

Using Big Data to Analyze 21st-Century Controversial Topic Evolution

Daniel DeMasi

Lehigh University

CSE 349/449 Final Project Writeup

djd225@lehigh.edu

ABSTRACT

The 21st century has been marked by transformative discussions surrounding controversial topics such as gender, identity, and discrimination. These conversations are shaped by evolving societal values, global movements, and the growing influence of digital platforms. Issues like gender equality, LGBTQ+ rights, and racial justice have not only fueled worldwide activism but have also sparked debates that challenge long-held norms. Movements such as #BlackLivesMatter, the resurgence of #MeToo, and the overturning of *Roe v. Wade* highlight how public discourse can influence policy, cultural perceptions, and individual lives. Understanding the evolution of these discussions is crucial, as they reveal the changing priorities of society.

Open, collaborative platforms like Wikipedia play a central role in documenting and shaping these debates, but their vast scale and dynamic nature make identifying trends complex. Furthermore, online discourse is often characterized by rapid growth, noise, and biases, which complicate meaningful analysis of societal change [1]. These factors underscore the need for robust methods to analyze evolving discourse.

This project leverages PySpark's distributed computing capabilities to process 388 Wikipedia pages spanning over 6,000 revisions from 2001–2024. Using Latent Dirichlet Allocation (LDA) for topic modeling, the project combines natural language processing, text analysis, temporal analysis, and machine learning to examine the evolution of online discourse. Through rigorous preprocessing with PySpark's NLP tools and fine-tuning of LDA parameters, the project identifies the most interpretable and well-fitting topics. This approach reveals temporal trends in language and topic prevalence, connecting them to real-world events such as #BlackLivesMatter and the overturning of *Roe v. Wade*. The project employs visualizations, including stacked bar charts and heatmaps, to provide data-driven insights into shifting societal focus over time.

Ultimately, the project seeks to answer the question: How has discourse on controversial topics evolved over time in Wikipedia revisions? The project's findings reveal increasing inclusivity and intersectionality in language, reflecting broader societal awareness and shifts in public dialogue. This project demonstrates the value of big data analytics and PySpark's scalability in understanding public discourse and its societal impact.

1. TOOLS

For this project, I used multiple tools like PySpark and Python for data analysis, preprocessing, and visualization. PySpark is a distributed computing framework for big data analytics that makes it easy to write efficient processing, querying, and machine learning scripts on my large-scale dataset using Python. NumPy was used to calculate summary statistics, while PySpark NLP handled tokenization, normalization, stopword removal, stemming, lemmatization, and n-gram generation for Wikipedia revision data. Regex (Python `re`) was also used to clean Wikipedia-specific elements like keywords, HTML tags, and unnecessary markup during preprocessing. PySpark SQL and UDFs helped in querying and making custom dataset transformations, and PySpark MLlib included CountVectorizer, IDF, and LDA for vectorization and topic modeling. Gensim was used for coherence scoring (`c_v`) and dictionary creation to evaluate topic quality approximating human judgement. For visualization, Matplotlib generated

stacked bar charts and heatmaps to illustrate topic distributions over time.

2. DATA

This project uses Wikipedia page dumps, exported in XML format, as the primary dataset. The data includes information such as titles, revision timestamps, comments, content, and authors, offering an opportunity for an analysis of revisions and the evolution of discourse over time. The dataset was retrieved using the Wikipedia API, which facilitates exporting selected pages [2]. Below, I provide an overview of the selected pages, revision history, and summary statistics for both the subsample and complete dataset.

Selected Wikipedia Pages

The dataset focuses on topics central to 21st-century discussions, ranging across themes such as gender, sexuality, race, and social movements. Examples include:

- A. Gender and Sexuality, Gender Studies, Non-Binary Gender, LGBTQ Rights, Intersectionality
- B. Feminism Movements and Ideologies, First-Wave Feminism, Ecofeminism
- C. Politics and Activism, Culture Wars, Identity Politics, Black Feminism, Critical Race Theory
- D. Workplace and Economic Issues, Gender Pay Gap, Glass Ceiling, Feminization of Poverty
- E. Violence and Oppression, Rape Culture, Misogyny
- F. Legal and Policy Issues, Roe v. Wade, Title IX
- G. Representation and Media, Gender Representation in Video Games, Media Portrayal of LGBTQ People
- H. Historic Events and Movements, International Women’s Day, Women’s Rights

The goal was to select pages discussing high-profile topics to analyze shifts in language and discourse using topic modeling.

Revision History

Wikipedia has over 22 million articles collaboratively written by 77,000 active contributors, making it an interesting source for analyzing language changes over time [3]. This project uses both a subsample dataset (11 pages, 210 revisions) for preprocessing experiments and a complete dataset (388 pages, 6,200 revisions) with one revision per year.

Addressed later on, a deeper dive into Wikipedia revisions revealed challenges in processing my data. Wikipedia revision history includes spam content, inappropriate, vulgar, and unrelated revisions; most revisions involve minor updates like spelling or formatting corrections; and pages occasionally include other languages (e.g., German terms like Geschlechtsidentität for "gender identity"). To address these issues and focus on overarching trends in discourse, I retained only one revision per year for each page. This strategy provides a general overview of temporal changes while avoiding near-duplicate content and computational inefficiencies. While this approach overlooks short-term edit conflicts, such as "wiki wars," analyzing only year-by-year changes is sufficient for understanding how discourse surrounding controversial topics like gender, identity, and discrimination evolved over the 21st century.

Subsample Dataset, 11 Pages (210 revisions):

Includes pages such as Feminism, Gender Studies, Cancel Culture, and Intersectionality. Provides a focused dataset for validating preprocessing steps, representative of the complete dataset.

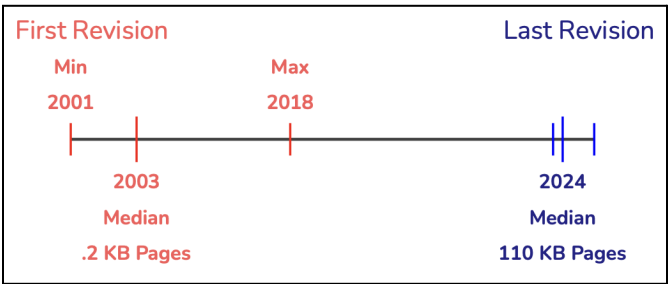


Fig. 2.2.a. Timeline of revision history for my 11 pages subsample, including the average size for the earliest revisions and latest.

The subsample dataset shows significant growth in content size from 0.2 KB (2001) to 110 KB (2024), reflecting the evolution of discourse over time. The staggered introduction of topics, like “Cancel Culture” (2018), highlights emerging concepts, which will be important to my analysis of shifting focus and language trends in key 21st-century issues.

Complete Dataset, 388 Pages (6200 revisions)

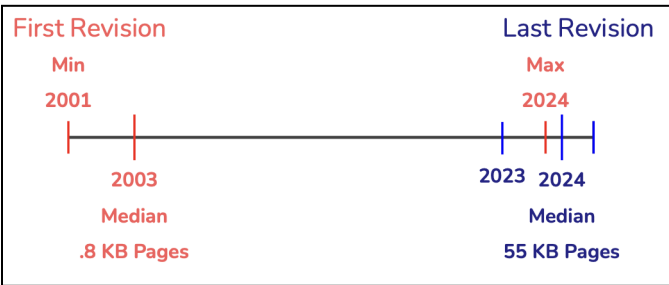


Fig. 2.2.b. Timeline of revision history for my 11 pages subsample, including the average size for the earliest revisions and latest.

The complete dataset (388 pages, 6,200 revisions) also shows steady content growth from 0.8 KB (2003 median) to 55 KB (2024 median). Early revisions contain minimal information, focusing on broad themes, while recent revisions provide detailed, nuanced content, improving topic modeling and coherence. The consistent activity in 2023–2024 highlights increased interaction with these pages, reflecting growing interest in controversial topics. This temporal content growth allows for identifying emerging themes and analyzing discourse evolution across key societal issues.

It is important to note, the dataset deliberately includes pages flagged by Wikipedia for bias, neutrality disputes, and controversy warnings.






	This article appears to be slanted towards recent events . Please try to keep recent events in historical perspective and add more content related to non-recent events. (October 2018)
	<p>This article has multiple issues. Please help improve it or discuss these issues on the [hide] talk page. (Learn how and when to remove these messages)</p> <ul style="list-style-type: none"> The neutrality of this article is disputed. (April 2020) This article may relate to a different subject or has undue weight on an aspect of the subject. (April 2020)
	<p>This article has multiple issues. Please help improve it or discuss these issues on the [hide] talk page. (Learn how and when to remove these messages)</p> <ul style="list-style-type: none"> This article's factual accuracy is disputed. (March 2019) This article's lead section may be too short to adequately summarize the key points. (March 2022)
	This section may lend undue weight to certain ideas, incidents, or controversies . Please help to create a more balanced presentation . Discuss and resolve this issue before removing this message. (August 2016)
	<p>This article has multiple issues. Please help improve it or discuss these issues on the [hide] talk page. (Learn how and when to remove these messages)</p> <ul style="list-style-type: none"> This article may lend undue weight to certain ideas, incidents, or controversies. (June 2022) This article possibly contains original research. (June 2022)

Fig. 2.3. Collection of wikipedia page warnings including page bias, neutrality disputes, and controversy warnings.

These pages are critical for studying how language and discourse evolve around controversial topics over time, aligning with the project’s focus on societal discourse in the 21st century.

Finally, after exploring my dataset by manually reviewing pages, though time-consuming, I learned valuable insights into the source data—a critical step in any data analysis project. The number of revisions per page, the earliest and latest revisions, and how page size of content evolved over time are crucial for interpreting the visualizations and ensuring accurate topic modeling, as page size and yearly variations will directly impact the analysis.

3. HISTORICAL TIMELINE

The timeline below highlights key 21st-century events shaping discourse on gender, identity, and discrimination. It is not complete, but I compiled it from sources such as Equaldex's LGBT Rights Timeline [4], Wikipedia's 21st Century Timeline [5], and Only Earthlings' Social Movements of the 21st Century [6]. These milestones provide perspective and will inform my analysis of how topics evolved over time using LDA topic modeling.

2001–2009: Early 21st Century

2001: 9/11; UN adopts *Millennium Development Goals*, prioritizing global gender equality.
 2003: Massachusetts legalizes same-sex marriage (first U.S. state).
 2005: *Intersectionality* gains prominence in feminist academic discourse.
 2006: *MeToo Movement* launched by Tarana Burke. 2008: California’s Proposition 8 bans same-sex marriage (later overturned).

2010–2016: Mid-21st Century

2010: Establishment of *UN Women*, spotlighting global gender issues.
 2013: U.S. Supreme Court strikes down *DOMA*, advancing LGBTQ+ rights.
 2014: *#BlackLivesMatter* emerges, emphasizing race and gender intersectionality.
 2015: *Obergefell v. Hodges* legalizes same-sex marriage nationwide.
 2016: Fourth-wave feminism focuses on digital activism, intersectionality, and sexual harassment.

2017–2020: Pre-COVID Era

2017: *Women’s March* becomes a global protest for women’s rights.
 2017: *MeToo Movement* resurges, igniting global discussions on sexual harassment.
 2018: *Toxic Masculinity* enters mainstream discourse, critiquing traditional male roles.
 2020: Global *Black Lives Matter* protests highlight systemic racism and intersectionality.

2021–2024: Post-COVID Era

2021: U.S. Supreme Court upholds LGBTQ+ protections under Title VII (*Bostock v. Clayton County*).
 2022: *Roe v. Wade* overturned, intensifying abortion rights debates.
 2023: Non-binary and transgender rights gain momentum amid legislative battles and advocacy. 2024: AI bias debates emerge, examining technology's role in systemic inequities.

4. PROJECT WORKFLOW

My project’s workflow followed an iterative process although the paper presents it linearly. Here is a summary of this workflow after many iterations.

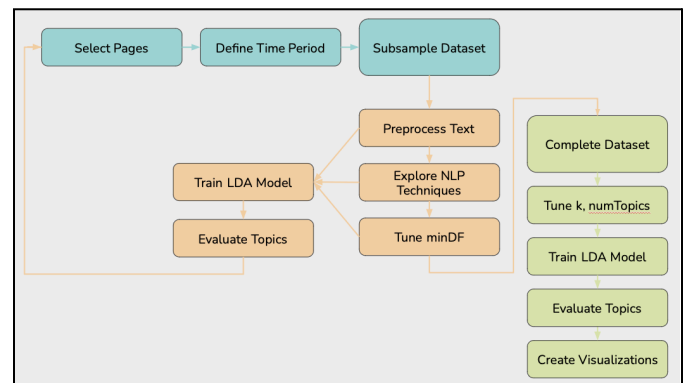


Fig. 4.1. Depiction of my project workflow. Note back arrows illustrating that this was an iterative process.

As already discussed, the process began with selecting Wikipedia pages focused on topics such as Gender, Feminism,

and Social Justice, followed by defining the time period of analysis from 2001 to 2024, opting for yearly granularity to balance detail and manageability. The dataset was then downsampled to include 11 key revisions, one per year.

Using this downsampled dataset for efficiency, my text preprocessing involved custom Python scripts and NLP pipelines, including regex-based text cleaning, tokenization, and term frequency vectorization. Various NLP techniques were explored and compared, with topics evaluated using metrics such as coherence, perplexity, and top words for interpretability. Parameter tuning was conducted by adjusting minDF in the CountVectorizer and re-evaluating performance. Using the complete dataset, an LDA model was trained with different number of topics. The final model incorporated my chosen “best” preprocessing pipeline and number of topics configuration. Visualizations were then created, including stacked bar charts to illustrate single-page trends and a heatmap showing topic evolution over time across all pages.

5. TOPIC MODELING

Latent Dirichlet Allocation (LDA) is a topic modeling technique used to identify hidden topics in a collection of documents. It works by analyzing word distributions to group words that frequently appear together into “topics.” My goal is to use LDA to visualize changes in topics over time and analyze the evolution of discourse on controversial subjects.

How LDA Works

After preprocessing and tokenization where text data is cleaned, tokenized, and processed using NLP techniques. My script uses CountVectorizer to convert tokens into a numerical representation by counting how often each word appears. Then uses IDF (Inverse Document Frequency) to weigh the words, giving less importance to common words and highlighting unique terms. Finally, LDA is applied to identify patterns in the weighted word distributions to generate topics—resulting in each topic represented as a list of keywords.

How My Project Uses LDA

LDA assumes a *bag-of-words* model, which ignores word order and context, see IBM’s LDA description [7]. This can be problematic when meanings evolve due to historical events. To address this, I group revisions by year, allowing me to track how terms and topics change over time while maintaining temporal alignment.

I will apply LDA on both the subsample dataset (11 pages, 210 revisions) for testing and the full dataset (388 pages, 6,200 revisions) for final analysis. When tuning preprocessing pipelines and LDA parameters, I will evaluate topics based on:

- A. Interpretability: Are the keywords in each topic meaningful and coherent?
- B. Fit to the Dataset: How well do the topics capture the nuances and changes in the dataset over time?

The focus is on overfitting the model to my dataset to capture trends specific to Wikipedia revisions, ensuring topics reflect meaningful changes in discourse without introducing external data.

Using LDA on Wikipedia revision data allows me to track the evolution of topics over time, uncovering shifts in societal discourse around controversial subjects. By analyzing revisions grouped by year, I can identify temporal patterns and emerging themes. To evaluate my LDA models, I will manually inspect the keywords for each topic alongside coherence scores and perplexity metrics.

Evaluation Technique: Coherence

The coherence score measures the semantic similarity between high-probability words within a topic. For this project, I use the *c_v* metric, which compares word co-occurrences against a reference dictionary (Gensim). This helps evaluate how “meaningful” or interpretable topics are. The *c_v* metric has several benefits such as it correlates well with human judgment and it captures semantic coherence, making it a useful marker for evaluating topics.

However, certain terms specific to my dataset, such as *cisnormativity* or *eco-feminism*, might not exist in the Gensim dictionary, potentially lowering the coherence score despite their relevance to topics on gender and feminism. Acknowledging this drawback, I will not be relying solely on coherence values to evaluate my models.

When tuning preprocessing pipelines and LDA parameters, I will prioritize topics with high coherence scores while conducting deeper analysis on the generated topics keeping in mind the limitations of the Gensim dictionary.

Evaluation Technique: Perplexity

The perplexity is a measure of how well a topic model predicts unseen data. Perplexity is calculated by evaluating the likelihood of the words in each document under the model, normalized by the total number of words in that document. This measure is then averaged across all documents in the dataset. Lower perplexity indicates that the model better explains the dataset, suggesting higher predictive performance. When evaluating preprocessing pipelines and tuning LDA parameters, I will generally aim for the lowest perplexity, as this reflects a model that is well-fitted to my data.

Since my goal is not generalization but capturing specific trends in discourse, perplexity alone will not dictate my choices. A lower perplexity may not always align with topics that are interpretable, so it will be used alongside coherence and manual inspection.

6. WIKIPEDIA PREPROCESSING

To optimize preprocessing steps, I evaluated using coherence scores (to assess topic interpretability), perplexity (to measure how well the model predicts unseen data), and manual inspection of generated keywords. For experimentation, I worked with a subsample of 11 Wikipedia pages, each with yearly revisions from around 2001 the earliest to around 2024, resulting in ~ 210 documents. The goal was not to generate perfectly coherent topics but to identify effective preprocessing strategies without overfitting to the smaller subset.

A smaller dataset reduces computational requirements, enabling faster iteration and testing of multiple preprocessing variations. As introduced in my data section, revisions spanning the full time range (2001–2024) maintain representativeness for testing preprocessing steps. The focus is on refining preprocessing pipelines, not final topic quality, where effective preprocessing steps identified here can be applied to the full dataset later. To avoid overfitting to the subset, I limited the topic model to 5 topics and used 30 max iterations with the aim of capturing overarching themes for validating preprocessing methods.

Wiki Specific Preprocessing

To find the best way to clean the text, I designed various preprocessing functions to remove Wikipedia-specific elements such as comments, references, and metadata while retaining meaningful content.

```

253 # Aggressive cleaning (removing Wiki-specific language)
254 def preprocess_all(text):
255     text = re.sub(r"", "", text) # Remove comments
256     text = re.sub(r"{{.*?}}", "", text) # Remove template content
257     text = re.sub(r"<ref.*?>.*?</ref>", "", text) # Remove references
258     text = re.sub(r"[[File:.*?]]", "", text) # Remove file links
259     text = re.sub(r"http(S+|www\.)S+", "URLS", text) # Replace URLs
260     text = re.sub(r"(?i)\bISBN\b", "", text) # Remove ISBN references
261     text = re.sub(r"(?i)\bREDIRECT\b", "", text) # Remove redirects
262     return html.unescape(text) # Decode HTML entities

```

Fig. 6.1. Function that includes all Regex based preprocessing techniques that I tried.

After applying different cleaning techniques, I found that removing frequent uninformative words like "edit," "page," and "article" helped improve topic clarity. Also, that using `html.unescape(text)` preserved meaningful terms like H  l  ne (from improperly decoded entities such as é and è), which would otherwise lower coherence scores.

Looking at my results for no preprocessing:

Keywords: transgender, lgbt, feminist, feminism, black, labor, culture, nametwsl, nametwsh, opportunityref, url

Perplexity: 9.32

Coherence Score: 0.665

and my results for preprocessing all:

Keywords: intersectionality, feminist, black, oppression, suffrage, nametwsl, accessdate, date, earnings, barrier

Perplexity: 8.72

Coherence Score: 0.664

Surprisingly, the "no preprocessing" approach yielded good coherence and relevant keywords despite the presence of noise (e.g., "url" and "opportunityref"). Preprocessing all achieved slightly lower perplexity (better generalization) with similar coherence score. However, terms like "nametwsl" and "accessdate" persisted, suggesting that the Wiki-specific preprocessing alone may not fully clean the data. Rather than aggressively filtering rare terms, I chose to retain them for now, as they may be addressed naturally by scaling to the full dataset.

Next, I will fine-tune my NLP preprocessing steps (e.g., tokenization, stopwords removal, and normalization) to eliminate any remaining noise while preserving meaningful terms. This iterative approach ensures the preprocessing pipeline is both effective and transferable when applied to the full dataset

7. NLP PREPROCESSING

The NLP preprocessing pipeline will be designed to handle the complexity of Wikipedia revision data, which includes varied language use, stylistic differences, and editing patterns. The descriptions of the Pyspark NLP included preprocessing steps can be found at JohnSnowLabs [8]. Using the output from the previous Wikipedia-specific preprocessing step, I tested different variations of each step in the NLP pipeline:

DocumentAssembler

The initial step in the pipeline, converting raw text into a format compatible with Spark NLP. For this step, I tested variations of document cleanup_modes: {shrink, disabled, inplace, each, delete full}.

Tokenizer

Splits and standardizes the text into individual tokens, handling punctuation and special characters. No variations were applied here.

Normalizer

Cleans tokens by applying text normalization techniques such as lowercasing and removing unwanted characters like

punctuation, numbers, and special symbols. I tested variations with and without lowercasing.

StopWordsCleaner

Removes common stop words, such as "the" and "and," using either built-in or custom stopword lists. I tested variations with and without stopword removal.

Stemmer

Reduces words to their root forms (e.g., "running" to "run") to consolidate word variations and simplify the dataset. I tested variations with and without stemming.

Lemmatizer

Applies linguistic rules to reduce words to their base forms (e.g., "better" to "good"). This step relies on pre-trained lemma dictionaries. I tested variations with and without lemmatization.

NGramGenerator

Generates n-grams, such as bigrams or trigrams, to capture contextual phrases that may add depth to the topic modeling. I tested variations with 1-grams, 2-grams, or 3-grams.

Finisher

Converts the processed annotations back into plain strings or arrays suitable for machine learning pipelines (LDA). No variations were applied here.

For this stage, I used parameters $k=5$, $\text{maxIter}=30$, $\text{minDF}=10$, and the subsampled dataset (11 pages, 210 revisions) for computational efficiency, as the focus is on testing preprocessing variations rather than final topic generation. Once again, each pipeline variation will be evaluated by generating topics using LDA and comparing the coherence scores, perplexity metrics, and output keywords. This process will help identify the most effective preprocessing pipeline for capturing meaningful topics.

8. NLP RESULTS

This section evaluates the impact of various preprocessing pipelines on my topic modeling results. The baseline pipeline (the most "standard" pipeline used to compare to my variations) applied lowercase conversion, stopword removal, and tokenization with no lemmatization, stemming, or n-grams, using the "shrink" cleanup mode. Keywords generated under this configuration included artifacts like *nametwsl*, *nametwsh*, and overly generic terms such as substantive and intersections. These results influenced me to try and find a better pipeline variation.

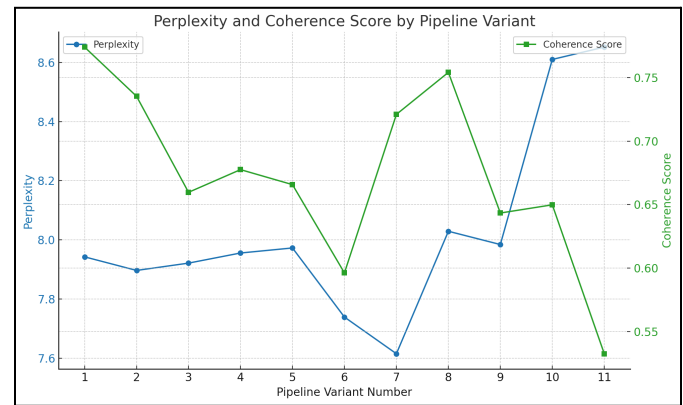


Fig. 8.1. A graph of coherence, perplexity for all relevant NLP pipeline variants. Pipeline Variants are 1: Baseline, 2-5: Document Cleanup Modes, 6: Stemmer, 7: Lemmatizer, 8: Case Sensitive, 9: No Stopword Removal, 10-11: 2-3 N-grams.

Document Cleanup Variants

When testing document cleanup modes (shrink, disabled, inplace, each, delete_full), distinct patterns emerged. The "disabled" mode retained broader domain-specific terms like patriarchy and thirdwave, but introduced noise with artifacts like *nametwsl*. The "inplace" mode improved representation of nuanced social issues, producing terms like transprejudice and misgendering, but failed to fully eliminate general artifacts. The "each" mode enhanced granularity, surfacing proper nouns such as *ettinger* and *butler*, which added contextual richness but increased sensitivity to individual names. The "delete_full" mode favored broader concepts like nationalism and intersections, but included artifacts like *natcult* due to its aggressive filtering.

From looking at the graph (pipeline variants 1-5 respectively), the "shrink" cleanup mode balances preprocessing efficiency and scalability while maintaining domain-specific terms critical to my analysis. It reduces noise without ignoring the nuanced language necessary for controversial topics.

Stemmer Vs Lemmatizer

Lemmatizer Keywords: *ettinger*, masculinity, psychoanalysis, nationalism, suffrage

Stemmer Keywords: intersect, ceil, bisexu, oppress, substant

Looking at the graph (variants 6 and 7 respectively), the stemmer performed worse than the lemmatizer with lower coherence and higher perplexity. Even though the choice may be clear, it is important for my project to take into account other factors as well. For instance, the lemmatizer relies on a pre-trained dictionary, which may fail to handle non-standard terms, abbreviations, or slang and performs better for small datasets where domain-specific interpretation is key. Whereas the stemmer truncates words to their root forms (e.g., "bisexu"), reducing variability and condensing similar terms. The stemmer may be sacrificing linguistic richness but scales

better to large datasets by reducing vocabulary size and noise. Stemming t is independent of pre-trained dictionaries, making it robust to non-standard or new terms.

Although stemming performed worse in coherence and perplexity, I chose it for its scalability to larger datasets. Stemming reduces vocabulary size and focuses on root-level clustering, which better supports high-level trend analysis. Its independence from pre-trained dictionaries ensures it handles diverse and evolving language, critical for my larger dataset spanning decades and including new lingo relevant to topics on gender and feminism. This trade-off prioritizes computational efficiency and scalability over human interpretability.

Case Sensitivity

Keywords: LGBT, Feminist, Americans, Glass, Intersectionality
Coherence Score: 0.754 (relatively high).

Introducing case sensitivity preserved term distinctions based on capitalization. These distinctions included proper usage of acronyms (e.g., LGBT) and proper nouns (e.g., Americans, Glass); variants like Intersectionality (title) versus intersectionality (concept); and a high coherence score suggesting that retaining capitalization improves clarity, especially for case-dependent terms.

While case sensitivity enhances semantic detail, it introduces redundancy and sparsity when capitalization differences are irrelevant, which is common in Wikipedia data. For my analyses focused on topic evolution, case-insensitive processing is preferable to me, as Wikipedia revisions lack consistent or meaningful capitalization patterns. This choice avoids overfitting and reduces complexity while maintaining focus on broader trends.

Without Stopword Removal

Keywords: movements, intersections, natcult, substantive, quote.

Disabling stopword removal resulted in broader, noisier topics. Inclusion of high-frequency terms like movements and quote diluted topic specificity. Artifacts like natcult and substantive added irrelevant terms, reducing clarity. Ambiguous words like intersections introduced multiple meanings, weakening thematic focus. Retaining stop words captures general patterns but reduces coherence and interpretability, making it less suitable for focused topic modeling.

N-gram Variants

Bigrams Keywords: ident polit, glass ceil, racial discrimin, gender pai, feminist movem

Terms like glass ceil and racial discrimin provide specific contextual pairs, enhancing topic granularity without sacrificing coherence (0.650). Keywords such as gender pai and feminist movem balance interpretability and detail, making bigrams effective for general topic modeling.

Trigrams Keywords: lesbian gai bisexu, jim crow law, gender pai gap, women right vote, equal opportun publish

Highly specific terms like lesbian gai bisexu and jim crow law deepen contextual insights but introduce sparsity and noise. Phrases like url url accessd and equal opportun publish highlight redundancy and artifacts, lowering coherence (0.533) and increasing perplexity. While trigrams offer richer phrases like gender pai gap and women right vote, they risk overfitting smaller datasets.

Purely out of exploration, I also tried using different combinations of NGrams, results not included in the graph.

Ngrams	P	C_v	Keywords
[1, 2]	10.713	0.365	movements as, gender class, intersectionality and, categoryracism, women feminist
[1, 2, 3]	11.424	0.439	gender politics, masculinity and, opportunity commission eeoc, umbrella term, misandry

Fig. 8.2. A table showing the perplexity (P), coherence score (C_v), and generated keywords for different combinations of Ngrams.

Testing [1,2]-grams captured paired phrases (e.g., gender class, intersectionality and) but introduced fragmented terms, yielding low coherence (0.365). Adding trigrams in [1,2,3]-grams improved descriptive detail (e.g., masculinity and, umbrella term), but increased perplexity (11.424) due to sparsity and redundancy (e.g., opportunity commission eeoc).

Unigrams (the default) are the most effective for this project, as they provide a clear and general representation of topics without introducing unnecessary complexity. While bigrams showed slightly higher coherence, they risk overfitting due to the sparsity of phrases in smaller revisions and add computational overhead when scaling to the larger dataset. Unigrams ensure consistent and interpretable results, making them better suited for analyzing broad topic changes over time across the full dataset.

From analyzing the results across pipeline variants, unigrams combined with stemming, the "shrink" cleanup mode, and case insensitivity emerge as the most effective configuration for my project. While lemmatization produced slightly higher coherence, stemming is better suited for scaling to my

complete dataset, as it reduces vocabulary size and handles new or evolving terms not captured by lemmatizer dictionaries. Bigrams offered improved contextual detail but risked overfitting smaller revisions and adding computational complexity. Ultimately, this NLP pipeline ensures a balance between interpretability, scalability, and capturing broad topic trends over time. Before scaling my pipeline to the complete dataset, I will test one final parameter on the subsampled dataset.

9. TUNING minDF

Minimum Document Frequency (minDF) determines the minimum number of documents a term must appear in to be included in the model. Adjusting minDF reduces the weight of rare terms, improving computational efficiency and enhancing interpretability by filtering out noise. However, for my project of analyzing topic changes over time, rare terms are essential, as they may highlight niche or emerging discourse, particularly in controversial topics spanning decades.

With this in mind, testing minDF on the subsampled dataset (11 pages, 210 revisions) demonstrates the trade-offs between coherence, perplexity, and topic granularity:

minDF	P	C_v	Keywords
3	8.155	0.497	mortal, anticategor, suffrag, natcult, bisexu
6	7.959	0.478	foucault, māori, blechner, griselda, patriarchi
11	7.658	0.512	misgend, jorgensen, transprejudic, substant
15	7.54	0.579	choctaw, segreg, antisemit, secondwav, suicid
21	7.293	0.599	antisemit, jew, intersex, wage, collin

Fig. 9.1. A table showing the perplexity (P), coherence score (C_v), and generated keywords for different minDF values.

Lower minDF values (e.g., 3, 6) retain rare and specific terms (e.g., mortal, māori), but result in sparse topics, increasing perplexity and lowering coherence due to noise. Higher minDF values (e.g., 15, 21) filter out rare terms, yielding more general and coherent topics (e.g., antisemit, wage) with improved coherence scores and reduced perplexity. While these results show the interpretive benefits of removing rare terms, tuning minDF demonstrates a limitation of hyperparameter optimization that improved metrics do not always align with the research goal.

When applying my pipeline to my complete dataset, I will not filter out rare terms using minDF. My goal is to analyze how discourse evolves over time, and rare terms often reflect niche, emerging, or context-specific language crucial to controversial topics. Removing these terms could obscure valuable insights, especially when analyzing historical changes in terminology and focus. As seen in my experiments, coherence and perplexity alone are not definitive markers of success; they must be balanced with interpretability and the research objective. With my experiments on the subsample dataset complete, I will now apply my finalized NLP pipeline to the complete dataset to analyze topic changes across 388 Wikipedia pages with revisions (6200 documents) over the 21st century.

10. TOPIC NUMBER OPTIMIZATION

To optimize the number of topics k for my final model, I applied my finalized NLP pipeline—Wikipedia-specific preprocessing with stemming—on the complete dataset (388 pages, 6,200 revisions spanning 2001–2024). I tested k values of 5, 7, 13, 16, and 20 with maxIter=100 to ensure the LDA model converged fully without premature stopping, especially when comparing results across different topic counts.

Fig. 10.1 illustrates the coherence scores and perplexity as k increases. Showing that coherence initially rises, peaking at $k=7$, before declining as topics become more fragmented. And that perplexity decreases with higher k , reflecting better statistical fits but often at the expense of interpretability.

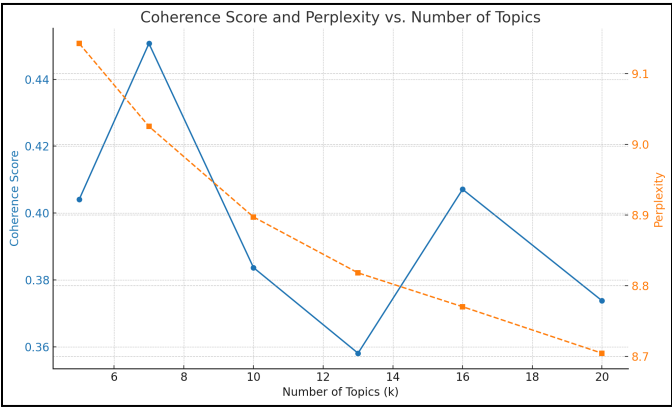


Fig. 10.1. Graph of coherence and perplexity vs number of topics for complete dataset 388 Pages (each with revisions 2001 earliest-2024 latest), 6200 documents in total.

Finding balance between coherence and perplexity is critical for my project. At $k=7$, coherence is maximized, suggesting semantically meaningful topics, while perplexity remains reasonably low, avoiding overfitting. Larger k values, such as 16 and 20, offer lower perplexity but reduced coherence, indicating overly granular or redundant topics that may

obscure broader trends. For my final model, I selected 7 topics to maintain interpretability and ensure meaningful insights into how discourse evolves across controversial topics over time. This choice aligns with the project's goals of visualizing high-level trends while balancing computational efficiency.

Next, I will use $k=7$ for all subsequent visualizations, including stacked bar charts and heatmaps, to analyze and illustrate topic evolution effectively. And I will keep $\text{maxIter}=100$ to ensure topic convergence, minimizing variability in results, sacrificing efficiency.

11. TOPIC EVOLUTION OVER TIME

Using $k=7$ number of topics and my preprocessing pipeline, I generated these topics:

- Topic 0: same-sex, harass, bulli, marriag, disabl, cyberbulli, polic, gai, drag, sexual
- Topic 1: same-sex, asexu, marriag, homosexu, gai, lgbt, rape, bisexu, lesbian, queer
- Topic 2: white, racial, race, black, racism, defam, nonviol, immigr, african, privileg
- Topic 3: antisemit, homosexu, gender, women, health, genet, intersex, therapi, sexual, eugen
- Topic 4: jew, antisemit, jewish, parti, islam, church, muslim, gerrymand, witch, hate
- Topic 5: women, abort, categorytreati, feminist, femin, steril, mascot, reproduct, sex, roe
- Topic 6: uyghur, milk, ethnic, bisexu, hrc, sectarian, cleans, xinjiang, muslim, chines

With $k=7$ topics, I visualized topic evolution across the dataset using stacked bar charts and a normalized heatmap to analyze changes in topic distributions over time, beyond just the dominant topic.

For example, in the *Gender equality* page (revisions from 2005–2024), the highest-distribution topic (Topic 5) remained consistent across years. However, changes in the relative proportions of other topics revealed subtle shifts in language and focus. This observation highlighted the need for visualizing full topic distributions, as analyzing only dominant topics may obscure meaningful trends.

The heatmap uses normalized prevalence, calculated as the sum of topic probabilities divided by the number of documents per year. This normalization addresses data imbalances, ensuring that years with more revisions do not dominate the visualization. The $k=7$ topics, their simplified representations, the stacked bar charts for *Gender equality* and

Intersectionality pages, and the heatmap for the complete dataset are following:

- Topic 0: Bullying
- Topic 1: LGBTQ+
- Topic 2: Race
- Topic 3: Health
- Topic 4: Religion
- Topic 5: Feminism
- Topic 6: Ethnicity

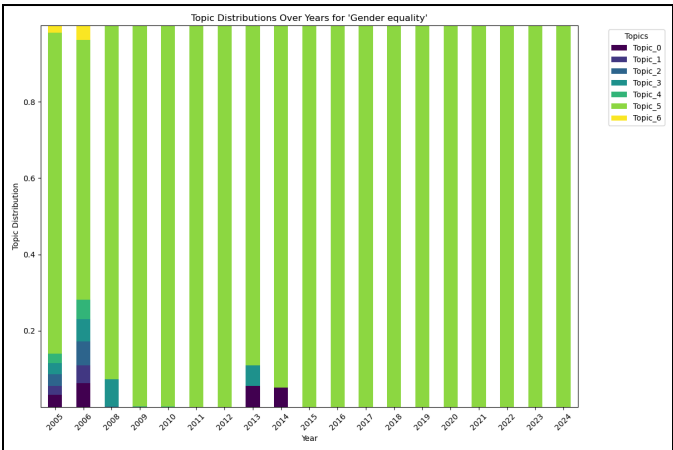


Fig. 11.1.a. Stacked bar chart for Topic Distributions over the years 2005–2024 for the *Gender equality* wikipedia page.

The *Gender equality* page shows a dominant Topic 5 (women, abortion, feminist) consistently from 2006 onward, reflecting a clear focus on feminist and reproductive rights discourse. While earlier years (2005–2006) show slight variations with other topics (e.g., Topics 0 and 6), these quickly diminish as the dominant topic solidifies. This persistence highlights the stability of feminist issues in gender equality discussions, aligning with my historical timeline such as the 2010 establishment of UN Women and 2016 fourth-wave feminism emphasizing reproductive rights and activism.

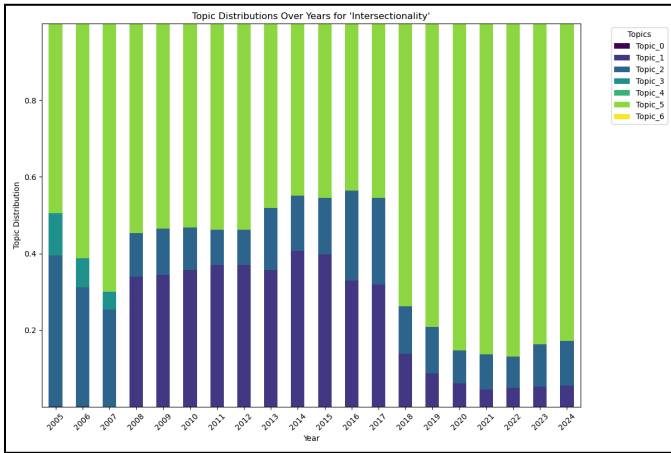


Fig. 11.1.b. Stacked bar chart for Topic Distributions over the years 2005-2024 for the *Intersectionality* wikipedia page.

In contrast, the *Intersectionality* page exhibits significant topic shifts over time. Early years (2005-2010) show diverse topic distributions, with Topics 2 (race, racial, black) and 1 (LGBTQ-related terms) contributing heavily. This reflects initial intersectionality discussions focusing on race and gender. Post-2010, there is a notable decline in Topic 1 (LGBTQ focus), while Topic 5 (feminism and reproductive rights) becomes increasingly dominant. These trends align with my historical timeline when considering real-world movements like #BlackLivesMatter (2014), MeToo’s resurgence (2017), and the recent evolving feminist discourse into mainstream digital activism.

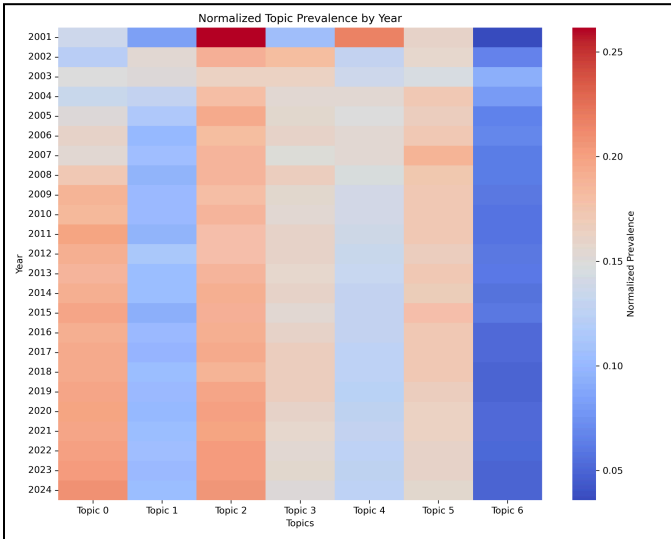


Fig. 11.2. Heatmap for the normalized topic prevalence over the years 2001-2024 by topic number across 388 wikipedia pages (6200 revisions).

The heatmap reveals overarching trends in topic prevalence across the full wikipedia dataset. Looking at the heatmap, Topic 2 (race-related terms) peaks from 2001-2010, aligning with events from my historical timeline like

#BlackLivesMatter and early discussions on systemic racism. Furthermore, Topic 5 (feminism and reproductive rights) consistently dominates post-2010, coinciding with *Obergefell v. Hodges* (2015) and *Roe v. Wade*’s overturn (2022). Other topics, like Topic 0 (harassment, cyberbullying), show minor yet notable increases post-2015, possibly reflecting growing discussions on digital harassment and social activism.

These visualizations add depth to the earlier historical timeline by quantifying discourse changes across Wikipedia revisions. While the timeline highlights major milestones like *Roe v. Wade* (2022) and *MeToo* (2017), the visualizations demonstrate how these events influenced topic prominence. For instance, Topic 2’s decline after 2010 aligns with shifts from racial to feminist-dominated discussions. Similarly, the stability of Topic 5 underscores the enduring significance of feminist and reproductive rights in online discourse. All in all, these figures provide a clear, data-driven view of how controversial topics evolve in public discourse.

12. CONCLUSION

This project taught me critical lessons in balancing overfitting versus generalization when analyzing topic changes over time. A significant portion of my effort was devoted to preparing and preprocessing the data, especially from Wikipedia dumps—a challenging task due to their unstructured nature and the nuances of natural language processing (NLP). In addition to my learnings from class, I learned the importance of establishing a strong baseline for comparing variations and the value of downsampling for debugging and testing scripts efficiently. I also learned the importance of ensuring the downsampled dataset was representative of the complete dataset.

Through this project, I was able to appreciate the critical role of visualizations in interpreting trends and communicating results effectively in big data analytics. While this project provides valuable insights into the evolving discourse on controversial topics, the results highlight the need for further experiments to draw stronger conclusions.

In conclusion, this project demonstrates the power of big data analytics and machine learning. Without these tools, manually analyzing the vast volume of Wikipedia revisions and text would have been impossible. By uncovering subtle shifts in topics over time, this analysis highlights a growing inclusivity and intersectionality in discourse—from gender equality to racial justice. These findings reflect broader societal changes and affirm the importance of data-driven methods in understanding the evolution of online discourse and its societal impact.

REFERENCES

- [1] Schmidt, J., Wiegand, M.: A survey on the role of Wikipedia for natural language processing, <https://journals.sagepub.com/doi/10.1177/0894439319828012>, last accessed 2024/12/16
- [2] Wikimedia Foundation: Wikimedia API, <https://en.wikipedia.org/w/api.php>, last accessed 2024/12/16
- [3] Wikimedia Statistics: Wikipedians Edit Count Tables, <https://stats.wikimedia.org/EN/TablesWikipediansEditsGt5.htm>, last accessed 2024/12/16
- [4] EqualDex: Timeline of LGBT rights (2000s), <https://www.equaldex.com/timeline/2000s>, last accessed 2024/12/16
- [5] Wikipedia Contributors: Timeline of the 21st century, https://en.wikipedia.org/wiki/Timeline_of_the_21st_century, last accessed 2024/12/16
- [6] Only Earthlings: Social movements of the 21st century, <https://onlyearthlings.com/14-social-movements-of-the-21st-century>, last accessed 2024/12/16
- [7] IBM: Understanding Latent Dirichlet Allocation, <https://www.ibm.com/think/topics/latent-dirichlet-allocation>, last accessed 2024/12/16
- [8] John Snow Labs: Spark NLP Documentation, <https://github.com/JohnSnowLabs/spark-nlp>, last accessed 2024/12/16

My data and PySpark scripts can be found at <https://github.com/demasiCodes/BigDataWikiTopicChange>.