Unguided and Guided LDA analysis of News Articles

Joel Franklin Stalin Vijayakumar

Introduction

National Digital Newspaper Program (NDNP) is a collaboration between the National Endowment for the Humanities (NEH) and the Library of Congress and it offers a comprehensive and reliable source for accessing an extensive archive of U.S. newspaper articles.

Chronicling America, a key output of the NDNP, allows users to search America's historic newspaper pages from **1756-1963.**

The NDNP has been ongoing since 2005, with regular updates and additions to the database. There's an ongoing effort to digitize and make available historic newspaper pages.

Navigating the Chronicling America Database: From Batches to Pages

The file https://chroniclingamerica.loc.gov/batches/1.json serves as part of a paginated directory listing all available batches in the Chronicling America database. This directory includes multiple batches, each containing a collection of digitized newspaper data.

The first batch in the https://chroniclingamerica.loc.gov/batches/1.json directory is identified by the URL https://chroniclingamerica.loc.gov/batches/nbu_bluejay_ver02.json and contains a total of 9902 pages. This batch, like others, comprises various issues of different newspapers.

Within the "nbu_bluejay_ver02" batch, the first newspaper listed is the edition from February 15, 1918, accessible via https://chroniclingamerica.loc.gov/lccn/sn83045201/1918-02-15/ed-1.json. This particular edition consists of 8 pages.

The first page of this February 15, 1918, edition is detailed at https://chroniclingamerica.loc.gov/lccn/sn83045201/1918-02-15/ed-1/seq-1.json, where specific resources related to the page are listed.

For this first page, we can find the PDF version at https://chroniclingamerica.loc.gov/lccn/sn83045201/1918-02-15/ed-1/seq-1.pdf and the OCR (optical character recognition) extracted text at https://chroniclingamerica.loc.gov/lccn/sn83045201/1918-02-15/ed-1/seq-1/ocr.txt, allowing for both viewing the original page image and accessing its text content.

Question 3

Selecting manually curated 5 topics that are likely to have been prevalent over the past 25 years (1960 – 1936) and developing a list of 10 seed words that encapsulate these 5 topics.

1. **Sports**: The seed words are "game", "team", "season", "play", "club", "win", "match", "score", "player", and "coach". I chose Sports as a topic because it has been a significant part of society's fabric, reflecting cultural, technological, and socio-political shifts from 1936 to 1960. This period saw historical events influencing sports, from league integrations to sports being used for political diplomacy. These words encompass the competitive nature, organizational structure, and key figures in sports, shedding light on its evolution during these years.

2. **Rentals & Real Estate**: The seed words are "house", "home", "room", "property", "rent", "estate", "apartment", "building", "lease", and "mortgage". This topic is relevant for understanding the changing living standards, urban development, and economic conditions between 1936 and 1960. The era marked by post-war recovery and suburban expansion significantly influenced housing policies and real estate development. These words capture the essence of housing arrangements and financial aspects, allowing an exploration of how societal attitudes towards homeownership and community layout transformed.

3. **Philosophy & Thought**: Seed words include "time", "life", "man", "world", "philosophy", "thought", "mind", "idea", "reason", and "belief". This topic was chosen to delve into the intellectual and cultural currents that shaped societal values and beliefs during a period marked by profound changes. Between 1936 and 1960, the world witnessed war, technological advancements, and shifts in social norms, which prompted deep philosophical and existential inquiries. The selected words aim to uncover the dominant ideologies, the quest for meaning, and the evolution of thought that characterized this era.

4. **Community Gatherings/Events**: The seed words are "church", "school", "event", "member", "community", "meeting", "ceremony", "celebration", "gathering", and "festival". Community gatherings and events play a pivotal role in social cohesion and cultural expression. During the years 1936 to 1960, such gatherings were essential for maintaining morale, fostering a sense of belonging, and preserving cultural heritage amid significant societal upheavals. These words were chosen to investigate the various forms of social and communal interactions, highlighting their importance in maintaining social fabric during times of change.

5. **Politics/Government**: The seed words are "president", "state", "government", "senate", "congress", "election", "policy", "law", "political", and "diplomacy". This topic is crucial for understanding the governance dynamics and political developments during a transformative period in history. The selected years encompass critical phases such as the aftermath of the Great Depression, World War II, and the early Cold War years. These words provide a framework to explore the political landscape, policy decisions, and international relations that defined the era, offering insights into the mechanisms of power and governance that influenced global and domestic affairs.

Data Extraction First Attempt

In the first attempt, I developed a function to fetch OCR text from a series of web pages, save it into a file, and halt the process once a specified maximum file size was reached. This task involved interacting with an API to navigate through various URLs, including batches, issues, and pages to extract the desired text data.

The process began by ensuring the target directory existed and then creating an empty file where the OCR text would be saved. I iterated through batch URLs to fetch issue URLs, and from each issue URL, I extracted page URLs. For each page, I retrieved the OCR text URL, which contained the actual text data I needed.

I made sure to keep track of the file size after appending new text to avoid exceeding the maximum file size limit I set. If the limit was reached, the function would stop fetching and saving new data, effectively controlling the volume of data being processed and stored. This mechanism was crucial for managing resources efficiently and avoiding unnecessary load on the server hosting the data.

Throughout the process, I handled exceptions gracefully to ensure the program's robustness against potential errors like failed HTTP requests or issues accessing or writing to the file system. This approach ensured that the task could complete successfully or fail cleanly, providing meaningful error messages to aid troubleshooting.

The entire operation was encapsulated in a single function, making it reusable for different URLs or file size limits. This design choice enhanced the flexibility of my solution, allowing it to be adapted easily for various data extraction and storage needs.

But I could extract less than 10 MB of data as the server didn’t allow me further to extract more data.

Data Extraction Second Attempt

To start my project on analyzing historical newspaper articles, I first needed to find a good source of data. I chose the "dell-research-harvard/AmericanStories" dataset because it's well-known for having a wide range of newspaper articles from over the years. This dataset gave me access to articles from the year I was interested in, allowing me to start my analysis with a solid foundation.

Next, I worked on getting the articles ready for analysis. This involved writing code to download the articles year by year, save them in a text format, and keep track of how much space they were taking up. I did this to make sure I didn't go over my storage limit. Once I had all the articles I needed, I could move on to examining them more closely.

To understand the data better, I calculated the average size of the articles. This helped me figure out how many articles I could work with without exceeding my storage limit. After getting a good sense of the data's size, I loaded articles from the most recent 25 years,

making sure the total didn't go over 50MB. This gave me a manageable subset of data to start with.

With this subset, I began preparing the data for analysis. This meant cleaning up the text in the articles, combining headlines with article bodies, and removing any unnecessary words or symbols. I also split the text into simpler forms, a process known as lemmatization, to make the analysis more straightforward.

Finally, I wanted to ensure the dataset's quality was high and covered a wide range of topics from different years. To do this, I manually checked random samples of the dataset, looking for any errors and making sure the metadata was complete. This step was crucial to ensure the data I'd be working with was reliable and comprehensive.

This whole process, from finding the right dataset to preparing the data for analysis, was a crucial first step in my project. It set the stage for the more detailed analysis I planned to do next, aiming to uncover the main topics discussed in American newspapers over the last century.

Data Preparation

To embark on my project, I started by acquiring a substantial dataset that spans the last 25 years, focusing on U.S. news articles. This dataset, housed in an Excel file named "Last_25_years_250_MB_dataset.xlsx," was pivotal for my analysis. My goal was to delve deep into the textual content of these articles to discern patterns, trends, and the evolution of societal discourse over the years.

Given the dataset's raw nature, the first step involved preparing the data for topic modeling. This process entailed concatenating article headlines with their corresponding body text to form a cohesive block of text for each article. Recognizing the importance of clean, analyzable text, I applied a series of preprocessing steps: stripping special characters and punctuation to minimize noise, converting all text to lowercase to ensure uniformity, and tokenizing the text into individual words for detailed examination.

To refine the dataset further, I leveraged the Natural Language Toolkit (NLTK), a powerful library for text processing. By removing common stopwords—words that, while frequent, offer little value in understanding the text's essence—I could focus on the meaningful words that contribute to thematic richness. Additionally, I employed lemmatization, a technique that condenses words to their base or dictionary form, thereby reducing the dataset's complexity without sacrificing depth.

With the data now clean and structured, I meticulously recorded each article's publication date, extracting the year, month, and day. This temporal breakdown would later enable a nuanced analysis of how topics and themes have shifted over time.

To ensure accessibility and ease of use for subsequent analyses, I consolidated the prepared data into a new Excel file, "Prepared_last_25_years_250_MB_dataset.xlsx," and stored it in a designated directory. This step marked a significant milestone in my project, setting the stage for deep dives into topic modeling and uncovering the narratives woven through decades of journalistic reporting.

Unguided LDA - Hyperparameter Tuning

After preparing the dataset by concatenating headlines and articles, cleaning, and preprocessing the text, I decided to embark on an exploration of unguided LDA (Latent Dirichlet Allocation) to uncover the underlying topics within the historical newspaper dataset covering the years from 1936 to 1960. This approach allowed me to dive into the content without preconceived notions about the topics, thus providing a broader understanding of the thematic landscape within the dataset.

To conduct the unguided LDA, I treated all entries in the 'prepared_text' column as strings and split them into lists of words. I also filtered out words with less than three characters to focus on more meaningful terms. This initial preprocessing was crucial in preparing the data for effective topic modeling.

I explored various configurations for the LDA model, adjusting the 'no_above' parameter to filter out words that appear too frequently across documents. This parameter was vital in ensuring that the topics discovered by the LDA model were not dominated by overly common words that could detract from the thematic specificity. After training models with different 'no_above' values, I evaluated each model's coherence score, which measures the semantic similarity between high scoring words in each topic, aiming to find the configuration that yielded the most interpretable and coherent topics.

The exploration revealed that adjusting the 'no_above' parameter had a significant impact on the coherence of the topics identified by the LDA model. The best model, with a 'no_above' value of 0.07, achieved the highest coherence score, indicating that it was able to uncover more meaningful and cohesive topics compared to models trained with other 'no_above' values. This model revealed topics related to court and county matters, government and committee discussions, automotive and sales, club and church events, and sports and leisure activities, reflecting the diverse content of the historical newspaper dataset.

Through this unguided LDA exploration, I gained insights into the varied thematic content of the newspaper articles from 1936 to 1960. The process not only highlighted the importance of parameter tuning in topic modeling but also showcased the potential of LDA to uncover hidden thematic structures in large text corpora, providing a foundation for further analysis and interpretation of historical news content.

Unguided LDA utilizing Bigrams and TFIDF - Hyperparameter Tuning

Continuing from the initial stages of my project, where I delved into the complexities of topic modeling and the preparation of the dataset, the journey led me to the critical phase of fine-tuning and evaluating the Latent Dirichlet Allocation (LDA) models. The meticulous approach adopted in selecting and preprocessing the dataset laid a solid foundation for the exploration of various hyperparameters to optimize the LDA model's performance.

As part of the optimization process, I conducted an extensive exploration of key hyperparameters such as 'no_below', 'no_above', and 'low_value'. These parameters were instrumental in refining the document-term matrix by controlling the word frequency thresholds and applying TF-IDF filtering. This step was crucial to ensure that the model focused on the most meaningful content, thereby enhancing its ability to uncover coherent and interpretable topics.

The training of multiple LDA models, each with a different set of hyperparameters, was an exhaustive yet enlightening exercise. It allowed me to assess the model's performance based on coherence scores, a reliable metric for gauging the semantic similarity among high-scoring words within topics. Higher coherence scores indicated a greater level of topic interpretability and semantic coherence, guiding me towards the optimal model configuration.

The resulting topics from the best-performing model provided a fascinating glimpse into the historical period covered by the dataset. The model successfully identified diverse themes ranging from political and economic discussions to community events and cultural narratives. Each topic served as a unique lens through which to view and interpret the societal values, concerns, and interests prevalent during the years 1936 to 1960.

In sum, the latter stages of my project underscored the significance of a well-considered and methodical approach to topic modeling. From preprocessing and hyperparameter tuning to model evaluation, each step was pivotal in harnessing the potential of LDA to reveal the hidden thematic structures within a rich historical newspaper dataset. This journey not only enhanced my understanding of topic modeling techniques but also highlighted the invaluable insights that can be gleaned from historical text data, offering a window into the past through the written word.

Unguided LDA utilizing Bigrams, TFIDF and NER

As my exploration advanced into the nuanced realms of topic modeling, integrating Named Entity Recognition (NER) and bigram phrase detection became pivotal steps. These methodologies, aimed at enriching the textual data's semantic quality, underscored my commitment to extracting coherent and robust topics. Through the application of NER, I

accentuated key entities within the corpus, transforming them into single tokens that preserved the essence of the original text. This emphasis on entities not only improved the interpretability of the topics but also elevated the granularity with which the model could capture thematic nuances.

Similarly, the incorporation of bigram phrase detection marked a methodological advancement, allowing for the identification and amalgamation of frequently co-occurring words into singular bigram tokens. This process significantly enhanced the contextual relevance of the topics unearthed by the LDA model, as it acknowledged the inherent value of word pairings that carried specific meanings often lost when considered in isolation.

The culmination of these preprocessing steps laid the groundwork for the final model training phase. Leveraging a meticulously fine-tuned Latent Dirichlet Allocation (LDA) model, enriched with the insights from NER and bigrams, I embarked on the task of identifying the latent topics within the historical dataset. The model was configured with carefully selected hyperparameters, ensuring a balance between specificity and generality, aimed at maximizing the coherence and interpretability of the resulting topics.

Upon training, the LDA model revealed distinct thematic clusters that resonated with the historical period's socio-cultural and political landscape. Each topic, a mosaic of terms enhanced by entity emphasis and bigram cohesion, offered a unique lens through which to interpret the dataset. The coherence of these topics not only validated the methodological choices made along the way but also illuminated the rich tapestry of narratives embedded within the corpus.

Reflecting on the journey, from dataset preparation through to the final modeling, it's evident that the integration of NER and bigram phrase detection was instrumental in achieving a nuanced understanding of the underlying themes. The completion of this phase not only signifies a milestone in my project but also reinforces the value of advanced preprocessing techniques in unearthing the latent structures that define large text corpora. As I look forward to the insights these topics will unveil, the groundwork laid throughout this process stands as a testament to the iterative, explorative nature of working with natural language processing and topic modeling.

Guided LDA Manual Approach – Hyperparameter Tuning

In a continuous stride toward enhancing topic discovery, I employed a strategic approach of amplifying the influence of pre-selected seed words across various thematic areas. This methodological refinement aimed to steer the topic modeling process toward generating themes closely aligned with the seed topics' essence. By meticulously choosing a set of seed words for each theme, such as "Sports" with words like "game," "team," and "match," and "Rentals & Real Estate," highlighting "house," "property," and "rent," I laid a focused foundation that echoes the significant discourse within the historical news article corpus spanning from 1936 to 1960.

The process involved adjusting the emphasis factor for these seed words, experimenting with different intensities to gauge the optimal influence on the model's output. This experimentation was not arbitrary; instead, it was deeply rooted in the hypothesis that a calibrated emphasis could refine the model's sensitivity to the nuanced lexicon of each theme. For instance, the factor adjustment directly impacted themes such as "Philosophy & Thought" and "Community Gatherings/Events," enabling the model to more precisely encapsulate the thematic depth through enriched lexical representations.

The subsequent training of the Latent Dirichlet Allocation (LDA) model, post seed word emphasis, revealed an intriguing array of topics. Each reflected a nuanced understanding of the underlying narratives present in the corpus, with themes ranging from socio-political dynamics captured in "Politics/Government" to the cultural and intellectual discourses within "Philosophy & Thought." The coherence scores, serving as a quantitative testament to the model's efficacy, underscored the beneficial impact of integrating seed word emphasis into the preprocessing pipeline.

This phase of the project, marked by the exploration of seed word emphasis and its influence on topic coherence, signifies a critical juncture in my quest to unveil the latent thematic structures within historical newspaper articles. The findings not only underscore the potential of tailored preprocessing techniques in enhancing topic model quality but also pave the way for further inquiries into the intricate balance between human intuition and algorithmic precision in the domain of natural language processing and topic discovery.

Guided LDA with CorEx – Hyperparameter Tuning

Continuing my exploration of guided topic modeling, I ventured into the realm of CorEx, an algorithm designed to discover latent themes with the assistance of domain knowledge through seed words. This part of my journey involved experimenting with a technique that integrates predefined seed words, such as "game" and "team" for sports, and "house" and "rent" for rentals & real estate, into the CorEx topic modeling framework. My intention was to steer the model towards uncovering topics that not only resonate with the selected themes but also encapsulate the essence of the discussions within the historical newspaper articles spanning from 1936 to 1960.

The procedure entailed transforming the dataset into a document-term matrix, a necessary precursor for CorEx. This step was meticulously executed, ensuring the optimal representation of the dataset's vocabulary through the careful selection of the maximum features parameter, which was varied across several iterations to evaluate its impact on the model's performance. This strategic manipulation aimed to balance comprehensiveness with computational efficiency, reflecting a nuanced approach to model tuning.

The introduction of seed words into the CorEx model as anchors—a unique feature of the algorithm—allowed me to infuse the model with my domain insights, effectively guiding the topic discovery process. For example, anchoring the theme of "Philosophy & Thought" with

words like "time," "life," and "mind" aimed to direct the model's attention towards capturing the intellectual and existential discourses prevalent in the articles.

Through iterative training sessions, each marked by the adjustment of the maximum features parameter and the evaluation of coherence scores, I was able to observe how different configurations influenced the emergent topics. The coherence scores, serving as a beacon of model efficacy, offered quantitative insights into the thematic clarity and relevance of the discovered topics, guiding my decisions on model refinement.

This exploratory phase, characterized by the application of CorEx guided topic modeling, underscored the potential of leveraging domain knowledge through seed words to enhance topic discovery. The insights garnered from this endeavor illuminate the intricate dance between algorithmic sophistication and human intuition, marking a significant milestone in my quest to unravel the latent thematic structures within the corpus of historical newspaper articles.

Report 5 sets of 10 seeding words and the corresponding resulting top 10 loading words. Also, report top 10 loading words for 5 unseeded topics.

Manually chosen 5 topics of 10 seeding words

1. **Sports** - game, team, season, play, club, win, match, score, player, coach
2. **Rentals & Real Estate** - house, home, room, property, rent, estate, apartment, building, lease, mortgage
3. **Philosophy & Thought** - time, life, man, world, philosophy, thought, mind, idea, reason, belief
4. **Community Gatherings/Events** - church, school, event, member, community, meeting, ceremony, celebration, gathering, festival
5. **Politics/Government** - president, state, government, senate, congress, election, policy, law, political, diplomacy

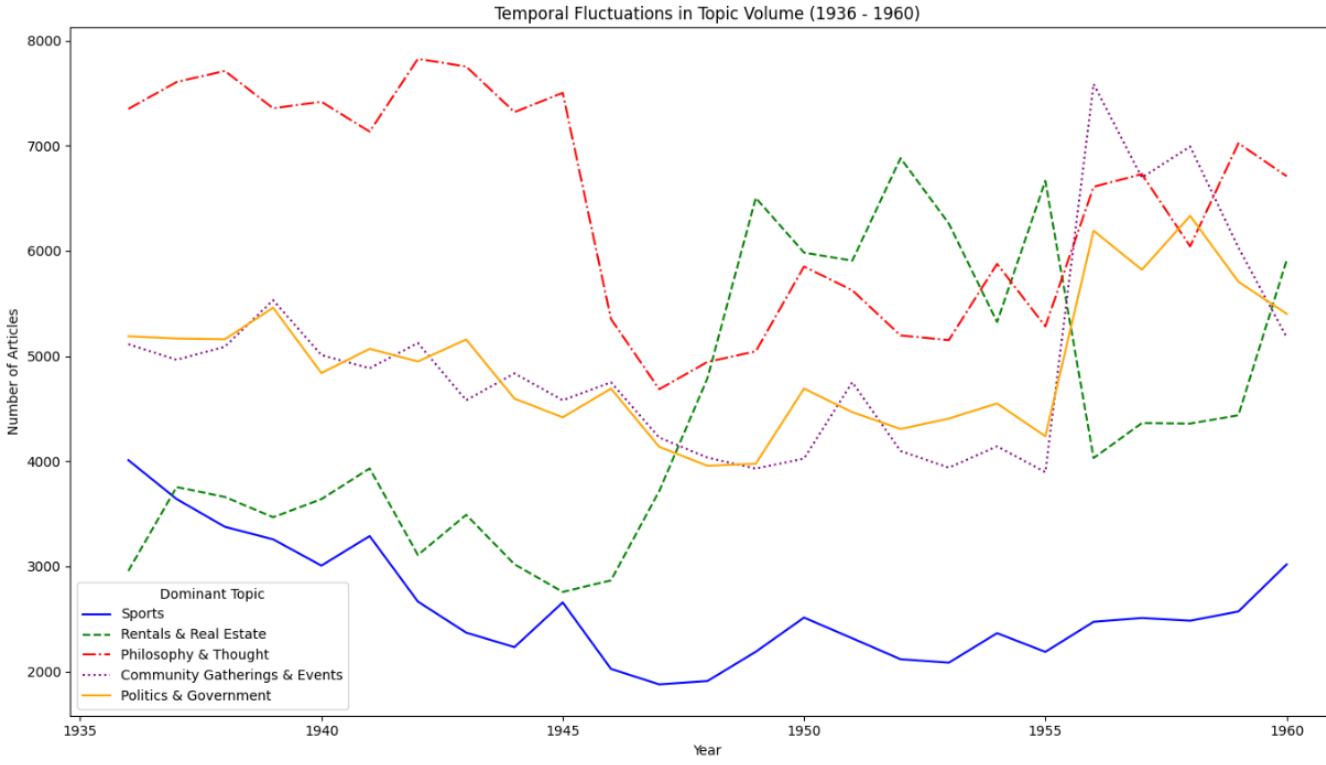
Guided LDA chosen 5 topics of 10 seeding words (Coherence score = 0.471)

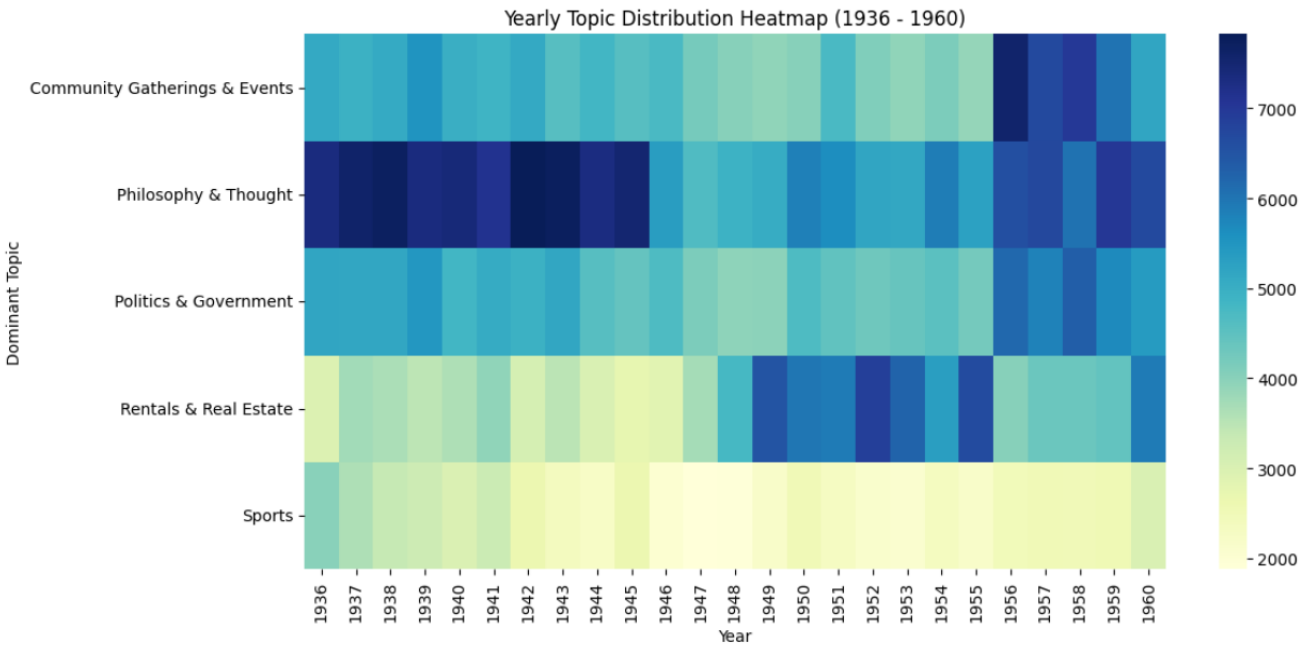
1. **Sports** - club, game, team, play, building, win, season, event, score, player
2. **Rentals & Real Estate** - house, room, home, property, rent, estate, apartment, new, car, lot
3. **Philosophy & Thought** - time, man, life, world, thought, one, idea, mind, reason, said
4. **Community Gatherings/Events** - home, school, church, member, meeting, john, ceremony, william, son, street
5. **Politics/Government** - state, president, government, law, member, congress, election, senate, political, policy

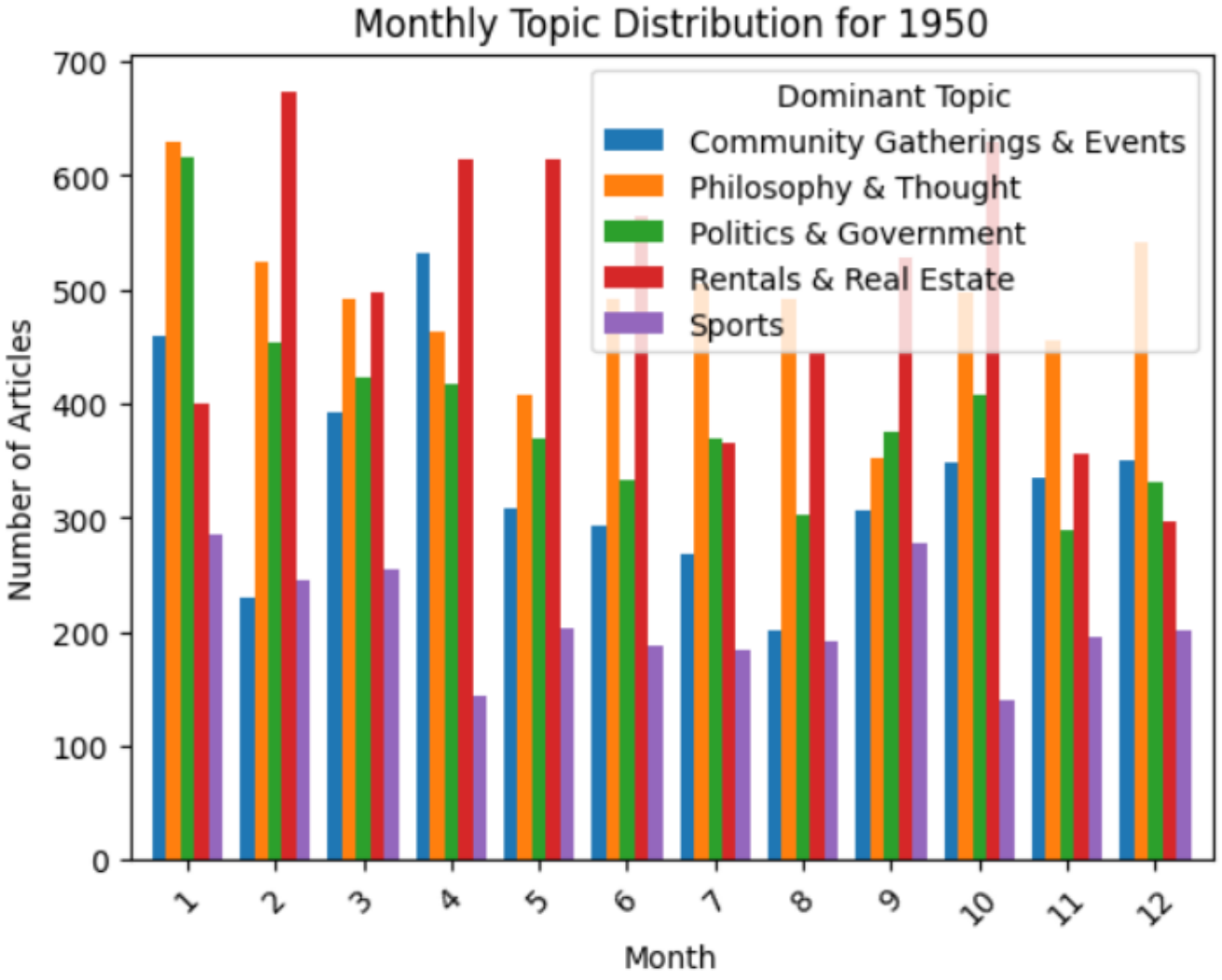
Unguided LDA chosen 5 topics of 10 seeding words (Coherence score = 0.53)

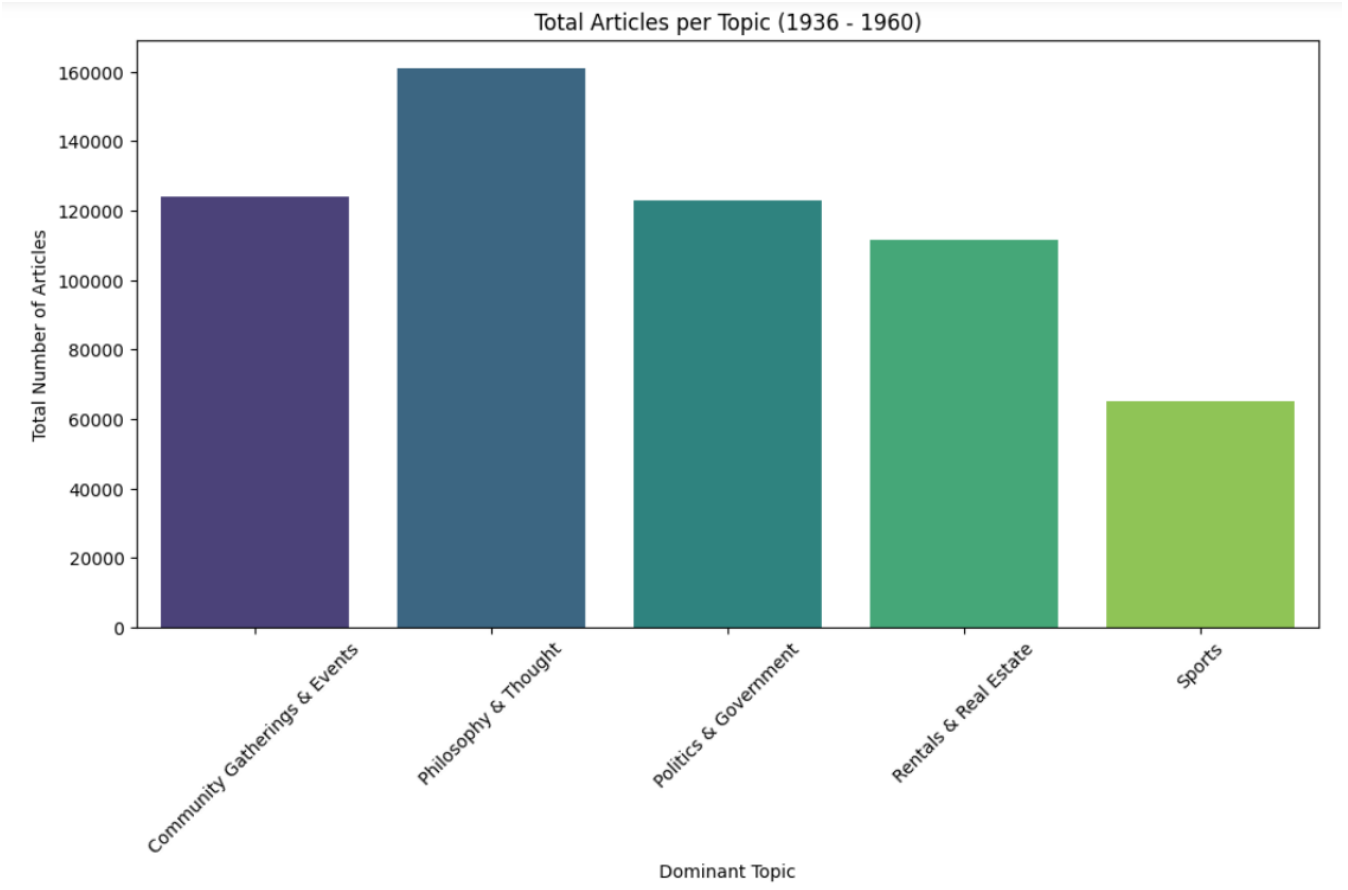
1. **Sports** - game, two, one, team, first, play, new_york, three, season, club
2. **Rentals & Real Estate** - room, new, car, lot, 500, home, phone, 350, house, ave
3. **Philosophy & Thought** - one, said, would, time, man, day, many, say, way, two
4. **Community Gatherings & Events** - home, john, church, son, william, school, street, member, miss, club
5. **Politics & Government** - state, said, year, new, would, today, president, district, may, committee

Graphics and Visualizations of Unguided LDA

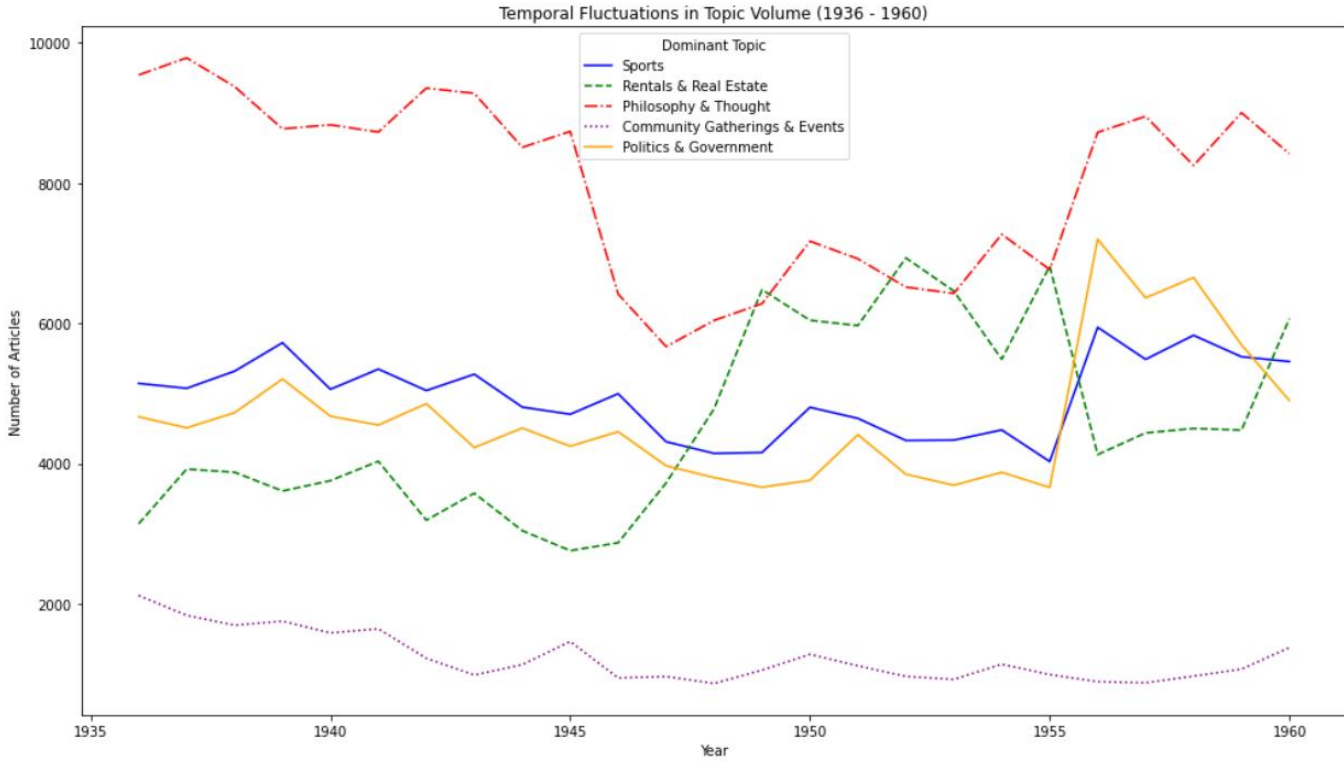


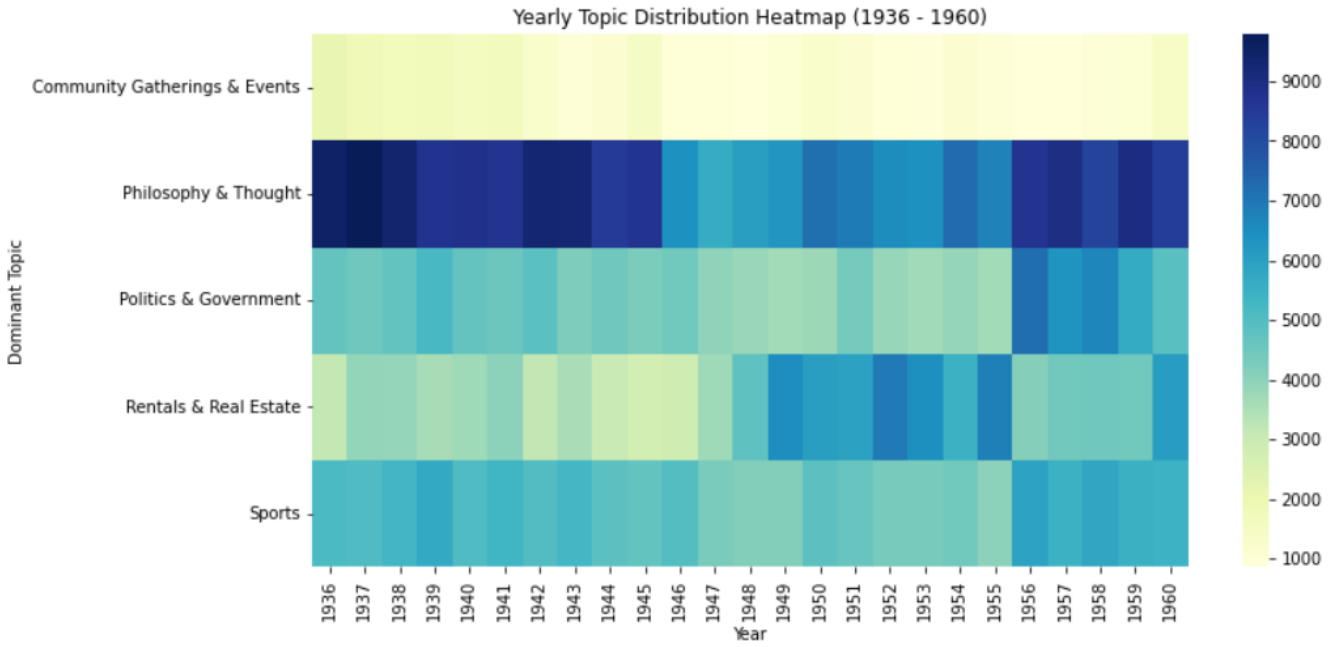

Yearly Topic Distribution Heatmap (1936 - 1960)
Dominant Topic
Community Gatherings & Events
Philosophy & Thought
Politics & Government
Rentals & Real Estate
Sports
1936
1937
1938
1939
1940
1941
1942
1943
1944
1945
1946
1947
1948
1949
1950
1951
1952
1953
1954
1955
1956
1957
1958
1959
1960
Year
7000
6000
5000
4000
3000
2000

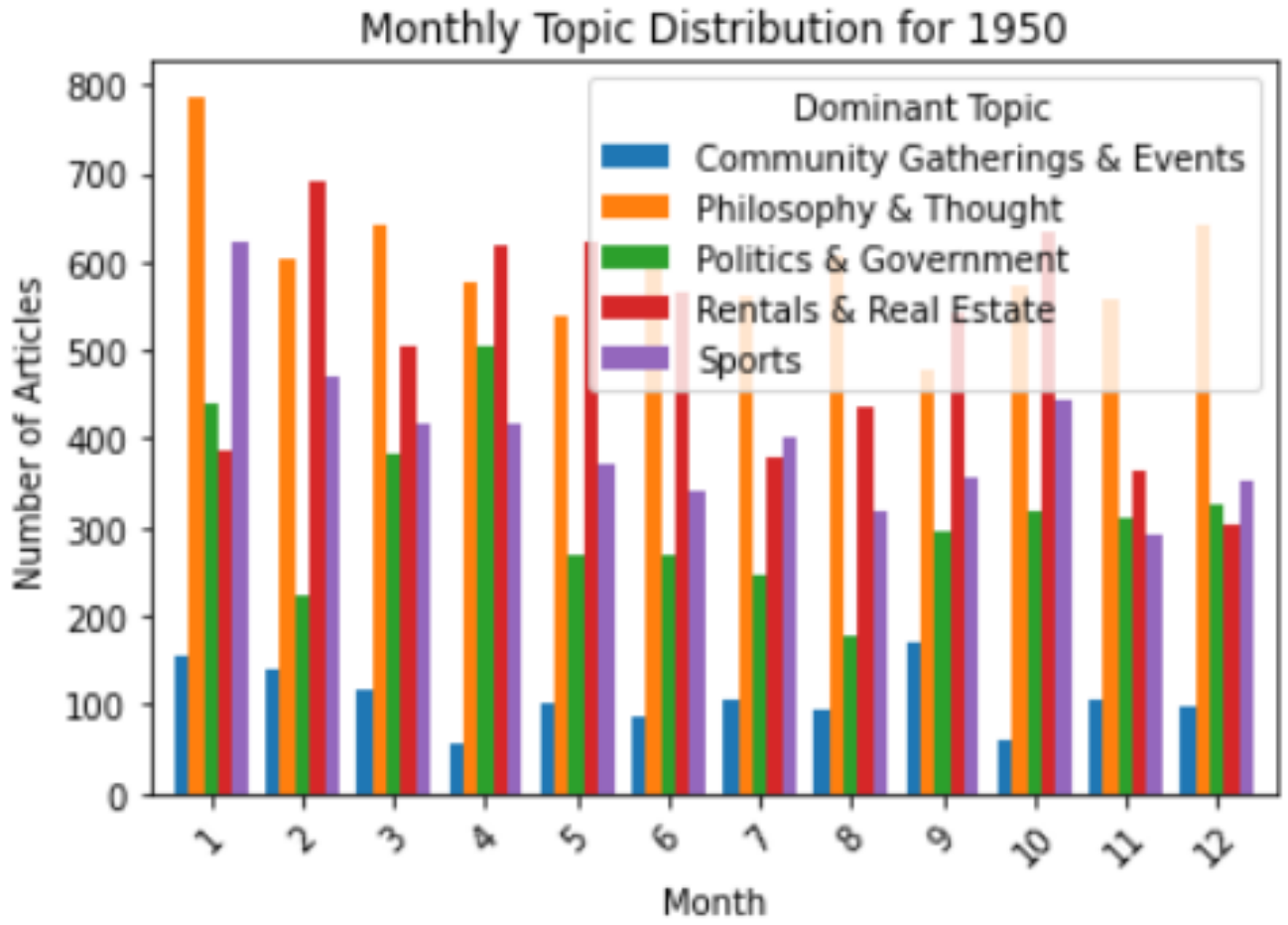

Monthly Topic Distribution for 1950
Number of Articles
0
100
200
300
400
500
600
700
Month
1
2
3
4
5
6
7
8
9
10
11
12
Dominant Topic
Community Gatherings & Events
Philosophy & Thought
Politics & Government
Rentals & Real Estate
Sports

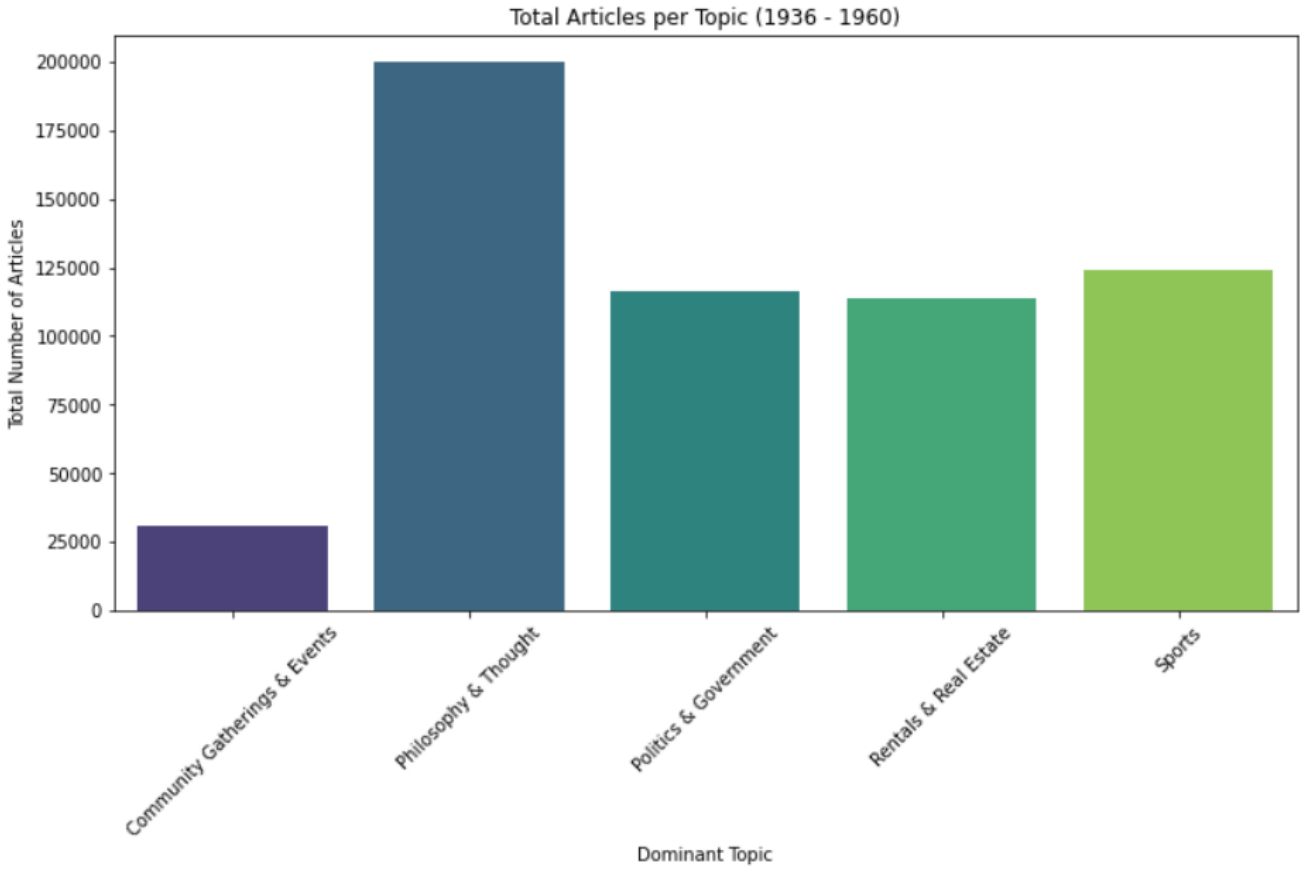


Graphics and Visualizations of Guided LDA




Yearly Topic Distribution Heatmap (1936 - 1960)
Community Gatherings & Events
Philosophy & Thought
Politics & Government
Rentals & Real Estate
Sports
Dominant Topic
1936
1937
1938
1939
1940
1941
1942
1943
1944
1945
1946
1947
1948
1949
1950
1951
1952
1953
1954
1955
1956
1957
1958
1959
1960
Year
9000
8000
7000
6000
5000
4000
3000
2000
1000


Monthly Topic Distribution for 1950
Dominant Topic
Community Gatherings & Events
Philosophy & Thought
Politics & Government
Rentals & Real Estate
Sports
Number of Articles
0
100
200
300
400
500
600
700
800
1
2
3
4
5
6
7
8
9
10
11
12
Month



I hereby affirm that the work presented in this report is entirely my own, derived from my personal efforts and research. No part of this work has been plagiarized or copied from any other source.

Joel Franklin Stalin Vijayakumar

03-18-2024