott Nolan—each have 14 FSRDC-driven papers. A few researchers account for a disproportionately large share of outputs; most authors appear only once or twice.

3.4 Citation Count Distribution

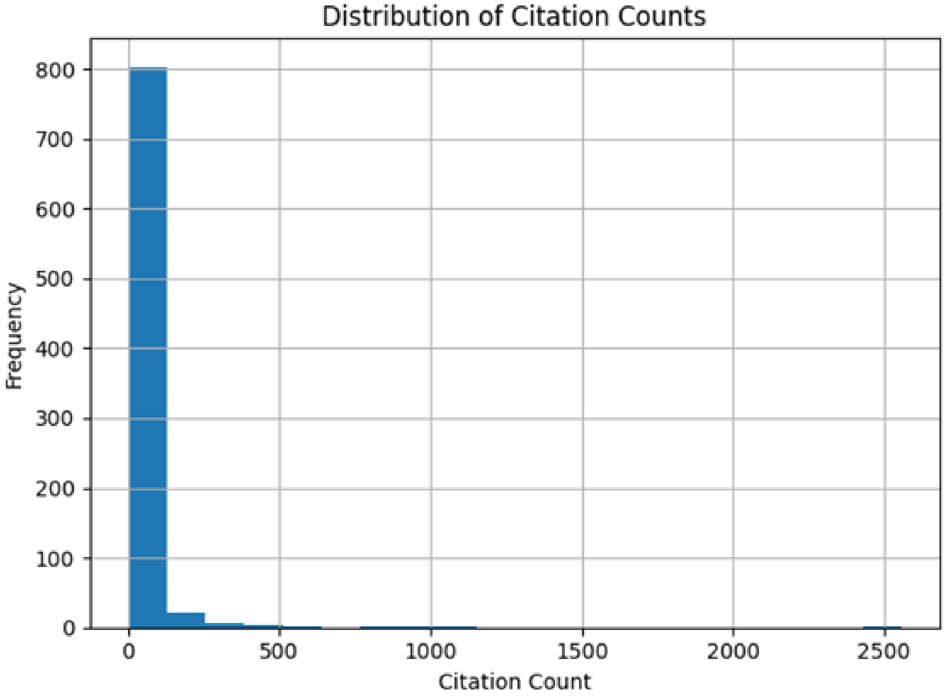

Figure 2: Citation Count Distribution of FSRDC Outputs

While most FSRDC research has modest citation impact, a handful of "landmark" studies drive the high end of the distribution. The heavy right-skew suggests citation filters (e.g., median ¿ 20) could help isolate the most influential outputs.

3.5 Additional Creative Insight: Time-to-Publication Lag

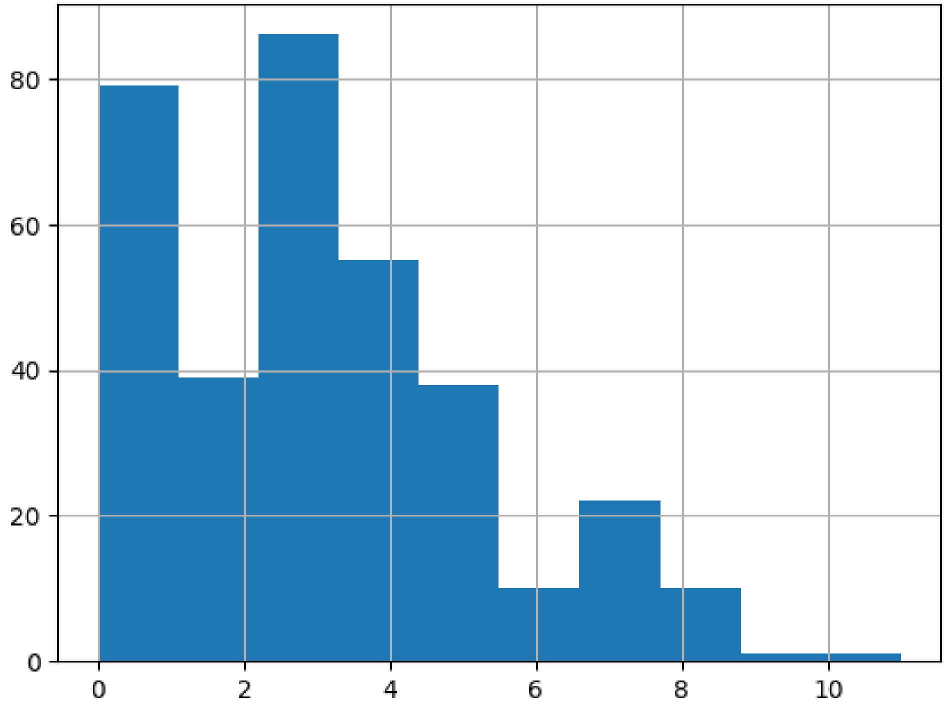

Figure 3: Time-to-Publication Lag Distribution

We computed `lag_years = OutputYear { ProjectYearStarted` for each record. Median lag: 3 years, with most papers published within 1–5 years of project start. A small number of projects take 7+ years to yield outputs, perhaps reflecting longitudinal or complex data work.

4 Using Python for Data Science Applications

4.1 Classification Models and PCA Analysis

We decided to create a classification model that would predict whether an output was published (PB), unpublished (UP), or forthcoming (FC). Initial analysis of the merged results showed that there were a lot of columns with null data, as shown in Table 4.

Table 4: Null Value Summary by Column

Column Name	**null_count**	**null_percent (%)**
ProjID	82	3.86
ProjectStatus	82	3.86
ProjectTitle	281	13.24
ProjectRDC	1	0.05
ProjectYearStarted	281	13.24
ProjectYearEnded	813	38.29
ProjectPI	1	0.05
OutputTitle	1	0.05
OutputBiblio	843	39.71
OutputType	0	0.00
OutputStatus	1	0.05
OutputVenue	301	14.18
OutputYear	9	0.42
OutputMonth	1208	56.90
OutputVolume	1279	60.24
OutputNumber	1144	53.89
OutputPages	1275	60.06

For this reason, only columns that had less than 5% of data missing were used in the model. There was 1 result in our target column (`OutputStatus`) with missing information. That row was removed.

The remaining information in the feature columns was imputed with `sklearn`'s `SimpleImputer`. The imputer used median for numerical columns (such as `OutputYear`) and 'missing' for categorical columns (`ProjectStatus`, `ProjectRDC`, `OutputType`). The `OutputTitle` column was also imputed, and TF-IDF scores were calculated.

Columns with less than 5% missing were not used in training since they were deemed unhelpful (`ProjID`, `ProjectPI`). Since FC outputs were very low compared to the other two, weights of 50:1:1 were applied to FC:PB:UP.

The data was used to train an SVC model with an RBF kernel. The parameters for `C` and `gamma` were tuned using grid search cross-validation to optimize the mean ROC-AUC one-vs-one score. The best parameters were then used to train the full dataset, and the resulting confusion matrix was obtained.

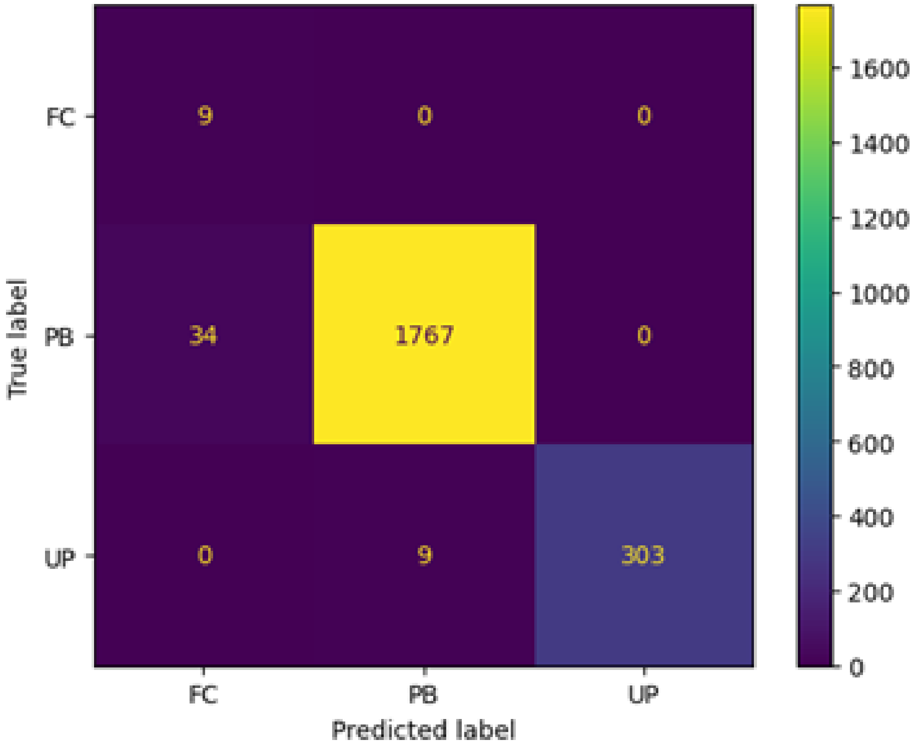

Figure 4: Confusion Matrix of the Classification Model

This showed that our model was able to distinguish between unpublished and published data well, but struggled with forthcoming data. This is likely because our training set did not evenly split the 3 categories, and even with the weights the model still struggled.

PCA analysis was performed using these columns: `OutputYear`, `ProjectStatus`, and `ProjectRDC`. However, when One Hot Encoding was used, it was found that there were too many categories and so it was difficult for PCA to reduce it to a small number of principal components with meaningful insights, as seen in the Scree Plot below:

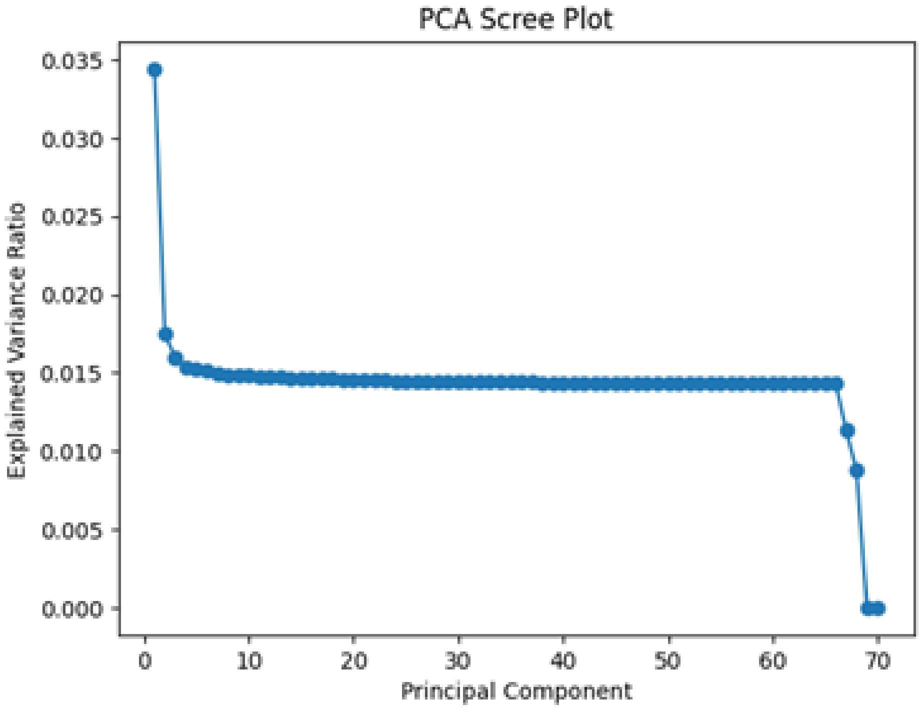

Figure 5: PCA Scree Plot of Encoded Project Metadata

The top 3 drivers for the first 2 principal components were:

Table 5: Top 3 Feature Loadings for PC1

Feature	**Loading**
ProjectStatus_Active	0.621601
ProjectStatus_Completed	0.621601
ProjectRDC_washington	0.121841

Table 6: Top 3 Feature Loadings for PC2

Feature	**Loading**
OutputYear	0.686427
ProjectRDC_nebraska	0.389875
ProjectRDC_baruch	0.303420

The variances were very low with these drivers outside of the Project Status. This shows there were too many categories to gain any true insights.

Page count and Project Time were calculated for rows that had that information. All rows with at least one column of missing information were dropped. PCA analysis was done on these columns: `ProjectYearStarted`, `ProjectYearEnded`, `OutputYear`, `PageCount`, and `ProjectDuration`. The Scree plot showed that the data could be explained with 2–3 components.

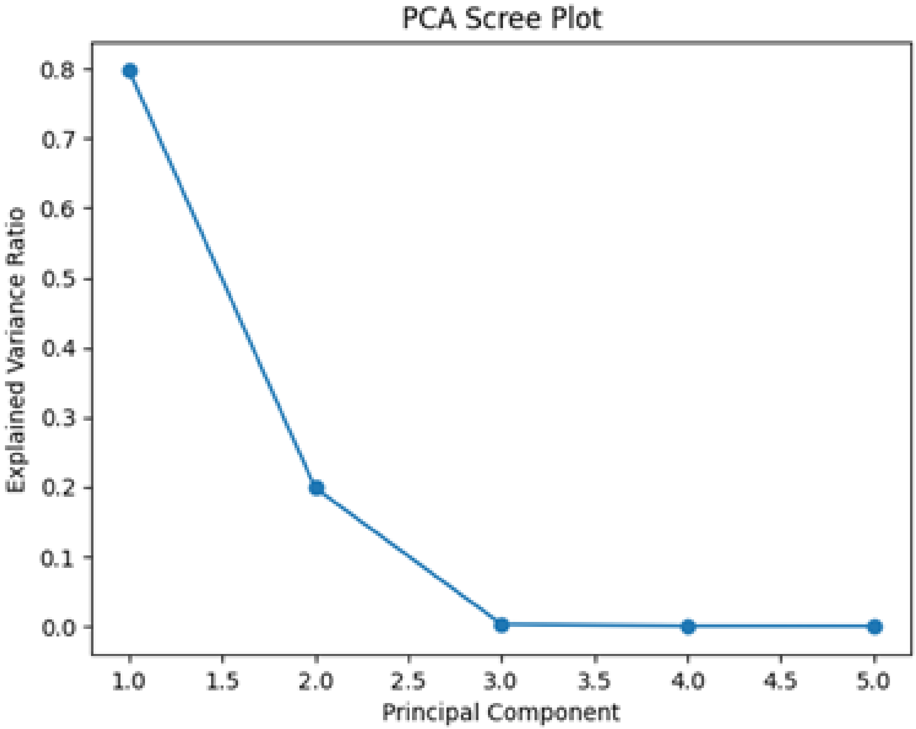

Figure 6: Scree Plot for Page Count and Project Time Features

The top contributors for each component can be found below:

Table 7: Top Drivers of PC1

Feature	**Loading**
OutputYear	0.999985
ProjectYearEnded	0.004231
ProjectYearStarted	0.002831

Table 8: Top Drivers of PC2

Feature	**Loading**
PageCount	0.999985
ProjectYearStarted	-0.003731
ProjectYearEnded	-0.003567

Table 9: Top Drivers of PC3

Feature	**Loading**
ProjectYearEnded	0.763383
ProjectYearStarted	0.632506
ProjectDuration	0.130877

4.2 Text Processing and Clustering Techniques

4.2.1 PCA Scatter Plot of Document Embeddings (K-Means Clusters)

Overview

In this analysis, we clustered the research abstracts to uncover natural groupings of documents based on their content. We used K-Means, a partitional clustering algorithm, on high-dimensional text embeddings of the abstracts and visualized the clusters in two dimensions using PCA. Each point in Figure 7 represents a research output, positioned according to its content similarity to others and colored by its K-Means cluster membership.

Objectives

- Identify distinct clusters of documents to see if the corpus breaks down into clear thematic groups.
- Use visualization to assess how well-separated these clusters are and to intuitively understand the relationship between clusters (e.g., which clusters are close together or overlapping in content space).

Methodology

- **Text Vectorization:** Transformed each document's abstract into a numerical vector using TF–IDF features. This represents each document in a high-dimensional term space.
- **Dimensionality Reduction:** Applied Principal Component Analysis (PCA) to reduce the vector space to 2 principal components for visualization. The PCA projection retains the most significant variance in the data while allowing 2D plotting.
- **Clustering (K-Means):** Performed K-Means clustering on the document vectors (in the original high-dimensional space). The number of clusters K was determined using silhouette analysis by testing various K values and choosing the one with the highest silhouette score (optimal separation at $K = 10$ clusters).
- **Visualization:** Plotted the documents on the 2D PCA scatter plot, coloring each point according to its assigned cluster. The cluster label of each document (from 1 to 10) is indicated by color so that clusters form colored groupings in the plot.

Results

The scatter plot (Figure 7) reveals the presence of ten document clusters in the abstract corpus. We observe that several clusters form tight groupings, whereas others overlap in the central region of the plot. For example, a few clusters (e.g., the orange and cyan points in Figure 7) are somewhat separated toward the periphery of the plot, indicating those documents form a distinct thematic group apart from the rest. In contrast, the majority of clusters (points colored red, green, blue, etc.) occupy an overlapping region near the center, suggesting those topics share similarities or gradual transitions with one another.

The overall cluster separation is moderate – the optimal K-Means solution of 10 clusters does partition the documents into meaningful groups, but not all clusters are entirely isolated. This implies the research outputs cover a range of topics that sometimes blend into each other.

The clustering provides a high-level thematic map of the corpus: documents within each color group are more similar in content to each other than to those in other groups. This visual confirmation of cluster cohesion (and areas of overlap) helps validate the choice of K and hints at underlying thematic structure in the research abstracts.

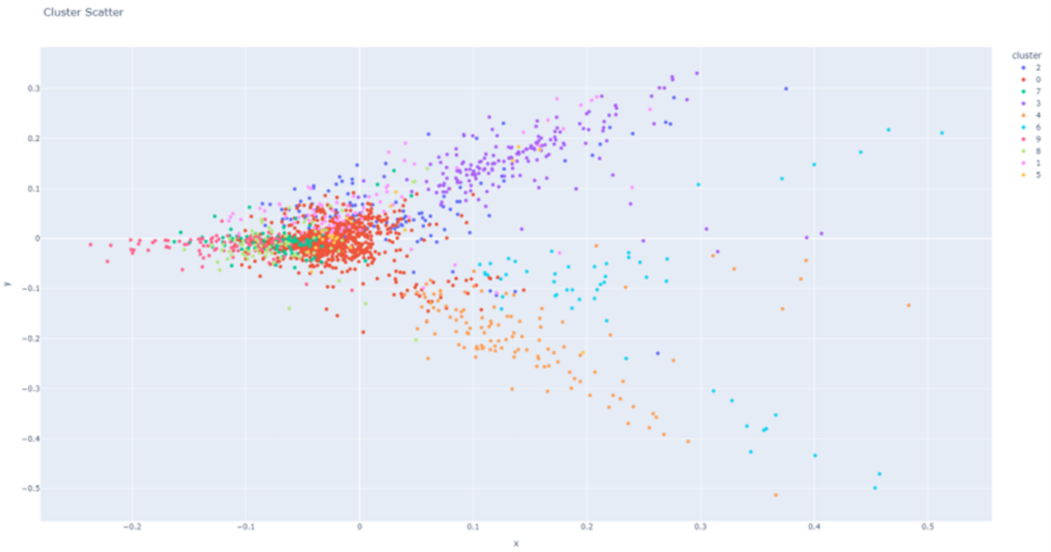

Figure 7: PCA Scatter Plot of Abstract Embeddings (K-Means Clusters)

4.2.2 PCA Scatter Plot of Document Embeddings (Agglomerative Clusters)

Overview

To compare with K-Means, we applied hierarchical clustering to the same document embeddings and visualized the results with PCA. Agglomerative clustering groups documents based on a hierarchy of merges, potentially yielding different cluster shapes or sizes. Figure 8 is a scatter plot of the documents (in the same PCA 2D space as Figure 7), but now colored by their agglomerative cluster assignments rather than K-Means clusters.

Objectives

- Evaluate whether a hierarchical clustering approach yields a different partition of documents compared to K-Means.
- Observe the consistency or differences in cluster composition and separation when using Ward's method, providing insight into the robustness of the identified thematic groups.

Methodology

- **Clustering (Agglomerative):** Performed Agglomerative Hierarchical Clustering on the document TF–IDF vectors, using Ward's minimum variance criterion to merge clusters. We set the number of clusters to $k = 10$ to allow direct comparison with the K-Means result.
- **Visualization:** Plotted the same PCA-reduced document points as in Figure 7, but this time colored each point according to its hierarchical cluster label. The cluster count was kept at 10 for consistency, and colors were assigned uniquely to each of the ten Ward clusters.

Results

The resulting scatter plot (Figure 8) shows that the hierarchical clustering yields a partition of the documents largely comparable to the K-Means solution, with some differences in cluster boundaries.

Many points that were grouped together in Figure 7 remain in the same colored group in Figure 8, indicating that there is a core thematic structure the two methods agree on. For instance, documents that formed a distinct outlying cluster in the K-Means plot (e.g., those on the far right of the PCA space) are still grouped together under the hierarchical scheme (similarly colored on the far right in Figure 8). This suggests a stable cluster of outlier documents with a unique theme.

However, there are also differences: hierarchical clustering, due to its binary merge process, produces one or two clusters that are more imbalanced in size. In Figure 8, one cluster (colored red) appears larger and more diffuse than in the K-Means result, whereas another cluster is very small (a few purple points tightly grouped). This indicates that Ward's method merged some groups that K-Means kept separate, and conversely isolated a tiny cluster that K-Means had merged into a larger group.

Overall, the clusters in the agglomerative approach show a similar overlapping pattern in the central region (indicating again that several themes are closely related), and only minor reassignments of documents to different clusters compared to K-Means. The comparison suggests that the ten-group structure is relatively robust, as major divisions in content are captured by both methods, with hierarchical clustering providing a slightly different perspective on cluster composition due to its tendency to create clusters of varying sizes.

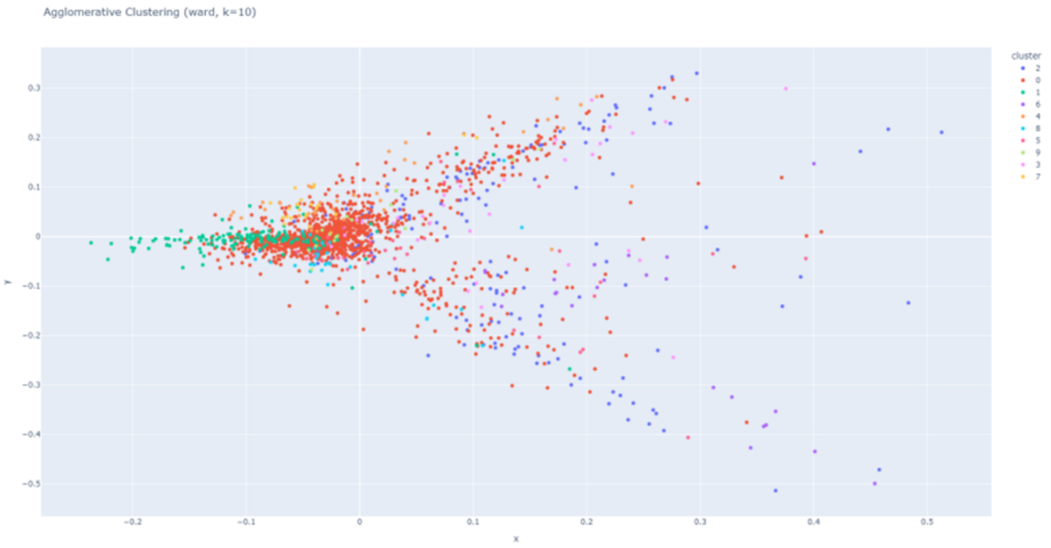

Figure 8: PCA Scatter Plot of Abstract Embeddings (Agglomerative Clusters)

4.2.3 Top Keywords of LDA Topics

Overview

In addition to clustering, we applied topic modeling to the abstracts to extract latent themes in an unsupervised manner. Using Latent Dirichlet Allocation (LDA), we decomposed the corpus into five topics and identified the most significant keywords for each topic. Figure 9 presents a bar chart of the top terms associated with each of the five LDA topics, with each topic distinguished by a different color. This visualization helps interpret the topics by highlighting the keywords that have the highest importance (i.e., the highest probability within the topic).

Objectives

- Reduce the complexity of thousands of documents into a small number of interpretable "topics," each represented by prominent keywords.
- Understand the thematic content of each derived topic by examining its top keywords, thereby validating whether the topics correspond to meaningful research themes in the FSRDC output corpus.

Methodology

- **Topic Modeling (LDA):** Performed LDA on the corpus of document texts, specifying 5 topics. The number of topics was chosen based on interpretability and model fit (e.g., topic coherence scores), and found to adequately cover the diversity of the corpus. Each document is modeled as a mixture of topics, and each topic as a distribution over words.

- **Extract Top Keywords:** For each of the 5 topics, we extracted the highest-weighted terms (keywords). We focused on the top 10 keywords per topic, which give a clear sense of the topic's subject matter. Common stopwords were removed prior to modeling.
- **Visualization:** A bar chart (Figure 9) compiles the top terms for all five topics. Each term on the x-axis is color-coded by topic (Topic 1 through Topic 5), and the y-axis indicates the relevance or weight of the term within its topic. This side-by-side comparison enables quick interpretation of each theme.

Results

The bar chart in Figure 9 highlights distinct keyword sets for each of the five LDA topics, confirming that the model discovered coherent themes within the research abstracts.

- **Topic 1** (blue bars): Terms such as "manufacturing," "energy," and "capital" suggest a theme related to industry and economic production.
- **Topic 2** (red bars): Includes "market," "firm," "trade," "growth," and "impact," pointing to markets, business economics, and possibly international trade.
- **Topic 3** (green bars): Includes "firm," "business," "dynamic," "cycle," "neighborhood," and "industrial," indicating productivity, collaboration, and economic structure.
- **Topic 4** (purple bars): Terms like "health," "income," "child," "rural," and "learning" suggest a socio-economic and demographic focus.
- **Topic 5** (orange bars): With "data," "census," "survey," "state," "community," "United," and "American," this topic revolves around census and survey-based demographic research.

These keyword sets have minimal overlap and exhibit strong thematic coherence. The LDA results complement the clustering analysis by offering an interpretable summary of topics through key terms, reinforcing the diversity and structure of the FSRDC research landscape.

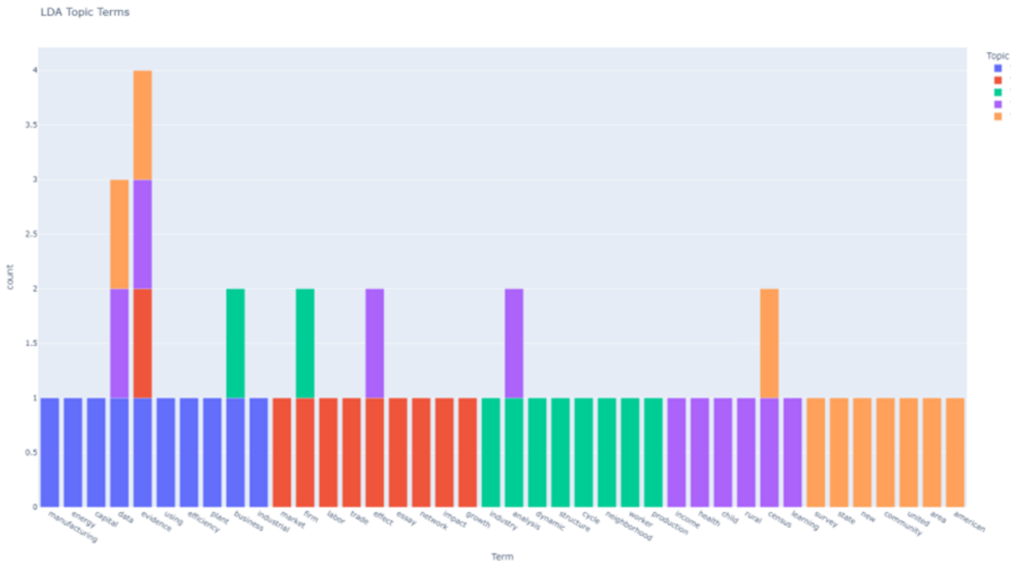

Figure 9: Top Keywords of LDA Topics

4.2.4 Sentiment Polarity Distribution of Documents

Overview

We conducted a sentiment analysis of all the research abstracts to gauge the emotional or subjective tone of the language used in these scholarly documents. Although academic abstracts are typically written in neutral, objective language, analyzing sentiment polarity can reveal subtle variations — for instance, whether abstracts lean slightly positive (e.g., highlighting benefits or improvements) or contain negative wording (perhaps when describing problems or challenges). Figure 10 shows a histogram of sentiment polarity scores for all documents, illustrating the distribution of sentiment from negative to positive.

Objectives

- Quantitatively measure the sentiment of each document's text to see the overall tone of the corpus and identify any outliers (documents that are unusually positive or negative).
- Verify the expectation that research abstracts are mostly neutral in tone, while checking if any documents deviate significantly, which could be worth further qualitative examination.

Methodology

- **Sentiment Scoring:** Employed a sentiment analysis tool (a lexicon and rule-based sentiment analyzer) to calculate a polarity score for each abstract. Each score ranges from −1 (very negative sentiment) through 0 (neutral) to +1 (very positive sentiment). The analyzer likely assigns scores based on the presence of positive or negative words/phrases in the text (adjusted for intensity and context).

- **Aggregation:** Collected the sentiment score for all documents in the corpus. Given the formal nature of abstracts, we anticipated many scores to cluster near 0, with fewer documents having extreme values.
- **Visualization:** Plotted a histogram of the polarity scores. On the x-axis are binned sentiment scores (from negative to positive), and the y-axis shows the count of documents falling into each sentiment bin. We chose a bin size fine enough to capture slight variations around neutrality.

Results

The sentiment score distribution in Figure 10 is highly skewed toward neutrality. The histogram displays a towering central bar at or near a score of 0, indicating that an overwhelming number of abstracts have a neutral sentiment (scores very close to 0). This confirms that the language in the research outputs is predominantly objective and formal, as expected in academic writing.

To the right of zero, we observe a secondary cluster of documents with mildly positive sentiment scores. There is a noticeable bump in the range of roughly +0.2 to +0.5, suggesting that a subset of abstracts use slightly positive or optimistic language — for example, using terms like “improved,” “significant,” or “benefit,” which can nudge the sentiment score upward. The peak of this positive cluster is still relatively modest (no abstract approaches a full +1 score), indicating that even the most positive abstracts are only moderately positive in tone.

On the negative side of the spectrum, the histogram shows very few documents with negative polarity scores, and those scores are only mildly negative (around –0.3 to –0.6 at most). These might be instances where abstracts describe limitations, challenges, or negative outcomes (using words like “bias,” “error,” or “cost”) that slightly tip the sentiment to the negative side. Importantly, there are virtually no strongly negative abstracts (e.g., below –0.7), implying that almost no research output is written with an overall negative or pessimistic tone.

In summary, the sentiment analysis demonstrates that the corpus is largely neutral, with a slight positive skew. This slight positivity could reflect authors emphasizing the contributions or positive implications of their research. The lack of negative sentiment outliers suggests consistency in maintaining an academic, impartial tone across the board.

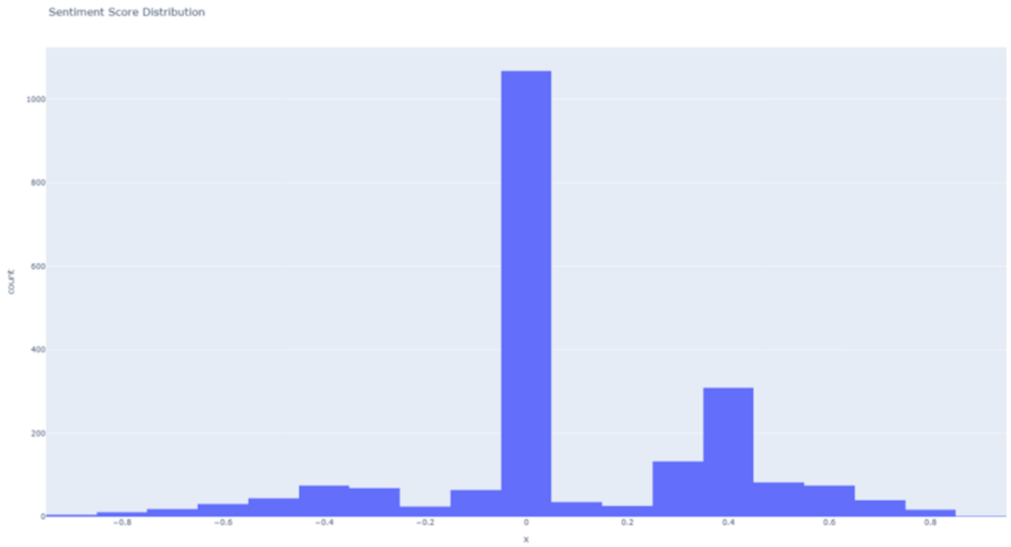

Figure 10: Sentiment Polarity Distribution of Abstracts

4.2.5 Authors per Cluster (Aggregated Author Texts)

Overview

To understand clustering at an author level, we aggregated each author's publications and then performed clustering on these aggregated texts. The rationale is that an author's body of work (considered as one combined text per author) can reveal that author's primary research themes. By clustering authors based on the content of all their abstracts, we can identify groups of authors who work on similar topics. Figure 11 shows a bar chart of the number of authors in each cluster, providing an overview of how authors are distributed among the discovered clusters.

Objectives

- Group authors into clusters according to the similarity of their research topics, effectively creating "communities of practice" or research domains among the authors.
- Examine the distribution of authors across these content-based clusters to see if most authors fall into a few large thematic groups or if they are evenly spread across many niche areas.

Methodology

- **Text Aggregation per Author:** For each author in the dataset, concatenate all the abstracts (or textual content) of that author's publications into a single document representing the author's collective research output. Authors with only one publication are represented by that single abstract.

- **Vectorization:** Transformed each author's aggregated text into a TF–IDF vector (similar to document vectors, but now each vector represents an author). This captures the key terms each author frequently uses or topics they write about.

- **Clustering (K-Means):** Applied K-Means clustering to the author vectors. Silhouette analysis suggested an optimal number of clusters $K = 9$. This clustering may differ from the document clusters due to different grouping structures in author-level data.

- **Visualization:** Counted the number of authors in each cluster and plotted a bar chart of cluster index (x-axis) versus number of authors (y-axis). Each bar in Figure 11 represents a cluster of authors.

Results

The bar chart in Figure 11 shows a markedly uneven distribution of authors across the 9 clusters, revealing that some research themes are pursued by many authors while others are much more niche.

Notably, one cluster contains an exceptionally large number of authors (the tallest bar towers near 290 authors, which is near half of the total authors considered). This suggests the existence of a dominant research theme or domain within the FSRDC outputs that attracts a large community of researchers.

In contrast, several other clusters are comparatively small. The smallest cluster contains only around 20 authors, indicating a very specialized area of research. Other clusters fall in between, each representing moderately popular subfields with 30–80 authors.

The presence of one large cluster alongside several smaller ones suggests that author research interests are not uniformly distributed. This pattern is common in research communities, where a few broad themes garner wide attention, while niche areas are explored by smaller groups. The large cluster likely serves as a central hub, while smaller ones represent satellite communities of focused interest.

These findings underscore the breadth of the FSRDC research landscape: a dominant thematic area is supported by a majority of authors, while diversity exists through multiple smaller, specialized clusters.

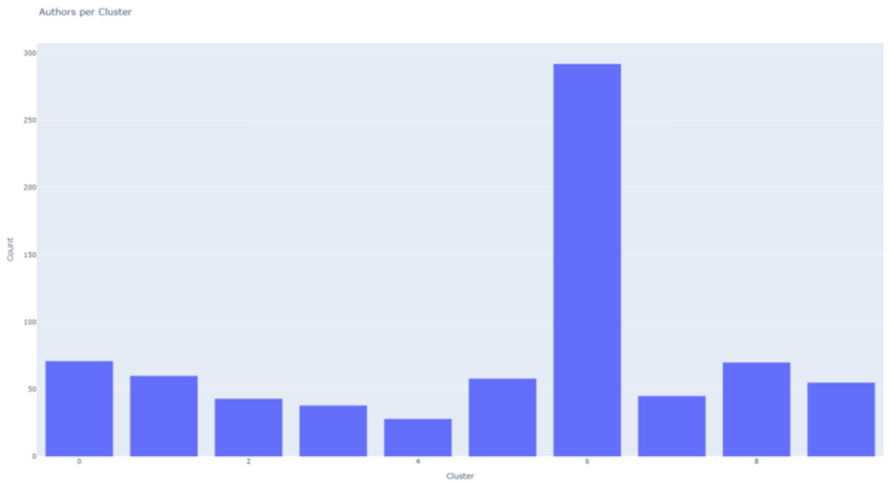

Figure 11: FNumber of Authors per Cluster Based on Aggregated Abstracts

4.2.6 Keyword Frequency Over Time

Overview

This analysis examines how the prominence of specific keywords has changed over the years in the corpus of research outputs. By tracking the frequency of important terms in the abstracts over time, we can identify trends such as emerging topics, growth or decline of particular research themes, and the temporal dynamics of the FSRDC research agenda. Figure 12 presents a multi-line plot showing the frequency of selected keywords by year of publication (`OutputYear`). Each line corresponds to a particular keyword and shows how often that word appears in abstracts each year.

Objectives

- Uncover trends in research topics over the project timeline by observing how keyword usage increases or decreases.
- Identify any notable inflection points or periods where certain terms became significantly more popular, which could correlate with external factors (e.g., policy changes, introduction of new data sources, or methodological innovations influencing many projects at once).
- Provide a temporal context to the earlier cluster and topic analyses: whereas those showed the structure of topics overall, this analysis shows how those topics evolve or fluctuate in importance over time.

Methodology

- **Keyword Selection:** Chose a set of representative keywords to track over time based on relevance and frequency — including domain terms (e.g., “census”, “data”), topic-specific terms from the LDA analysis (e.g., “health”, “trade”, “survey”), and other notable trends.

- **Frequency Counting:** For each keyword, counted its occurrences in all abstracts per year. This produced a time series for each keyword.
- **Normalization (if applicable):** Although document counts vary by year, we used raw counts to reflect absolute interest. This highlights how rising output volume may affect keyword presence.
- **Visualization:** Created a multi-line plot with year on the x-axis and count on the y-axis. Each line corresponds to a keyword, with distinct colors for differentiation.

Results

Figure 12 shows clear temporal trends for several important keywords, reflecting evolving research priorities.

For example, the term *"census"* rises sharply in the 2010s and early 2020s, peaking around 2017 and 2022 — likely aligned with expanded data access and the 2020 Decennial Census. *"Data"* similarly trends upward, reflecting the growth of data-centric research and methodological advances.

Meanwhile, *"health"* shows steady growth post-2005, suggesting increased attention to public health issues in the FSRDC community. Conversely, *"manufacturing"* declines or fluctuates, possibly due to a shift away from industrial themes toward policy, demographic, or social issues.

The plot also shows periodic spikes in *"survey"*, possibly corresponding to major survey releases or analytical focus. The overall trend shows increasing research output — many keywords show minimal counts pre-2000, while post-2010 usage accelerates.

These divergences reveal that some themes have surged while others remain flat or decline. The keyword timelines thus provide dynamic insight into how research interests have shifted with external events, data availability, and emerging priorities in the FSRDC research ecosystem.

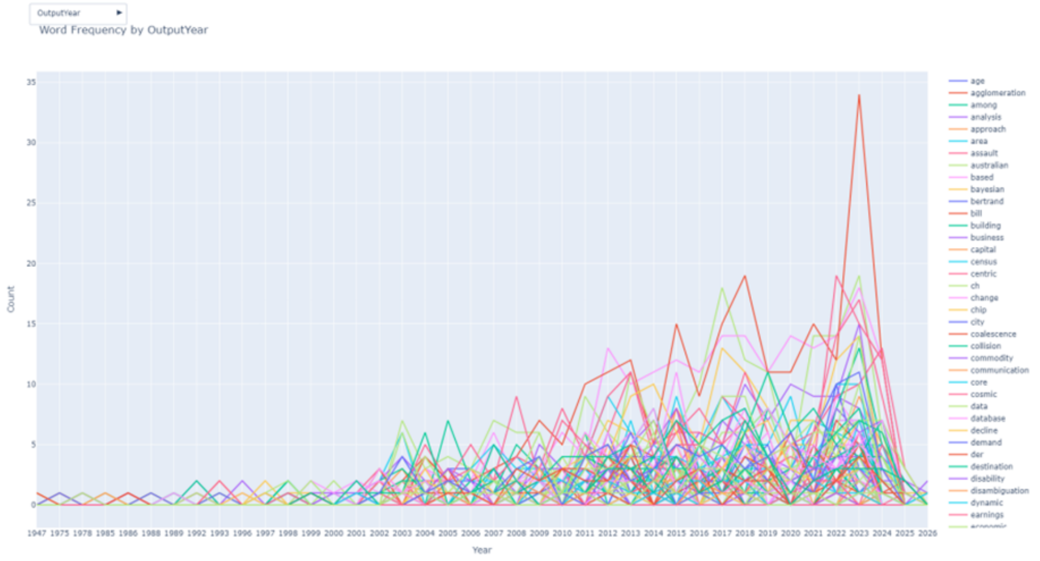

Figure 12: Keyword Frequency Over Time in FSRDC Abstracts

4.2.7 Document Similarity Network (Degree Distribution)

Overview

We constructed a document similarity network to examine how interconnected the research outputs are in terms of content. In this network, each node represents a research document (abstract), and edges connect documents that are very similar in content (based on cosine similarity of their text embeddings). By analyzing this network, we gain insight into whether the corpus consists of mostly isolated works or if there are clusters of very related papers. Figure 13 is a histogram of node degrees in this network, which summarizes how many connections each document has. The degree of a node here corresponds to the number of other documents that are highly similar to it (above a chosen similarity threshold).

Objectives

- Assess the overall connectivity of the document similarity network: do most documents share content links with multiple others, or are they mostly isolated with perhaps only one or two similar peers?
- Identify whether there are "hub" documents that are connected to many others (which could indicate survey papers or broadly thematic papers that overlap with multiple studies) versus documents that stand alone.
- From a content perspective, understand if the knowledge base is fragmented into distinct clusters or forms a giant connected component of interrelated studies.

Methodology

- **Similarity Computation:** Computed cosine similarity between every pair of document TF–IDF vectors. Cosine similarity measures the content overlap between two documents on a scale from 0 (unrelated) to 1 (identical).
- **Network Construction:** We set a similarity threshold — if cosine similarity exceeded this value, we connected the document pair with an edge. To avoid isolated nodes, each document was guaranteed a connection to its most similar neighbor using k-nearest-neighbors logic.
- **Degree Calculation:** Counted the number of strong similarity links for each document. The resulting histogram shows how many documents are similar to 1, 2, 3, etc., other documents.
- **Visualization:** Plotted a histogram of node degrees. The x-axis is degree (number of connected similar documents), and the y-axis is the number of documents with that degree.

Results

The degree distribution shown in Figure 13 indicates that most documents have only one or two strongly similar peers. The tallest bar appears at degree = 1, meaning the majority of abstracts are connected to exactly one other abstract — typically a paper with very similar content or methodology.

Only a few documents have degrees of 3 or 4, and no document connects with more than 4 others. This suggests a right-skewed, sparsely connected network — typical for research corpora where most studies are highly specific. A few documents act as weak hubs (e.g., survey papers or methodologically broad work), but these are rare.

This pattern implies that FSRDC research outputs, while thematically aligned, rarely overlap strongly in content. Many papers form isolated thematic pairs — possibly by the same team or using the same unique dataset. There is no evidence of a highly connected core paper or set of papers dominating the literature.

In summary, the document similarity network highlights the specificity and diversity of the FSRDC research landscape. It complements clustering and LDA topic modeling by emphasizing that shared themes exist mostly in small, local groups rather than across large, unified topic areas.

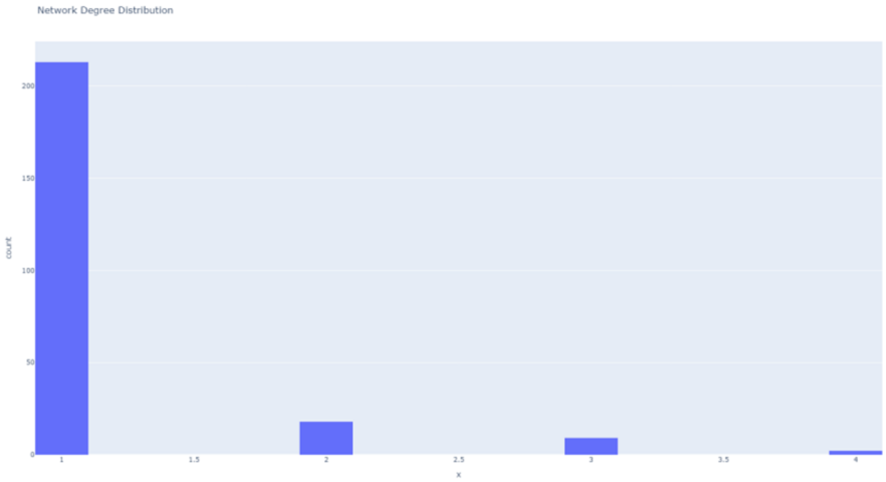

Figure 13: Degree Distribution in the Document Similarity Network

5 Create a GitHub Pages Site

5.1 Overview

GitHub Pages is widely used to showcase research projects. It is a useful platform that can host websites directly from a GitHub repository. It can store codes, data, figures, interactive dashboards, and instructions to tell the readers how to understand the project contents. In Group 7 Project 3, we also created a GitHub Pages website to display our data analysis.

5.2 Method

Our computer operating system is Windows, so we chose to use GitHub Desktop software to assist in the creation of our website. After creating a repository through this software, files can be arranged in the local system. When all of the codes, data, and other results were prepared, we directly uploaded them to the GitHub platform using GitHub Desktop.

The GitHub repository for our project is available at: `https://github.com/ChenxuLi2001/Group-7-Project-3`.

5.3 Structure

To produce a GitHub Pages website, we first edited an HTML file. The titles of the website itself and the content were set. According to the project requirement file, we divided the whole work into pieces, such as EDA, PCA, clustering techniques, and so on. The result analysis was then displayed block-by-block. Figures generated by Python codes were attached with brief explanations. Tables were also added to make our work more organized and verifiable.

6 Error Handling and Testing

Robust error handling and thorough testing were integral to ensuring the reliability and reproducibility of our data pipeline and analysis framework.

During pipeline development, we implemented multiple layers of exception handling to manage unexpected data inconsistencies, missing fields, and API response failures. For example, we introduced safeguards to prevent overwriting existing valid metadata when enriching records via external APIs such as CrossRef and OpenAlex. Whenever a required column (e.g., `DOI`, `URL`, `OutputTitle`) was missing, the system flagged it with a clear warning instead of failing silently. Where possible, fallback strategies were used—such as querying both APIs in sequence or applying conservative fuzzy matching thresholds.

In addition to defensive programming practices, we conducted modular testing at each processing stage. Individual pipeline functions (e.g., normalization, deduplication, metadata enrichment, TF-IDF vectorization) were tested independently using subsets of data with known outputs to verify correctness. We also validated final outputs against schema expectations to ensure compliance with the required format for merged deliverables.

These practices helped us isolate and resolve edge cases, such as malformed DOIs, non-ASCII characters in abstracts, or conflicting metadata from multiple sources. Overall, our testing strategy ensured that data integrity was preserved while maintaining compatibility with downstream analysis stages.

7 Insights and Takeaways

This project offered extensive hands-on experience with real-world data integration, exploratory analysis, and unsupervised learning methods applied to a complex and fragmented research dataset.

Some of the key insights include:

- **Metadata inconsistencies are a major bottleneck in multi-source research**

aggregation. Normalizing and deduplicating across eight independent datasets required careful schema unification, fuzzy matching, and custom heuristics.

- **API enrichment is powerful but must be used judiciously.** External APIs like CrossRef and OpenAlex filled critical metadata gaps, but their accuracy varied. Our strategy of using them only to populate missing values helped retain source fidelity.
- **Clustering and topic modeling revealed meaningful thematic structures.** Both K-Means and Agglomerative clustering showed stable groupings, while LDA topics aligned well with real research themes (e.g., census data, business economics, public health).
- **Most abstracts are neutral in tone, with slight positive sentiment skew.** Sentiment analysis validated expectations about the academic tone of abstracts while identifying subtle trends in language.
- **Keyword trends change over time.** Terms like “census” and “data” have grown in frequency, reflecting broader interest in data-centric methodologies and new dataset availability.

Overall, the pipeline and analytical techniques we developed can be generalized to similar metadata-rich academic corpora, offering scalable insights into publication behavior, collaboration networks, and research trends.

8 Project Conclusion

Project 3 served as the capstone for our FSRDC research data pipeline, culminating in the integration, validation, and in-depth analysis of thousands of publication records. We successfully merged diverse group submissions into a unified dataset, enriched it with metadata from reliable APIs, and used machine learning tools to uncover thematic structures and evolving research dynamics.

Our analysis highlighted the richness and diversity of the FSRDC research landscape. Despite varied topics and methodologies, shared patterns emerged—both in document similarity and author behavior. Clustering, PCA, LDA, and sentiment analysis provided multidimensional perspectives on this body of work.

The combination of rigorous data cleaning, validation logic, reproducible code, and interpretive visualizations not only satisfied the project’s technical goals but also built foundational skills in collaborative, end-to-end data science. The final GitHub Pages site showcases our work for a wider audience, providing an accessible summary of our methods, results, and visual insights.

We conclude that the techniques used here—when combined with careful validation—offer a powerful toolkit for mapping large, unstructured research spaces and can serve as a model for future interdisciplinary projects.