An Analysis of US Legislative Focus Through Public Comment Text

Puran Dou
Yuyan He
Georgetown University, USA

Public comments submitted during U.S. federal rulemaking offer insight into stakeholder participation but are often dominated by repetitive, template-based submissions. This project applies text-as-data methods to distinguish semantic diversity from volume and to examine thematic patterns and potential advocacy tendencies in public comments related to energy and healthcare regulations. Using SBERT embeddings and cosine similarity, we identify semantically repetitive comments and retain one representative from each cluster while preserving all unique submissions.

The resulting corpus is analyzed using Latent Dirichlet Allocation (LDA) to compare topic structures across unique and repetitive comments. We further apply the Wordfish model to explore whether comment language reflects ideological or advocacy-oriented positioning, adopting different methodologies for energy and healthcare dockets given their institutional differences. Topic modeling reveals clear differences in emphasis between repetitive and unique comments, particularly in healthcare regulations. In contrast, Wordfish identifies limited ideological variation in healthcare comments, reflecting their technical and docket-specific nature. Overall, the findings demonstrate how preprocessing choices and policy context shape what text-based methods can reveal about public participation in regulatory processes.

Keywords: public comments; text as data; SBERT; LDA; Wordfish; energy; healthcare

I. Introduction

Public comments submitted during federal rulemaking are one of the most widely used channels through which stakeholders attempt to shape how policies are implemented; policymakers also use these submissions to refine proposed rules. Yet comment volume can be driven by coordinated mass campaigns that generate near-duplicate or template-based submissions. Nevertheless, even when many comments are repetitive, a docket may still contain meaningful disagreement and substantive discussion. This makes it essential to extract the "signal" from a large corpus which might be impractical to review comment-by-comment. This project uses text-as-data methods to distinguish variation from volume, characterize what commenters talk about, and assess whether their language reveals an underlying advocacy structure.

A growing methods literature emphasizes that naive deduplication can miss mass participation because sophisticated campaigns often paraphrase rather than copy and paste. Prior work therefore proposes embedding-based approaches that (i) encode each comment into a semantic vector, (ii) measure cosine similarity, and (iii) use density-based clustering algorithms (DBSCAN) to group near-duplicate or template-like submissions (Federal Chief Data Officer Council. 2021). By identifying whether the comments share similar information and narratives using word embedding, we can have a more comprehensive understanding of the commenters, instead of only looking at the same broad themes or exact features which don't necessarily convey meanings (Waight et al., 2025). More recent work highlights a key issue in long-form text analysis: many comments exceed transformer token limits, so default truncation can distort embeddings. Sentence or block-level aggregation methods provide lightweight alternatives that preserve more information from long texts while remaining computationally feasible for large corpora (Ortakci et al., 2025). Building on these ideas, we combine embedding-based clustering, which is used to identify repetitive vs. unique content, with topic modeling and Wordfish scaling, which are used to summarize substantive themes and to test for latent advocacy axis, respectively.

Concretely, we ask three questions in this research:

1. Do these dockets exhibit mass-campaign clustering, and how concentrated are the clusters?
2. After de-duplication, what topics dominate each corpus?
3. Does language-based scaling separate comments or actors along a meaningful dimension, and does this differ between contested and more harmonized dockets?

The report proceeds as follows. Section II describes how we collected and cleaned comment-level data from Regulations.gov via the API, defines the unit of observation (a single comment), and summarizes the key variables and text construction. Section III presents our analysis pipeline: (1) SBERT-style sentence-level embeddings with DBSCAN clustering to identify repetitive submissions and construct a de-clustered corpus; (2) LDA topic modeling to summarize thematic structure; and (3) Wordfish scaling to recover latent advocacy dimensions. Section IV reports results for the Env/Energy (IRS) and Social/Medical (CMS) corpora, highlighting how cluster concentration, topic diversity, and ideological separation vary across domains and across specific dockets. Section V concludes with the project's contributions, limitations, and extensions.

II. Data and Methods

Data source and Unit of observation

All of our data comes from public comments submitted to U.S. federal rulemakings on Regulations.gov. We collected these comments through the Regulations.gov API (v4) (https://open.gsa.gov/api/regulationsgov/), which lets us pull both (1) docket-level information (e.g., docket IDs and basic metadata) and (2) comment-level records, including the comment text, basic commenter information (such as organization name and location when provided), and any attached documents.

Our project centers around the Inflation Reduction Act (IRA), and we built two separate comment collections:

(i) Energy & Environment. This subset focuses on IRS rulemakings where commenters discuss clean energy tax incentives and how those incentives interact with Prevailing Wage and Apprenticeship (PWA) requirements. In the raw scrape, this dataset spans 17 IRS dockets and totals about 28,078 comment records (See Figure 1). Docket

IRS-2023-0066 was skimmed from 23,501 to 735 to match the Healthcare & Social Programs total number of comments. The final count yields 5,330 entries.

(ii) Healthcare & Social Programs. This subset consists of comments on Centers for Medicare & Medicaid Services (CMS) proposed rule for the Calendar Year 2025 Medicare Physician Fee Schedule and related policies in a single CMS docket (CMS-2024-0256-0045), with 6,101 comments in the raw dataset.

The unit of observation is a single public comment. Each row corresponds to one comment (identified by commentId) submitted under a specific docket (identified by docketId).

Variables of interest

To keep the analysis consistent from start to finish, we focus on three sets of fields: (1) identifiers and rulemaking metadata, (2) basic commenter information, and (3) the comment text itself. First, the identifiers and metadata let us reliably place each comment inside the broader rulemaking process—every record includes a comment ID, the docket ID it belongs to, the agency ID, and the related document ID, along with practical "paper trail" fields such as the posted date, received date, title, and a tracking number. Second, when available, we kept the basic information about who was speaking: the commenter's organization name, first and last name, and coarse location fields (city, state/province/region, and country). Finally, because the substance of many submissions lives either in the web form or in attachments, we were able to store multiple versions of the text: the raw HTML comment text returned by the API, a cleaned version of that text, the text extracted from PDF attachments when present, and a single combined text field that merges the cleaned comment text with the extracted PDF text so that each record has one analysis-ready body of content.

Data wrangling

To make the API output usable for analysis, we built a Python pipeline that produces a clean, consistent comment corpus across dockets.

We first scraped comments docket-by-docket through the Regulations.gov API, using pagination to ensure full coverage and basic rate-limit handling so large dockets can be collected reliably. Because attachment information is not always complete in the listing response, we then fetch the full comment record for each comment, including any attachment metadata. Next, we standardized the text content: we cleaned the HTML form comment body into plain text, extracted text from PDF attachments when available, and merged both into a single combined text field so each record has one analysis-ready version of its substantive content. Finally, we exported the results in a fixed schema (CSV/JSON), for later merging across dockets and all other downstream methods.

III. Analysis

1. Embedding, calculating similarity, and deduplicating

The pipeline of the analysis was inspired by a federal comment analysis toolkit developed in 2021 (https://github.com/kenambrose-GSA/CDO-Council-Public-Comment-Analysis-Project). They argued that exact-match deduplication was not adequate enough in terms of detecting mass campaigns, due to the fact that those who utilize such approaches often paraphrase the repetitive comments. Their proposed fix was rather straightforward: encode each comment into an embedding, measure pairwise similarity with cosine similarity, and then either (i) flag near-duplicates using a similarity cutoff or (ii) group similar comments using a density-based clustering algorithm such as DBSCAN.

We actively took on and adapted their approach with the latest research. Ortakci and Borhan (2025) directly tested two lightweight alternatives: Sentence-Level (SL) and Block-Level (BL) document embeddings, built on top of standard SBERT-family sentence transformers. Their key point was exactly the problem we faced in Regulations.gov: once a comment exceeds a model's max token limit, the "default" approach effectively truncates the rest of the text, which can quietly distort the embedding. Instead of truncating, SL splits a document into sentences and aggregates sentence embeddings into a single comment embedding; BL does a similar idea, but using token-length chunks. Empirically, they found SL/BL generally outperform the default truncation-based baseline, and the resulting clustering performance is competitive with long-text transformers like Longformer/BigBird—while SL tends to outperform BL.

So, rather than copying the toolkit's "MPNet vs. BigBird" choice directly, we used these findings to select a practical alternative for IRA comments: Sentence-Level (SL) embeddings built with a sentence-transformer model (all-mpnet-base-v2). This helps keep the embedding step lightweight while avoiding truncation problems for long submissions.

In implementation, we split each comment into sentences, embedded the sentences with all-mpnet-base-v2, averaged the sentence vectors into a single comment-level embedding, and then computed cosine similarities and run DBSCAN to identify dense clusters (mass-campaign or template-like submissions) versus low-density unique comments. Under this setup, DBSCAN assigned unique/unclustered comments to cluster -1, while clustered comments were assigned integer cluster labels accordingly.

For later topic modeling, we used the DBSCAN results to de-cluster the corpus: we kept all comments labeled -1 (unique/unclustered), and for each other cluster we selected a single representative "signature" comment, typically the one with the richest metadata in other fields (e.g., organization name and geographic information).

2. Topic Modeling

We applied topic modeling, specifically Latent Dirichlet Allocation (LDA), to analyze the substantive topics covered in the public comments. As discussed earlier, prior to topic modeling, we used the SBERT model to generate semantic embeddings for each comment and calculated their proximity in embedding space. This process allowed us to distinguish comments that were semantically repetitive from what were unique. For comments identified as repetitive, we retained one representative comment from each cluster, while all unique comments were preserved. These two components together formed the dataset used for the LDA analysis.

Before implementing the LDA model, we tested different values of the number of K by comparing the log-likelihood across different K. Based on this evaluation, we identified K = 30 as a robust choice for both the energy-related and healthcare-related datasets (See Figure 2)

We then ran the LDA topic model on the joint dataset consisting of the unique comments and the representative repetitive comments for the energy dataset, and ran the unique comments separately for the healthcare dataset. The different approach comes from the distinct data nature of the two datasets: the energy dataset consists of more than 10 different dockets, and the comments do not appear to be highly repetitive, while the healthcare dataset is only from one docket and similar template comments still exist after de-duplication. It makes sense to check the topics and content separately for the healthcare dataset, otherwise the unique component would take up most of the topic results; the energy dataset however, does not have this issue. For the repetitive comments in the healthcare dataset, since the number of repetitive comments was substantially reduced after de-duplication, we were able to compare the topics generated by the LDA model with the observed content of the repetitive comments.

Additionally, when organizational information was available, we examined the background of the commenters to assess whether certain types of stakeholders were more likely to contribute repetitive or unique submissions.

3. Wordfish

We next applied the Wordfish model to explore whether the public comments exhibit any ideological tendency. Similar to the topic modeling analysis, this step used the combined dataset consisting of all unique comments and the representative comments from repetitive clusters. The specific implementation of Wordfish differed across regulation dockets in the energy and healthcare domains.

For the energy-related dockets, we used Wordfish to test whether comment language contains a latent “advocacy axis” consistent with a Bootlegger–and–Baptist coalition structure. The idea is that if firms and climate-advocacy groups systematically emphasize different claims, risks, and priorities, a scaling model should recover those differences from word choice. We focused on two energy dockets, IRS-2023-0066 (clean hydrogen 45V credits, highly contested) and IRS-2023-0042 (PWA bonus-credit requirements, more harmonized). We kept only comments with non-empty organization names and assigned “true coalition” labels, classifying major climate NGOs as Baptists and energy or industry firms as Bootleggers, with additional buckets such as unknown, other, or hybrid. After standard preprocessing, we built a document–term matrix and ran an unanchored Wordfish model to get initial document positions (θ) and inspect extreme-loading terms. We then used the “true coalition” labels to select anchors: the most “Baptist-like” document (the minimum-θ document among labeled Baptists) and the most “Bootlegger-like” document (the maximum-θ document among labeled Bootleggers). We re-estimated Wordfish with these anchors, then checked whether θ separates the two coalitions using distribution plots and simple statistical tests. As the last step, we projected unions, hybrid groups, and other organizations onto the same axis to see where they fall.

In contrast, there’s no similar opposing structure for healthcare docket comments. In addition, the healthcare dataset contains several hundred comments, which made it difficult to identify reasonable anchors at the outset. To address this issue, we applied Wordfish to the full set of healthcare comments using a two-step procedure. The first run was used to identify potential anchor documents by examining extreme positions, while the second run generated the final scaled document positions.

In total, the Wordfish analysis included 516 healthcare comments and 41 plus 57 organization-submitted comments from two energy dockets.

IV. Results

1. Env/energy

1.1 After DBSCAN result

To make the Env/Energy corpus comparable to the Social/Medical corpus (≈5,000 comments), we reduced the Env/Energy raw dataset from 28,078 comments to roughly the same scale using the embedding + DBSCAN declustering process.

We started with the largest docket, IRS-2023-0066 (23,501 comments), and ran the embedding–DBSCAN process with a 0.95 similarity threshold, and declustered it down to 753 comments by keeping all "-1" (unique/unclustered) comments and keeping one representative comment per cluster for the remaining clustered submissions. We then combined this declustered IRS-2023-0066 subset with the other Env/Energy dockets, resulting in 5,330 comments in total.

We ran embedding and DBSCAN again on this preprocessed pooled corpus and obtained 32 clusters, with -1 (unique comments) making up the largest share. A quick look at the DBSCAN cluster distribution (top clusters by count) shows the pattern that most comments are unique (57%), and the clustered portion (43%) is highly concentrated: the largest cluster alone (cluster 15) accounts for 63% of all clustered comments, with a long tail of small clusters. (See Figure 3)

1.2 Topic result

Figure 4 compares each docket's share of comments in the topic-modeling corpus (Comment Perc) with its share of topic representatives (Topic Perc, i.e., how many of the 30 topic representatives come from that docket). From the comparison, we can see whether a docket is over-represented in topics or under-represented.

A clear example of high topic representation is IRS-2023-0066 (clean hydrogen, 45V). Although it accounts for only 14.13% of the final corpus after preprocessing (753 out of 5,330), it contributes 26.67% of topic representatives (8 out of 30). This pattern is consistent with what we observed qualitatively: the docket contains active debate activity and multiple stakeholder voices, which translates into more distinct thematic clusters in the text. Several mid-sized dockets also show a similar "over-representation" pattern: e.g., IRS-2022-0021 (48C/45X) and IRS-2022-0025 (Notice 2022-51), suggesting richer topic diversity relative to their size.

In contrast, IRS-2024-0026 (45Y/48E clean electricity) is strongly under-represented in topic representatives: it makes up 35.91% of the corpus (1,914 comments) but contributes only 6.67% of topic representatives (2 out of 30). This gap suggests that the docket's large volume is driven by highly concentrated or repetitive content, so additional comments add scale but not many new "themes" from the topic model's perspective. A similar, though smaller, under-representation appears in dockets like IRS-2023-0054, which contribute a meaningful share of comments but only one topic representative.

Finally, several very small dockets have zero topic representatives. This does not mean they contain "no themes"; rather, given their small size (and the fact that their language often overlaps with larger dockets), they are simply not selected as the representative documents for any of the 30 topics.

One important takeaway is that preprocessing for IRS-2023-0066 did not shrink away the substantive diversity. In the raw data, IRS-2023-0066 dominates the dataset (86.76% of original comments), but after embedding and DBSCAN de-clustering, it was reduced to 753 comments and became only 14.13% of the topic-modeling corpus. Despite this downsampling, it still produces 8 topic representatives, which is more than any other docket. This potentially suggests the declustering step mainly removed repetitive template submissions while preserving (and effectively highlighting) the genuinely diverse, debate-driven content that carries distinct policy arguments.

Across the Env/Energy corpus, the LDA topics are largely interpretable and docket-relevant (See Figure 5). For IRS-2023-0066, the eight topics capture the core 45V debate over how to define and operationalize "clean" hydrogen and reflect stakeholders' main concerns, but the presence of noise topics such as boilerplate and fragmented tokens suggests stronger preprocessing is needed. For IRS-2022-0021, the topics align with the rule's direction on domestic manufacturing and supply chains, energy-community eligibility, and community clean-energy projects, showing commenters' focus on eligibility, geographic targeting, and local delivery. For IRS-2022-0025, the model highlights two clear areas of interest, thermal energy systems and storage and CCS, suggesting that commenters use the notice to discuss how wage and bonus rules will interact with concrete technologies.

1.3 Wordfish result

IRS-2023-0066: 45V Clean Hydrogen

IRS-2023-0042: Tax Credit + PWA (See Figure 5)

In this setting, we use "Baptists" to refer mainly to climate- and public-interest organizations that argue from environmental integrity, justice, or community-protection frames, and "Bootleggers" to refer to firms and trade associations with direct financial stakes in IRA tax credits and flexibility in compliance rules.

In the more contentious 45V hydrogen docket (IRS-2023-0066), the Wordfish scale reflects this divide: industry and trade groups cluster on the Bootlegger side of the axis, while advocacy groups sit toward the Baptist side. When we project "other" organizations and individual commenters onto the same axis, most of them also fall closer to the Baptist side, suggesting that the Baptist organizations are broadly aligned with the language and positions of unaffiliated commenters in this highly contested docket.

In the relatively more harmonized PWA bonus-credit docket (IRS-2023-0042), the Wordfish positions are much more compressed, and the Baptist/Bootlegger split is harder to detect, which is consistent with a docket where the basic direction of the rule is less in dispute. Union and hybrid organizations tend to sit closer to the Baptist side, and Wordfish does a poor job separating these groups from industry Bootleggers. However, once we map individual commenters (labeled as "other") onto the same axis, a distinct cluster appears at the opposite end from the organized groups, closer to the Bootlegger side. This pattern suggests that many individual commenters, often writing from the perspective of apprentices whose livelihood is directly affected, use more traditional Bootlegger-style language that emphasizes their own earnings and job security, rather than program-level or fighting for higher-level collective benefits.

2. Medical/Healthcare

2.1 After DBSCAN, result

Coming from one docket, a large number of highly duplicated comments were detected in the CMS-2024-0256-0045 healthcare docket, using the embedding + DBSCAN declustering process (Figure 7). More than 4,000 comments were identified into one single cluster; after keeping one representative in all the repetitive clusters, we got 15 comments. Combined with 501 unique comments in the "-1" cluster, in total 516 comments were reserved after the de-duplication.

Take a quick look at the unique comments and the reserved repetitive comments, we may observe that the unique comments tend to be elaborative and extensive, sometimes pages long, while template like comments still exist.

2.2 Topic result

After running LDA on the unique comments and close observation over the repetitive representatives, we have detected a clear pattern that both unique and repetitive comments address similar issues related to the physician fee schedule, but their substantive emphases differ. The repetitive comments are heavily concentrated on payment-related concerns such as physician's income and reimbursement (e.g. CMS-2024-0256-5510, CMS-2024-0256-4044 etc.), which are central to service providers. After looking into the organizational information, we noticed that many of these comments are primarily submitted by organized interest groups, including organizations such as the National Academies of Practice and the National Psoriasis Foundation. In addition, among the repetitive representatives, we identified clear template-style texts, which further underscores the organized nature of stakeholder participation in the comment process.

In contrast, the unique comments place greater emphasis on insurance clause changes and the potential implications of these changes for patients, treatment options, and overall costs (e.g. CMS-2024-0256-1498, CMS-2024-0256-6984, etc.). Judging from their organizational information and their self identification, these submissions are more likely to come from medical research companies, laboratory groups, and individual clinician-researchers.

2.3 Wordfish result

For the healthcare comments, the Wordfish analysis did not reveal strong ideological differences across submissions (Figure 8). Although we selected a left-leaning anchor, a comment advocating for canceling copays to protect patients (CMS-2024-0256-5610), and a right-leaning anchor, a comment opposing the government's proposed work values (CMS-2024-0256-5769), the resulting distribution of document positions remained largely centered. Most comments cluster closely around the middle of the scale, indicating limited separation along an ideological dimension.

This outcome is consistent with the nature of the dataset. All comments respond to the same regulatory docket and generally advocate for similar outcomes. As a result, ideological variation across comments is minimal. Moreover, the comments are predominantly technical and detail-oriented, which further constrains the ideological space that the Wordfish model is able to detect.

V. Discussion

Main results

In the Env/Energy (IRS) corpus, the main results are as follows. First, embedding + DBSCAN successfully separates volume from variety: most comments are unique, while the repetitive share is highly concentrated in a small number of clusters. Second, the LDA topics are policy-relevant and docket-specific: The 45V clean hydrogen docket (IRS-2023-0066) remains disproportionately topic-rich even after declustering, and the recovered topics track core disputes while highlighting commenters' main points of interest. Third, Wordfish reveals an advocacy axis only where the docket comments present heated debates: in IRS-2023-0066, industry and public-interest organizations show a clear but somewhat unstable separation along a Bootlegger–Baptist dimension, whereas the more harmonized PWA docket (IRS-2023-0042) lacks such meaningful separation.

As for the Healthcare (CMS) corpus, we have the following results: first, there are large sums of repetitive comments in the single healthcare docket, which lead to a considerate decrease of total comments after the embedding + DBSCAN process. Second, as a result of the LDA, unique comments and repetitive representatives have distinguishable content emphasis and background, where unique comments are patient focused and more likely to be initiated by individuals and academic institutes, and repetitive comments are practitioner focused and more likely to be initiated by industry interest groups. Third, given the single docket structure, we failed to detect obvious ideological differences among comments.

Main Contributions

This project makes three main contributions. Methodologically, we build a practical text-as-data workflow for federal rulemaking comments that goes beyond exact-match deduplication. By combining sentence-level SBERT embeddings with DBSCAN clustering, which is different from the government's current approach of using word-level BigBird embedding, we provide a scalable way to distinguish "volume" from "variety" in large dockets where mass campaigns are common. Substantively, applying the same pipeline to two policy domains (IRA energy tax-credit implementation vs. Medicare fee schedule rulemaking) shows how comment processes can differ in structure and in analysis results: the energy corpus retains substantial thematic diversity after declustering, while the healthcare corpus is dominated by highly concentrated template clusters, and these differences shape what topic models and scaling models can recover. Analytically, we demonstrate when and where a latent advocacy dimension is detectable: Wordfish yields clearer coalition-aligned separation in contested energy dockets but compresses in more harmonized settings, highlighting the conditions under which ideological scaling of regulatory comments is informative.

Next Steps

Moving forward, several extensions could further strengthen the analysis and provide additional insights. One promising direction is to apply the Wordfish model across multiple years of healthcare docket comments. The healthcare regulation is renewed annually and receives a large volume of public comments each year. Examining comments across different periods may introduce broader variation in content, which could make ideological patterns easier to detect than in a single-docket, single-year setting.

Another valuable extension would be to compare our current results with public comments related to the One Big Beautiful Bill (OBBB) Act. Such a comparison is meaningful because the OBBB Act and the regulations examined in this project share common policy focuses on energy and healthcare, while pursuing opposing policy goals. This contrast provides an interesting opportunity to examine whether topic modeling and Wordfish results differ systematically when policy objectives diverge, potentially revealing clearer thematic or ideological distinctions across comment sets.

In addition, further verification is needed regarding the impact of comment condensation using word embeddings on document representativeness. As aforementioned, de-duplicated IRS-2023-006 accounted for 27% of the topics generated by LDA, even though they represented only about 14% of the original comments. Based on this result, we suspect that using word embeddings to remove duplication does not substantially diminish document representativeness. However, additional testing is needed to confirm this conclusion more rigorously.

Finally, in the current analysis, we did not combine the energy-related and healthcare-related comment datasets, largely because the topics covered in these two domains differ substantially. Combining them could lead to unclear or unstable results in both LDA and Wordfish. Moving forward, a more targeted cross-reference analysis within the energy and healthcare datasets may be useful, allowing us to explore whether any meaningful patterns emerge when the two domains are examined in relation to each other rather than fully merged.

References

Federal Chief Data Officer Council. 2021. Implementing Federal-Wide Comment Analysis Tools. resources.data.gov. Accessed December 17, 2025. https://resources.data.gov/resources/cdoc_comment_analysis/.

Ortakci, Y., & Borhan, B. (2025). Optimizing SBERT for long text clustering: two novel approaches with empirical insights. The Journal of Supercomputing, 81(8). https://doi.org/10.1007/s11227-025-07414-4

Waight, H., Messing, S., Shirikov, A., Roberts, M. E., Nagler, J., Greenfield, J., Brown, M. A., Aslett, K., & Tucker, J. A. (2025). Quantifying Narrative Similarity Across Languages. Sociological Methods &Amp; Research, 54(3), 933-983. https://doi.org/10.1177/00491241251340080

Appendix

Figure 1

Docket ID	Docket (Official Name)	# Comments (Rows)
IRS-2023-0042	Increased Credit or Deduction Amounts for Satisfying Certain Prevailing Wage and Registered Apprenticeship Requirements (REG-100908-23)	344
IRS-2022-0025	Request for Comments on Prevailing Wage, Apprenticeship, Domestic Content, and Energy Communities Requirements Under the Act Commonly Known as the Inflation Reduction Act of 2022 (Notice 2022-51)	304
IRS-2022-0021	Request for Comments on Energy Security Tax Credits for Manufacturing Under Sections 48C and 45X (Notice 2022-47)	293
IRS-2023-0063	Section 45X Advanced Manufacturing Production Credit (REG-107423-23)	193
IRS-2023-0066	Section 45V Credit for Production of Clean Hydrogen; Section 48(a)(15) Election To Treat Clean Hydrogen Production Facilities as Energy Property (REG-117631-23)	23,501
IRS-	Guidance on Clean Electricity Low-Income Communities Bonus Credit	47

2024- 0045	Amount Program (REG-108920-24)	
IRS-2024-0026	Section 45Y Clean Electricity Production Credit and Section 48E Clean Electricity Investment Credit (REG-119283-23)	1,914
IRS-2025-0002	Section 45Z Clean Fuel Production Credit; Request for Public Comments (Notice 2025-10) / Section 45Z Clean Fuel Production Credit; Emissions Rates; Request for Comments (Notice 2025-11)	286
IRS-2022-0028	Request for Comments on the Credit for Carbon Oxide Sequestration (Notice 2022-57)	122
IRS-2024-0049	Section 30C Alternative Fuel Vehicle Refueling Property Credit (REG-118269-23)	38
IRS-2023-0054	Definition of Energy Property and Rules Applicable to the Energy Credit (REG-132569-17)	347
IRS-2023-0029	Section 6417 Elective Payment of Applicable Credits (REG-101607-23)	154
IRS-2023-0028	Section 6418 Transfer of Certain Credits (REG-101610-23)	81
IRS-2023-0014	Energy Community Bonus Credit Amounts under the Inflation Reduction Act of 2022 (Notice 2023-29)	25
IRS-2020-0013	Credit for Carbon Oxide Sequestration (REG-112339-19)	85
IRS-2022-0029	Request for Comments on Credits for Clean Hydrogen and Clean Fuel Production (Notice 2022-58)	258
IRS-2023-	Additional Guidance on Low-Income Communities Bonus Credit Program (REG-110412-23)	77

0025		
IRS-2024-0023	Domestic Content Bonus Credit Amounts under the Inflation Reduction Act of 2022: Expansion of Applicable Projects for Safe Harbor in Notice 2023-38 and New Elective Safe Harbor to Determine Cost Percentages for Applicable Percentage Rule (Notice 2024-41)	86

Figure 2

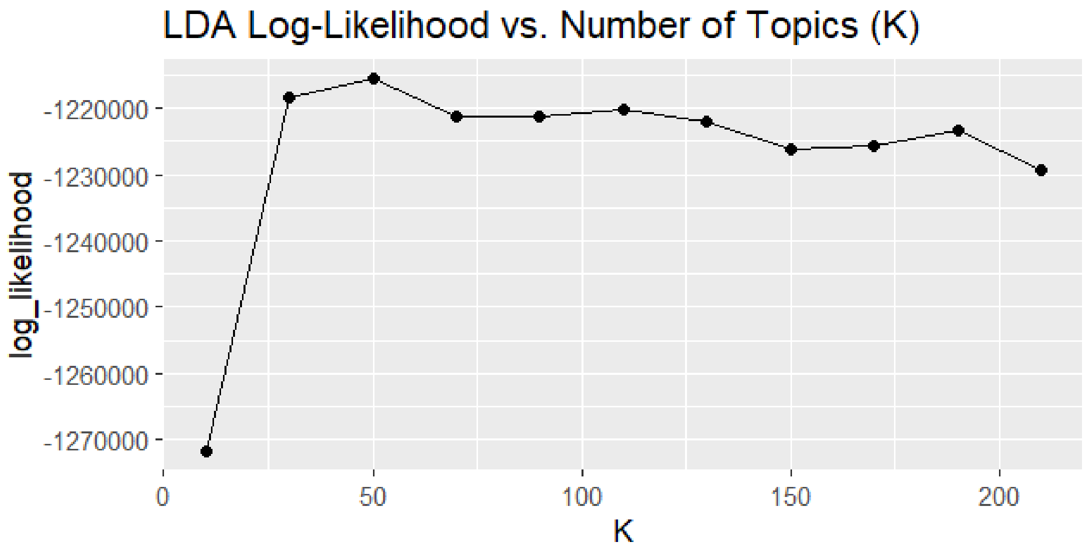
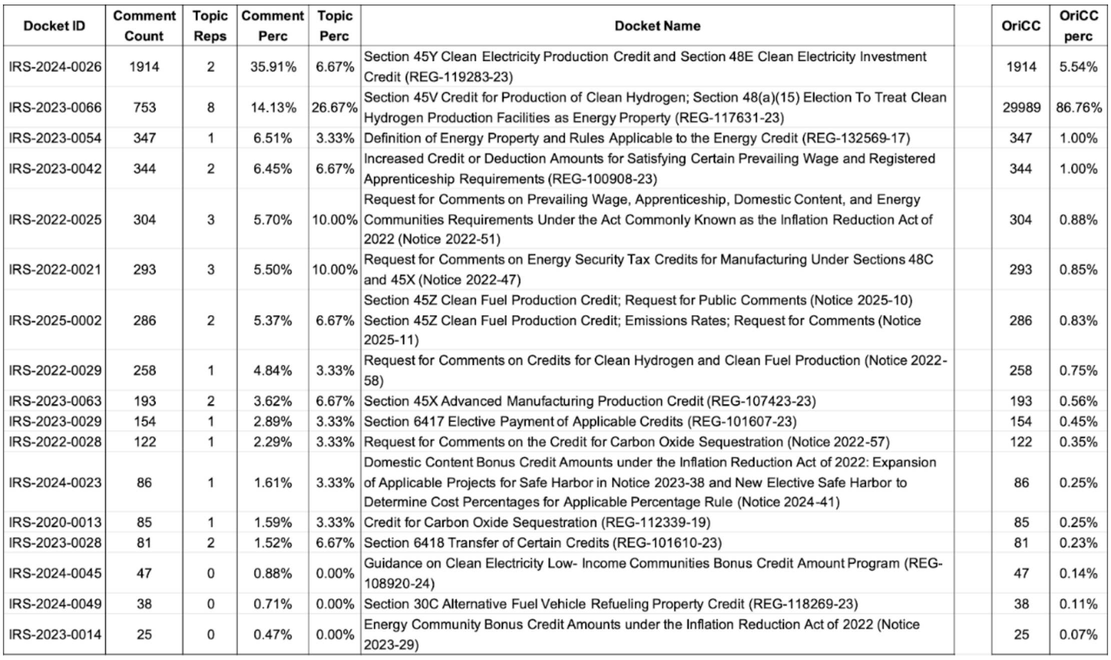


Figure 3

dbscan_cluster	count
-1	**3045**
15	**1441**
24	**220**
16	**170**
7	**58**
27	**55**
1	**37**

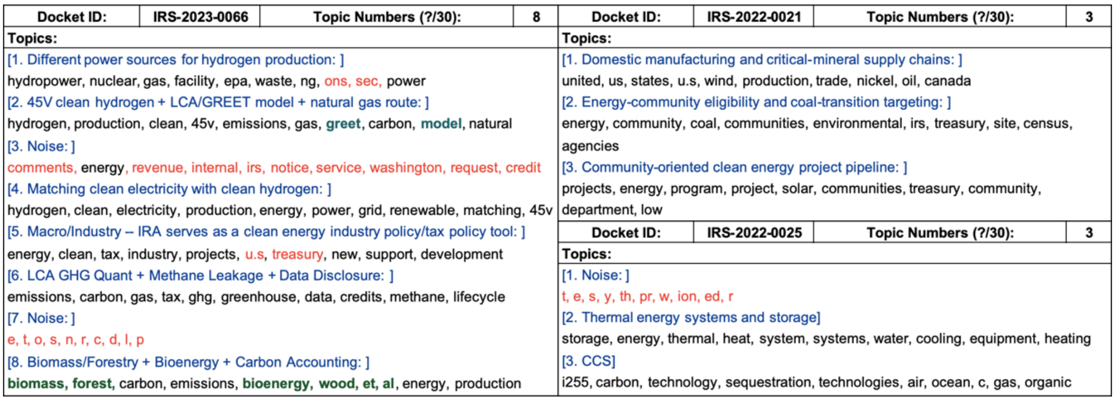
29	**32**
2	**19**
3	**18**
21	**14**
19	**12**
11	**12**
5	**12**
17	**11**
0	**11**
18	**10**
22	**10**
13	**9**
26	**8**

Figure 4



Docket ID	Comment Count	Topic Reps	Comment Perc	Topic Perc	Docket Name	OriCC	OriCC perc
IRS-2024-0026	1914	2	35.91%	6.67%	Section 45Y Clean Electricity Production Credit and Section 48E Clean Electricity Investment Credit (REG-119283-23)	1914	5.54%
IRS-2023-0066	753	8	14.13%	26.67%	Section 45V Credit for Production of Clean Hydrogen; Section 48(a)(15) Election To Treat Clean Hydrogen Production Facilities as Energy Property (REG-117631-23)	29989	86.76%
IRS-2023-0054	347	1	6.51%	3.33%	Definition of Energy Property and Rules Applicable to the Energy Credit (REG-132569-17)	347	1.00%
IRS-2023-0042	344	2	6.45%	6.67%	Increased Credit or Deduction Amounts for Satisfying Certain Prevailing Wage and Registered Apprenticeship Requirements (REG-100908-23)	344	1.00%
IRS-2022-0025	304	3	5.70%	10.00%	Request for Comments on Prevailing Wage, Apprenticeship, Domestic Content, and Energy Communities Requirements Under the Act Commonly Known as the Inflation Reduction Act of 2022 (Notice 2022-51)	304	0.88%
IRS-2022-0021	293	3	5.50%	10.00%	Request for Comments on Energy Security Tax Credits for Manufacturing Under Sections 48C and 45X (Notice 2022-47)	293	0.85%
IRS-2025-0002	286	2	5.37%	6.67%	Section 45Z Clean Fuel Production Credit; Request for Public Comments (Notice 2025-10) Section 45Z Clean Fuel Production Credit; Emissions Rates; Request for Comments (Notice 2025-11)	286	0.83%
IRS-2022-0029	258	1	4.84%	3.33%	Request for Comments on Credits for Clean Hydrogen and Clean Fuel Production (Notice 2022-58)	258	0.75%
IRS-2023-0063	193	2	3.62%	6.67%	Section 45X Advanced Manufacturing Production Credit (REG-107423-23)	193	0.56%
IRS-2023-0029	154	1	2.89%	3.33%	Section 6417 Elective Payment of Applicable Credits (REG-101607-23)	154	0.45%
IRS-2022-0028	122	1	2.29%	3.33%	Request for Comments on the Credit for Carbon Oxide Sequestration (Notice 2022-57)	122	0.35%
IRS-2024-0023	86	1	1.61%	3.33%	Domestic Content Bonus Credit Amounts under the Inflation Reduction Act of 2022: Expansion of Applicable Projects for Safe Harbor in Notice 2023-38 and New Elective Safe Harbor to Determine Cost Percentages for Applicable Percentage Rule (Notice 2024-41)	86	0.25%
IRS-2020-0013	85	1	1.59%	3.33%	Credit for Carbon Oxide Sequestration (REG-112339-19)	85	0.25%
IRS-2023-0028	81	2	1.52%	6.67%	Section 6418 Transfer of Certain Credits (REG-101610-23)	81	0.23%
IRS-2024-0045	47	0	0.88%	0.00%	Guidance on Clean Electricity Low- Income Communities Bonus Credit Amount Program (REG-108920-24)	47	0.14%
IRS-2024-0049	38	0	0.71%	0.00%	Section 30C Alternative Fuel Vehicle Refueling Property Credit (REG-118269-23)	38	0.11%
IRS-2023-0014	25	0	0.47%	0.00%	Energy Community Bonus Credit Amounts under the Inflation Reduction Act of 2022 (Notice 2023-29)	25	0.07%

Figure 5



Docket ID:	IRS-2023-0066	Topic Numbers (?/30):	8
Topics:			
[1. Different power sources for hydrogen production:]			
hydropower, nuclear, gas, facility, epa, waste, ng, ons, sec, power			
[2. 45V clean hydrogen + LCA/GREET model + natural gas route:]			
hydrogen, production, clean, 45v, emissions, gas, **greet**, carbon, **model**, natural			
[3. Noise:]			
comments, energy, revenue, internal, irs, notice, service, washington, request, credit			
[4. Matching clean electricity with clean hydrogen:]			
hydrogen, clean, electricity, production, energy, power, grid, renewable, matching, 45v			
[5. Macro/Industry -- IRA serves as a clean energy industry policy/tax policy tool:]			
energy, clean, tax, industry, projects, u.s, treasury, new, support, development			
[6. LCA GHG Quant + Methane Leakage + Data Disclosure:]			
emissions, carbon, gas, tax, ghg, greenhouse, data, credits, methane, lifecycle			
[7. Noise:]			
e, t, o, s, n, r, c, d, l, p			
[8. Biomass/Forestry + Bioenergy + Carbon Accounting:]			
biomass, forest, carbon, emissions, **bioenergy, wood, et, al**, energy, production			

Docket ID:	IRS-2022-0021	Topic Numbers (?/30):	3
Topics:			
[1. Domestic manufacturing and critical-mineral supply chains:]			
united, us, states, u.s, wind, production, trade, nickel, oil, canada			
[2. Energy-community eligibility and coal-transition targeting:]			
energy, community, coal, communities, environmental, irs, treasury, site, census, agencies			
[3. Community-oriented clean energy project pipeline:]			
projects, energy, program, project, solar, communities, treasury, community, department, low			

Docket ID:	IRS-2022-0025	Topic Numbers (?/30):	3
Topics:			
[1. Noise:]			
t, e, s, y, th, pr, w, ion, ed, r			
[2. Thermal energy systems and storage]			
storage, energy, thermal, heat, system, systems, water, cooling, equipment, heating			
[3. CCS]			
i255, carbon, technology, sequestration, technologies, air, ocean, c, gas, organic			

Figure 6

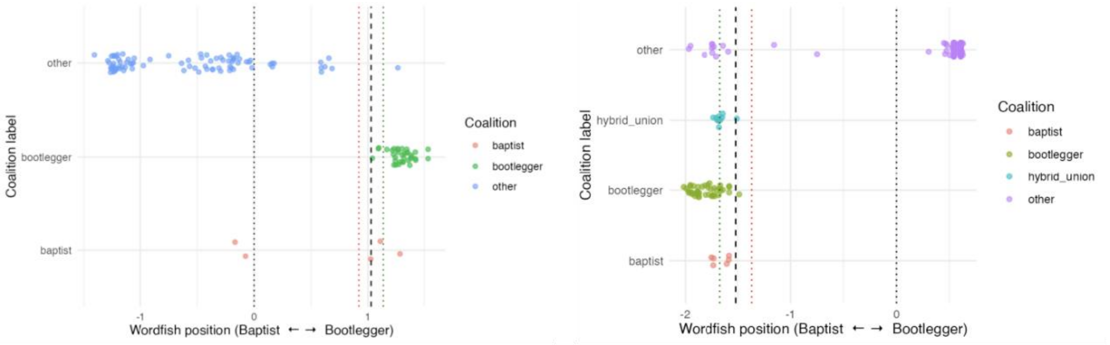
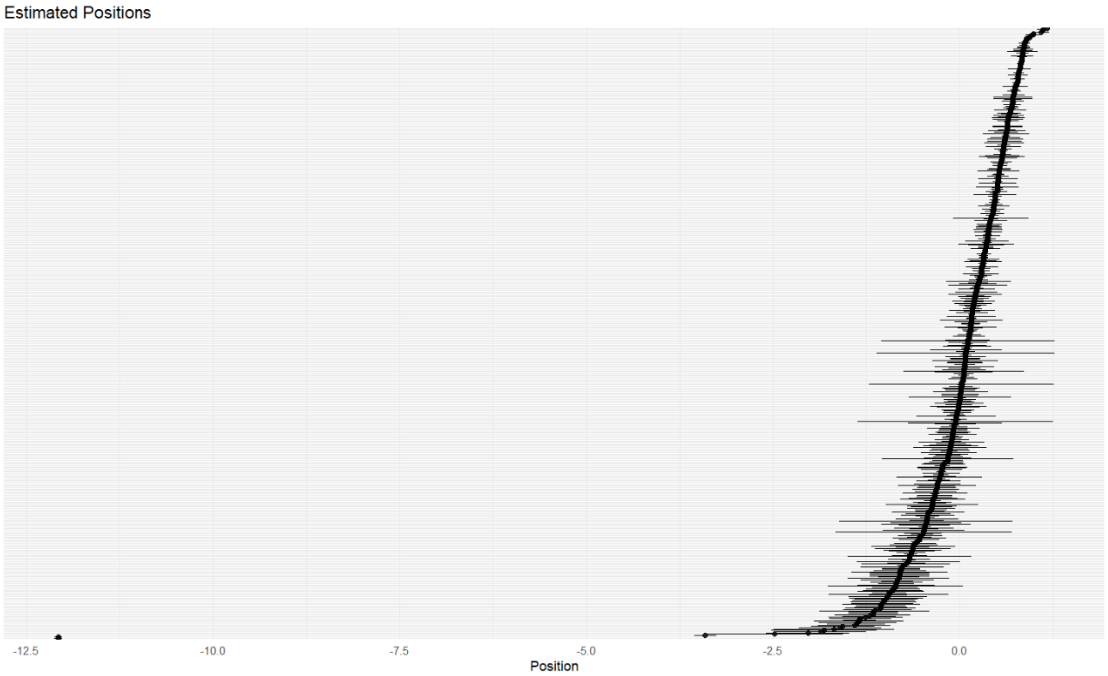


Figure 7

dbscan_cluster	**count**
0	4482
-1	501
4	185
3	101
7	44
10	13
6	12
9	12
11	9
12	8
1	8
2	8
5	6
13	5

dbscan_cluster	count
0	4482
-1	501
4	185
3	101
8	5
14	5

Figure 8


Estimated Positions
Position