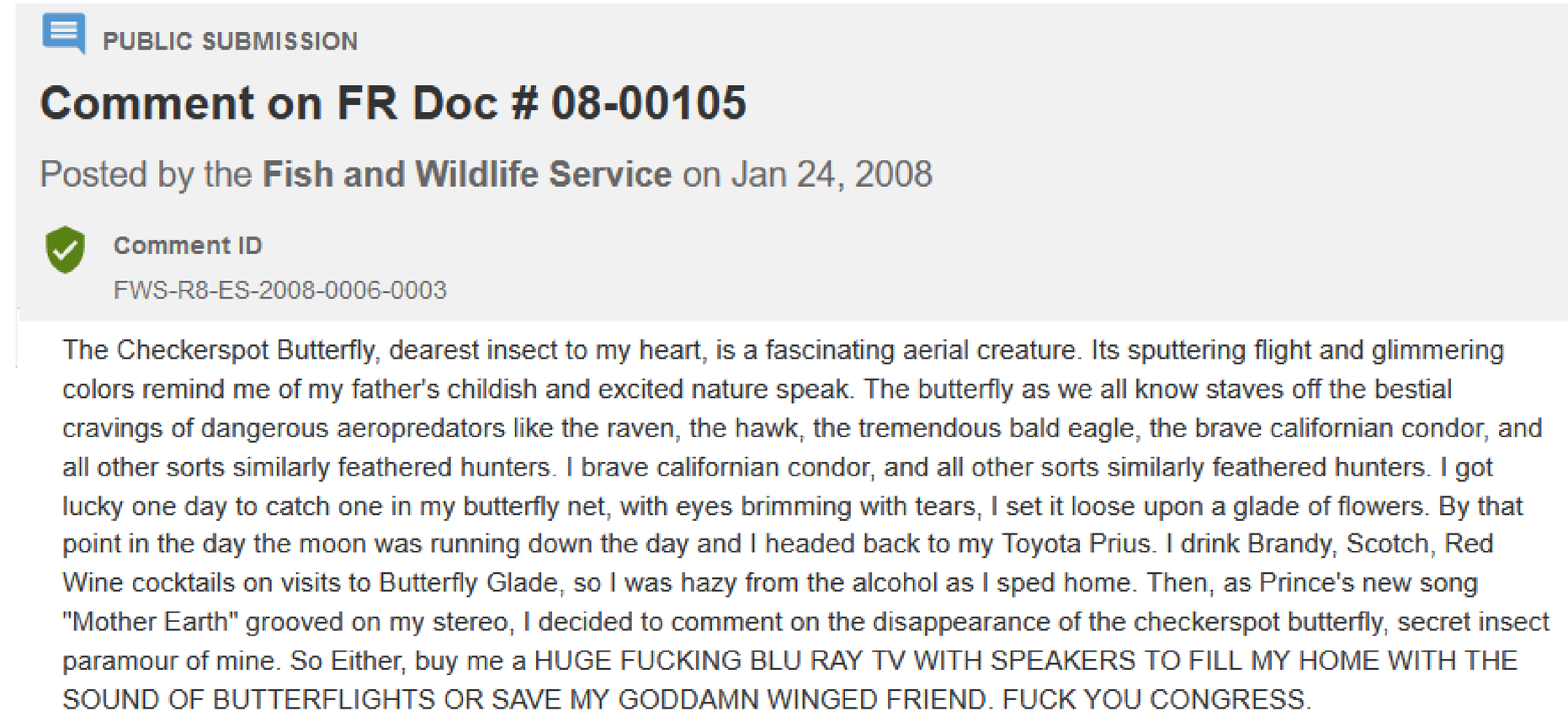
Streamlining the Processing of Public Comments

Will Jobs

University of Massachusetts, Amherst

BACKGROUND

Federal agencies are required to submit potential new regulations through a formal review process before putting these regulations into effect. Usually, this process includes a period in which citizens and organizations can submit public comments expressing their position on proposed regulations and providing perspectives and data the government may not have. The site *Regulations.gov* serves as a repository for over 250 federal agencies to submit proposed regulations and solicit public comments (notably, many "non-participating agencies" do not receive comments at *Regulations.gov* and may have their own public comment system). Based on the sample collected in this project, proposed regulations ("documents") receive **2,000 comments, on average**, and the maximum number of comments I observed in a single document was 223,585. Therefore, in order to process and respond to the comments (not individually) , agencies are likely to group comments into common themes that they can then respond to in bulk.


PUBLIC SUBMISSION

Comment on FR Doc # 08-00105

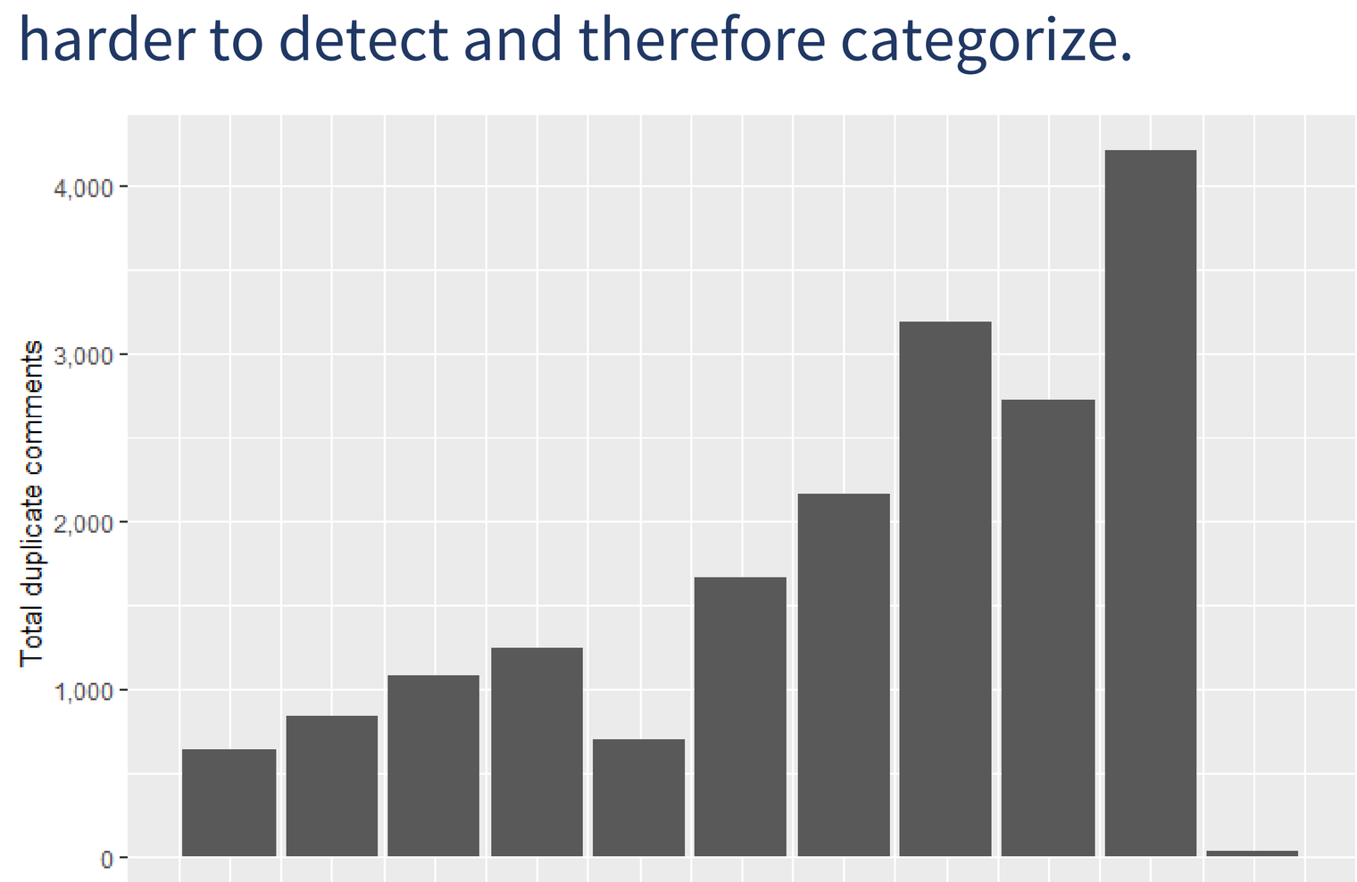
Posted by the **Fish and Wildlife Service** on Jan 24, 2008

Comment ID
FWS-R8-ES-2008-0006-0003

The Checkerspot Butterfly, dearest insect to my heart, is a fascinating aerial creature. Its sputtering flight and glimmering colors remind me of my father's childish and excited nature speak. The butterfly as we all know staves off the bestial cravings of dangerous aeropredators like the raven, the hawk, the tremendous bald eagle, the brave californian condor, and all other sorts similarly feathered hunters. I brave californian condor, and all other sorts similarly feathered hunters. I got lucky one day to catch one in my butterfly net, with eyes brimming with tears, I set it loose upon a glade of flowers. By that point in the day the moon was running down the day and I headed back to my Toyota Prius. I drink Brandy, Scotch, Red Wine cocktails on visits to Butterfly Glade, so I was hazy from the alcohol as I sped home. Then, as Prince's new song "Mother Earth" grooved on my stereo, I decided to comment on the disappearance of the checkerspot butterfly, secret insect paramour of mine. So Either, buy me a HUGE FUCKING BLU RAY TV WITH SPEAKERS TO FILL MY HOME WITH THE SOUND OF BUTTERFLIGHTS OR SAVE MY GODDAMN WINGED FRIEND. FUCK YOU CONGRESS.

One difficulty posed by the public comment system is the frequency of form letter comments. Form letters are those that are created as a template for others to send with carefully written language reflecting their, and a group's, opinion. People and organizations may send these en masse to try to convey broad consensus and, usually, disagreement and disdain for a proposed regulation. This is prevalent enough that *Regulations.gov* has a notice in their instructions explaining that **form letters "do not constitute a 'vote'"** and that **"a single, well-supported comment may carry more weight than a thousand form letters."** Even still, comments on a given document can be flooded with form letters, which may take various forms:

1. **Exact duplicates**
2. **Near-duplicate** (addition of a signature, a phrase like "thank you", etc.)
3. **Prompt-driven** (comments including the same key information, but often re-worded and rephrased to seem more individualized). This kind is *much* harder to detect and therefore categorize.



RESEARCH QUESTION

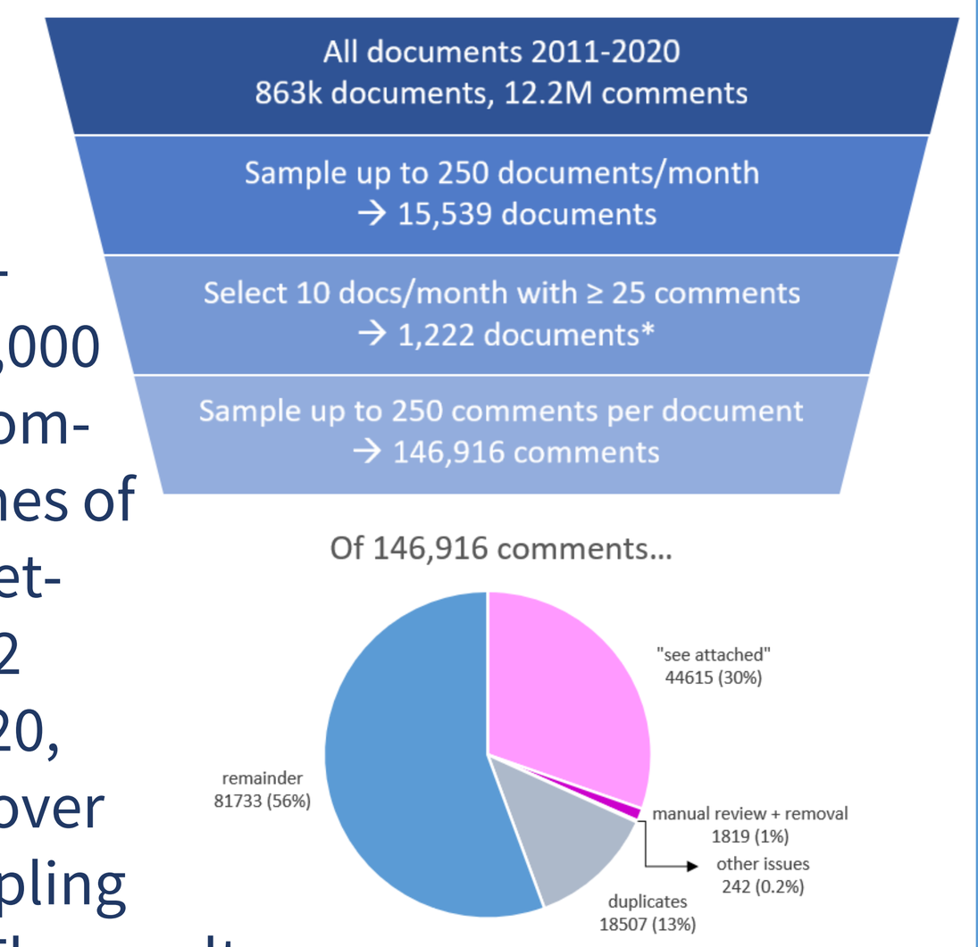
The goal of this project is to find ways to expedite the process federal agencies use to review public comments. Of primary importance is the ability to categorize comments into "topics" which can be responded to as a whole. This is difficult for a variety of reasons, including the fact that a comment may discuss multiple topics. Despite this, it may be possible to substantially reduce the number of comments that require more in-depth review.

Therefore, the questions underlying this project include:

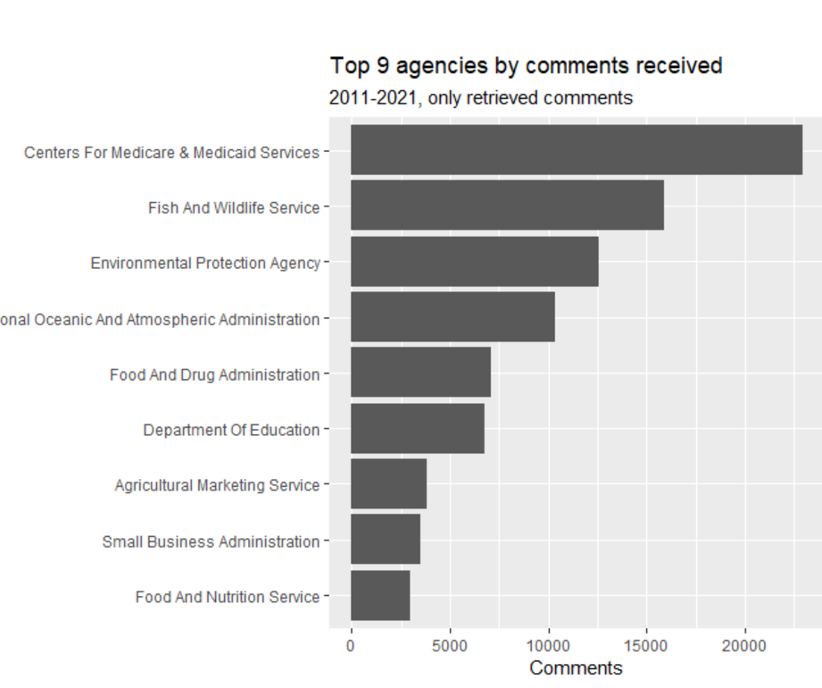
1. **Is it possible to develop a model to accurately and efficiently classify public comments into the categories of "form letter" and "not form letter"?**
2. **Can we use this information in tandem with a topic model to reduce the workload required by federal agencies to review public comments?**

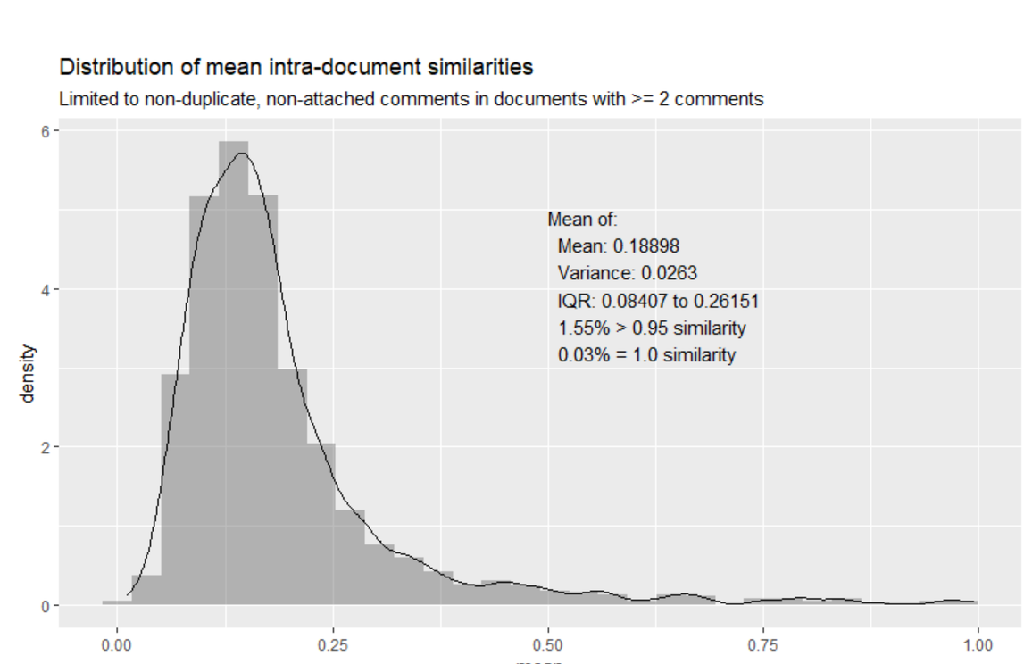
DATA

Comments on Regulations.gov are available via a public API that is well-documented but has its limitations. At most, 250 items can be retrieved per request, and an API key has access to only 1,000 requests per hour. Worse still, only the comment *metadata* can be retrieved in batches of 250; the text of each comment must be retrieved one request at a time. Given the 12 million comments between 2011 and 2020, these comments would have taken well over 500 days to download. As a result, a sampling scheme was devised, illustrated above. The result was **146,916 comments from 1,222 documents** over the 10-year period, with an **average of 120 comments collected per document**.



Comments that were attached and not provided in plain-text were excluded from downstream analyses, reducing the dataset by about 30%. Exact duplicates were identified based on lowercased comments; **18,507 duplicate comments (18.5% of the clean comments)** were identified. **Over 92% of documents** had at least one pair of duplicate comments.



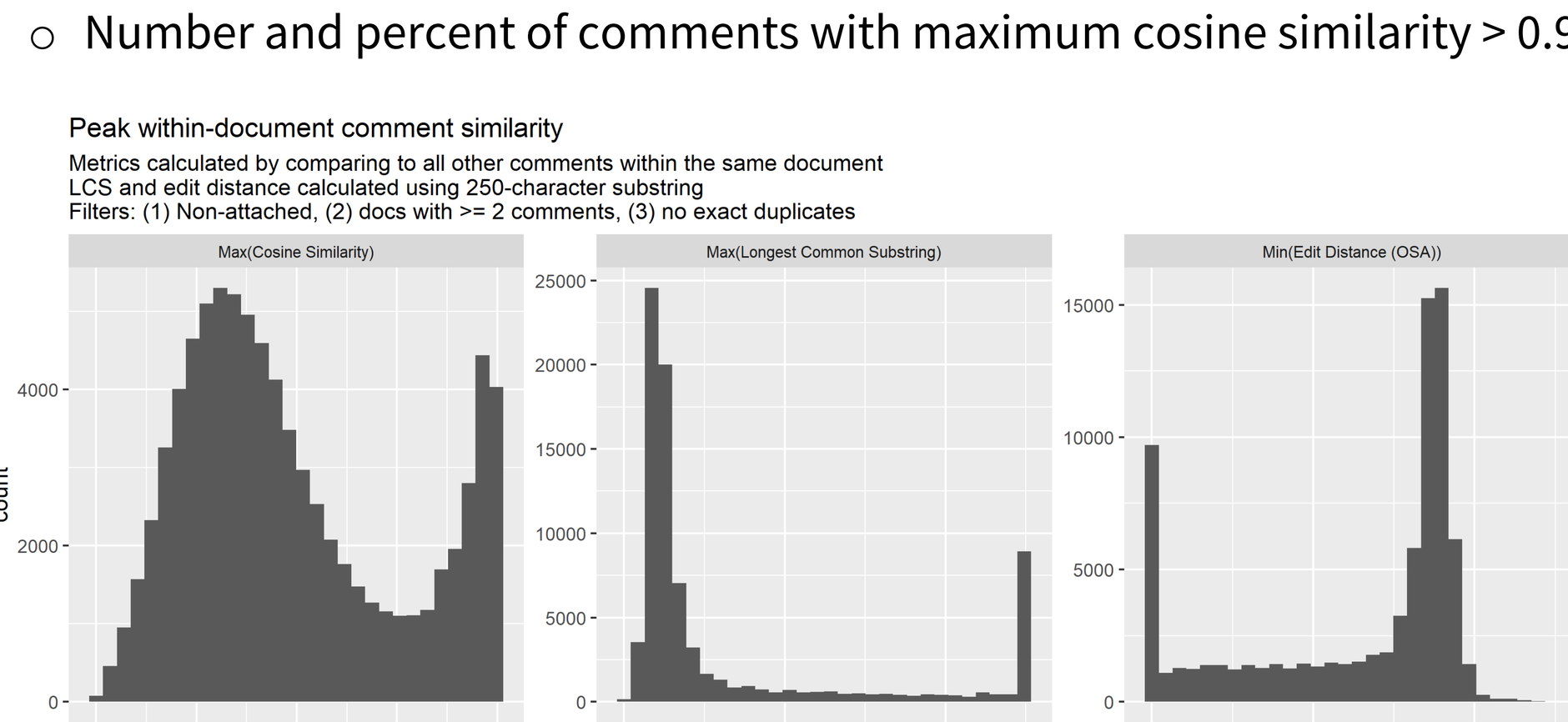


METHODS

1. Feature Engineering

The language of a comment is not itself useful for identifying form letters; rather it is the relationship between comments. The following features were computed for each comment:

- Comment-specific features:
  - Comment length
  - Total number of tokens
  - Number of unique tokens
  - Number and percent of "rare" tokens (in at least 10 comments but fewer than 250)
  - Emotional polarity (via NRC Word-Emotion Lexicon)
- Relational features
  - Maximum cosine similarity
  - Maximum longest common substring (LCS)
  - Minimum edit distance (Damerau-Levenshtein optimal string alignment)
- Document (regulation)-wide features
  - Total number of comments (including those not collected)
  - Number and percent of exact duplicates
  - Number and percent of comments with maximum cosine similarity > 0.9



2. Training a Classifier

1,600 selected comments were hand-labeled to train a supervised model, balanced across the top nine most-commented agencies, all years, and a diversity of document-comment structures. A logistic regression model was fit to all standardized features of the hand-labeled data. The feature set was reduced via stepwise selection and further simplified by identifying features that convey similar information and trying to choose a minimal set of comprehensive, roughly orthogonal features. All models fit well according to the Hosmer-Lemeshow test and were evaluated in terms of their accuracy, recall, precision, and F_1 score.

3. Topic Modeling and Clustering

A topic model (LDA) was fit to the comments from the top nine most-commented agencies, tuning the number of topics based on the perplexity evaluated on a **test set of 20%**. The model was then re-fit to 100% of the comments from the nine agencies and estimates of the proportion of agency comments by topic were computed. The topic vectors for each comment were clustered via K-means, with the optimal *K* chosen based on a plot of the within-sum-of-squares (WSS). This allowed comparison of the partitioning of comments by "form letter" vs. "not form letter" by the topic and topic-and-cluster methods.

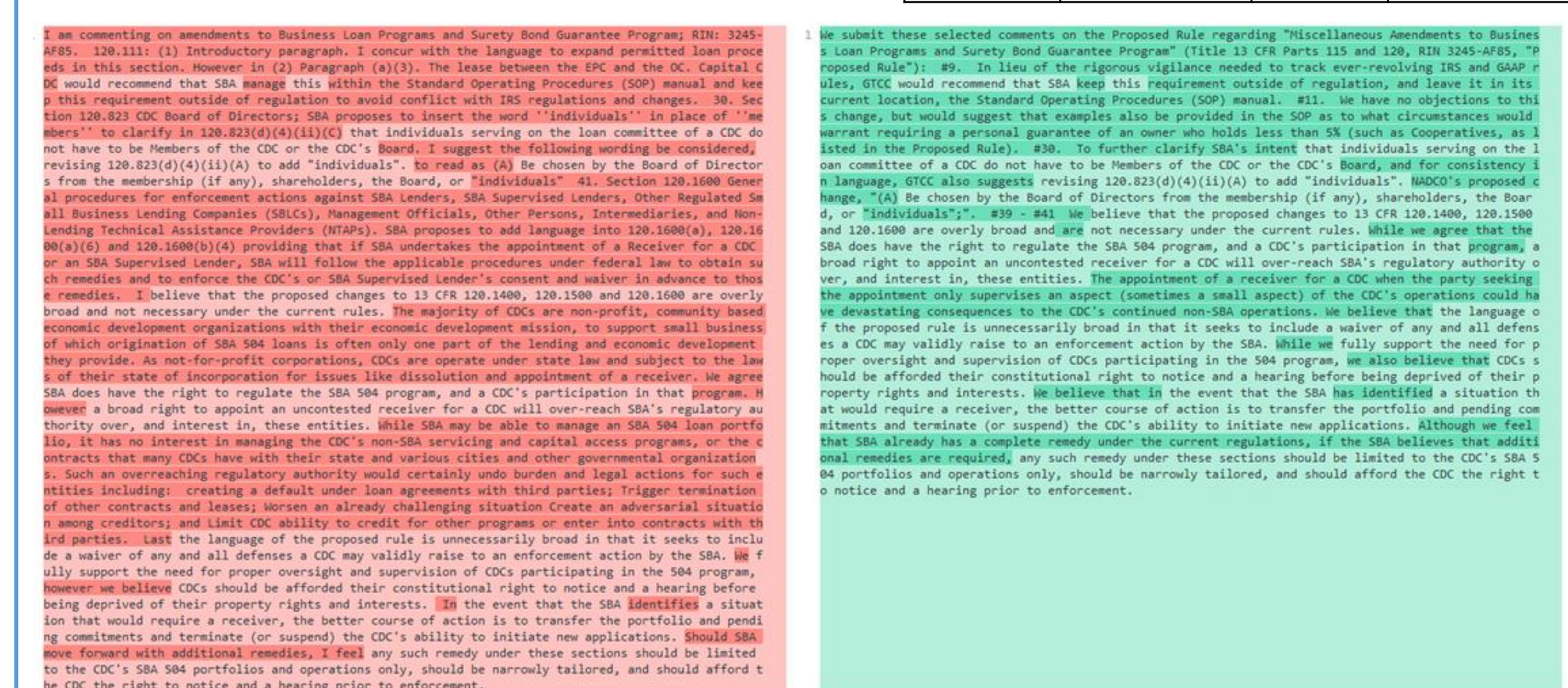
Finally, a separate topic model was fit to the comments of a single document. Because only 250 comments were available, the topic models were unstable, so additional comments were retrieved to bring the total up to approximately 5,000 comments for this document. The same topic and topic-and-cluster approaches were compared on the basis of the partitioning of the comments.

RESULTS

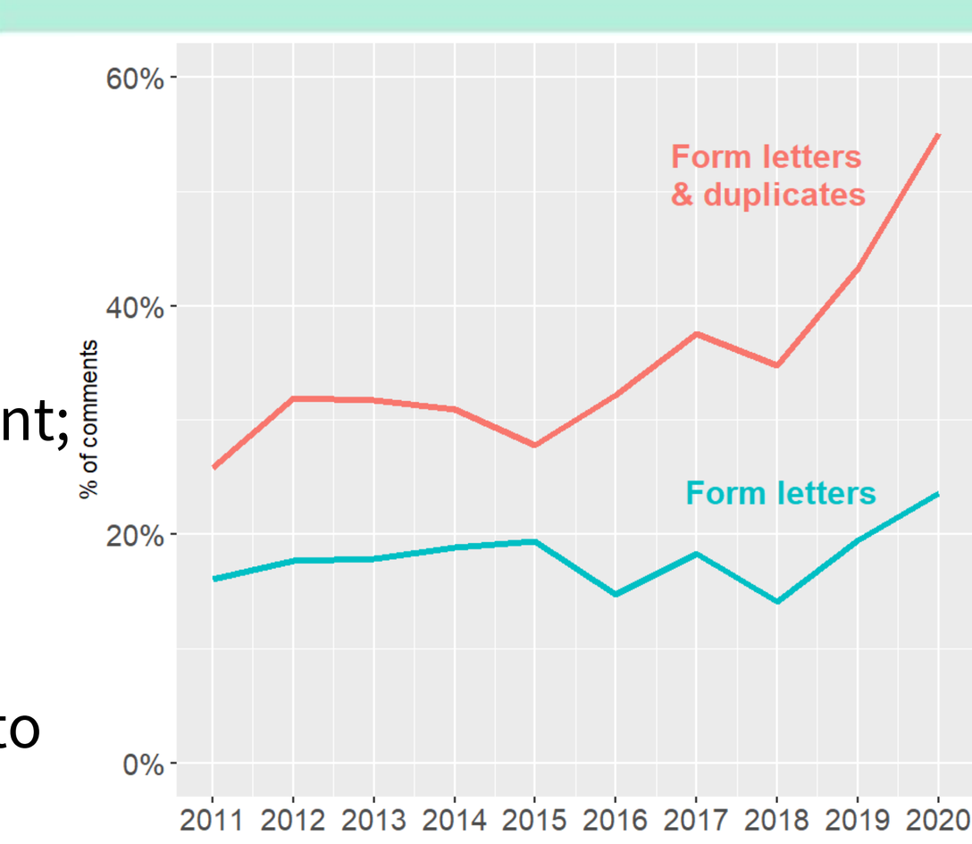
Form Letter Modeling

Feature	Coefficient	p-value
Max cosine similarity	4.2415	$< 10^{-15}$
Number of unique tokens	1.5210	0.02
Comment Length	-1.3382	0.06
Maximum LCS	0.8394	0.00002
Document: duplicates: % of all comments	0.7054	0.002
Document: number of comments	-0.4564	0.01
Emotional Polarity	0.1938	0.3

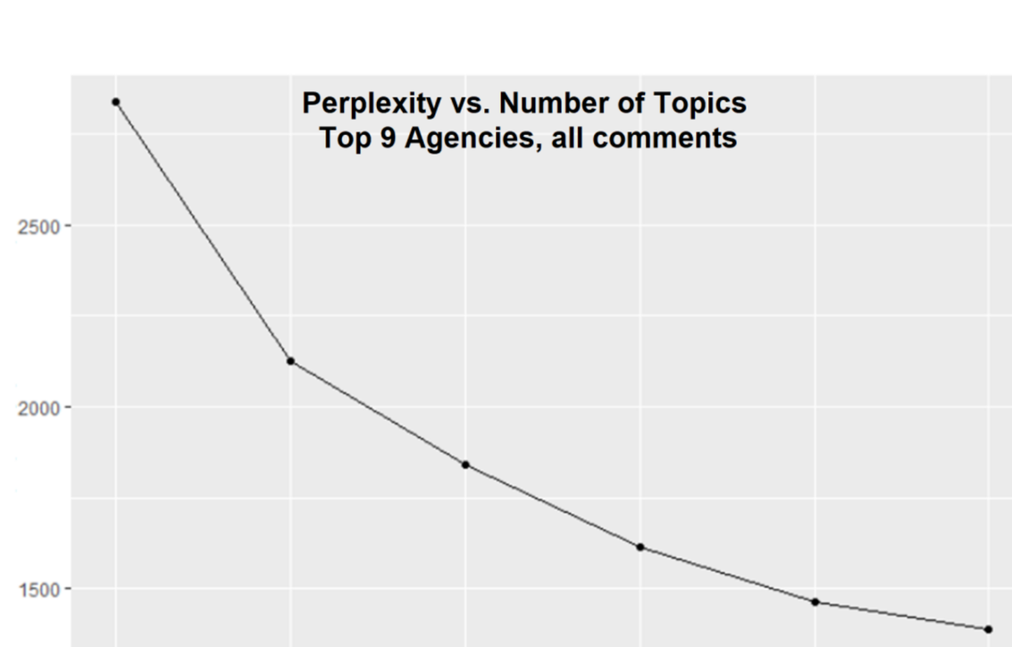
Agency	Precision	Recall	Accuracy
Overall	0.96	0.91	0.96
FWS	0.92	0.97	0.95
AMS	0.91	0.89	0.96
CMS	0.98	0.93	0.96
EPA	0.94	0.97	0.98
NOAA	0.93	0.83	0.96
FDA	0.98	0.98	0.98
ED	1.00	0.98	0.99
FNS	1.00	0.92	0.99
SBA	0.66	0.95	0.91

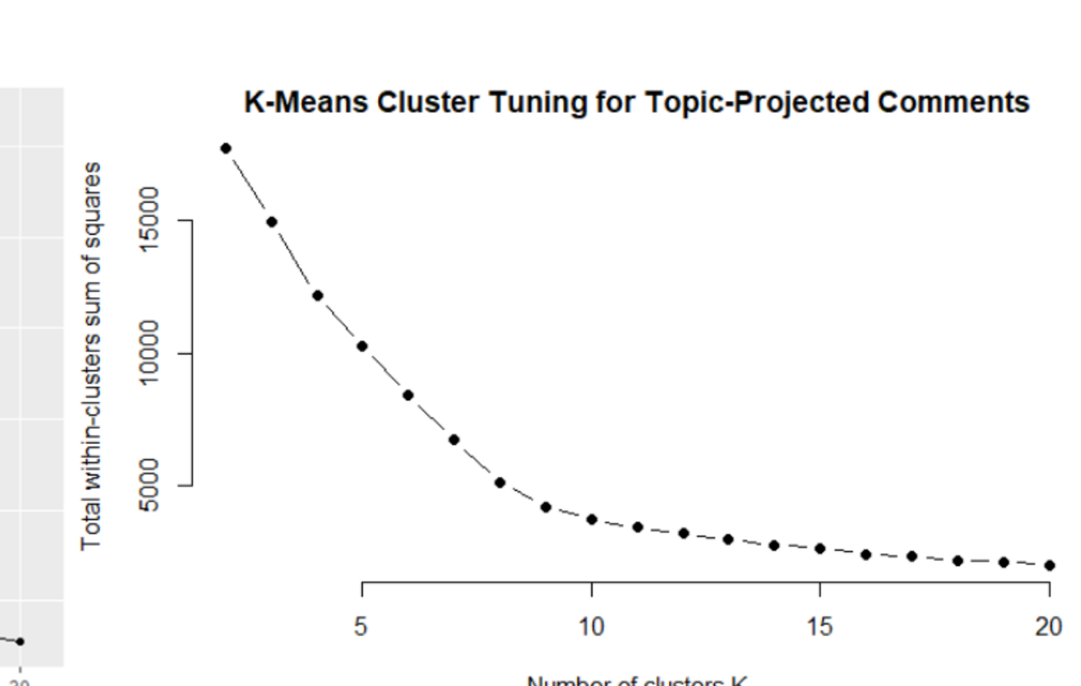


- Form letter analysis identified over 18,000 comments (threshold = 0.5)
- Overall, **over 37% of comments are form letters or exact duplicates**
- Median of 5 form letters per document; 64% of documents have some form letters
- Successfully identified form letter comments I might have missed due to significant re-wording
- Model somewhat less certain when there are many form letter comments that are very long, even if they are highly similar

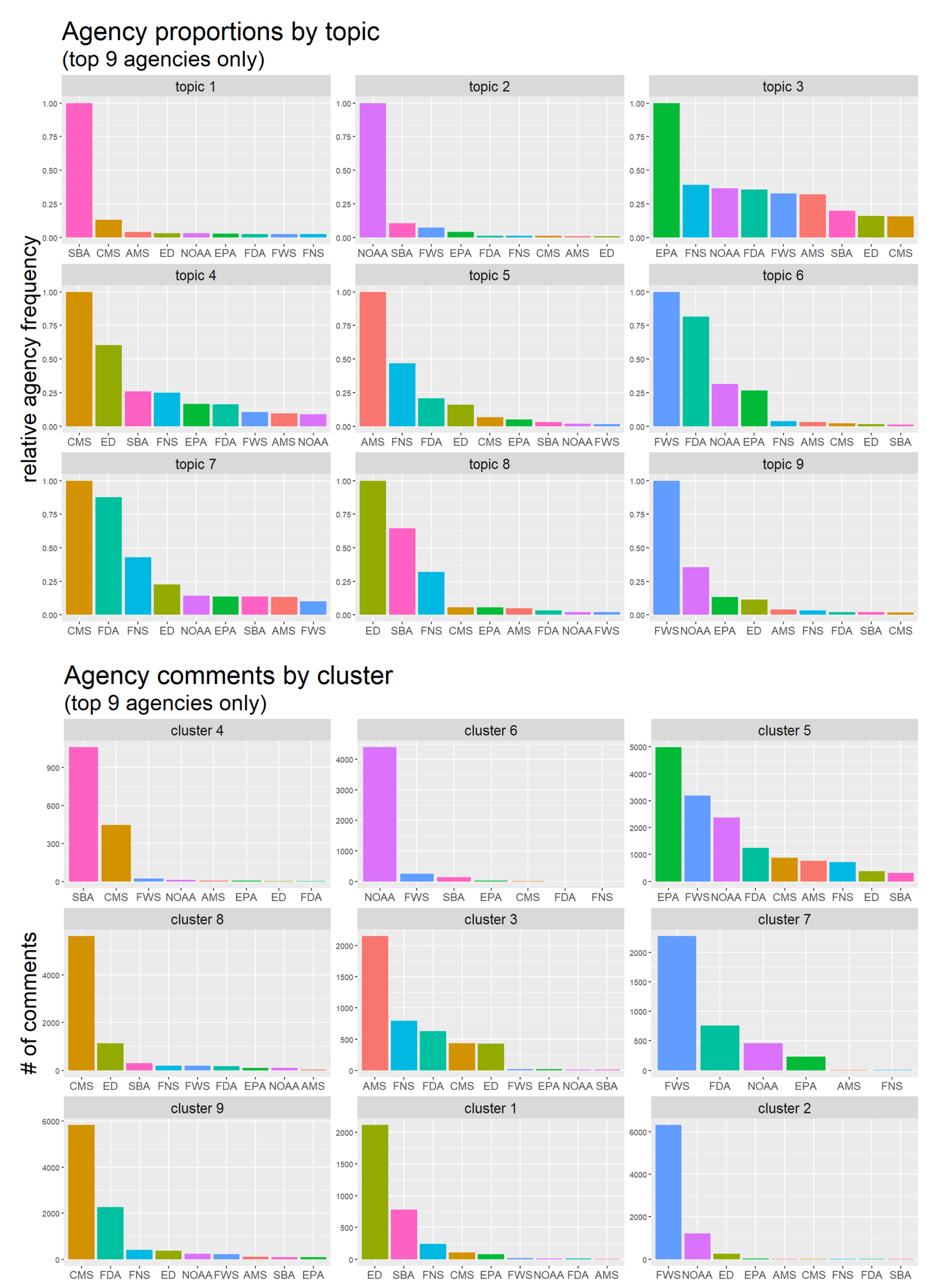


Topic Modeling





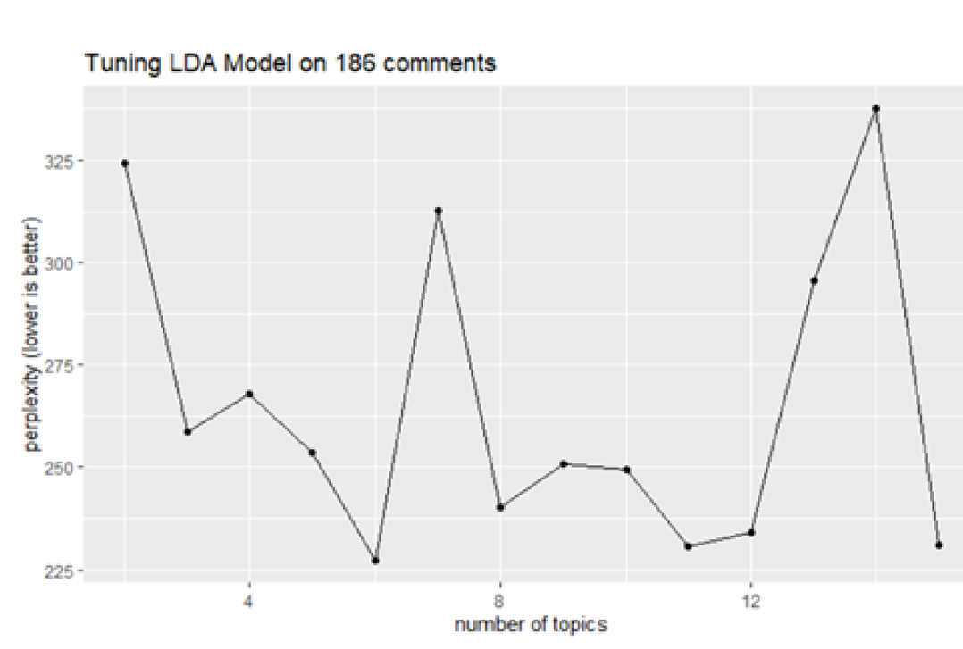
Partitioning top 9 agencies' comments by topic and cluster

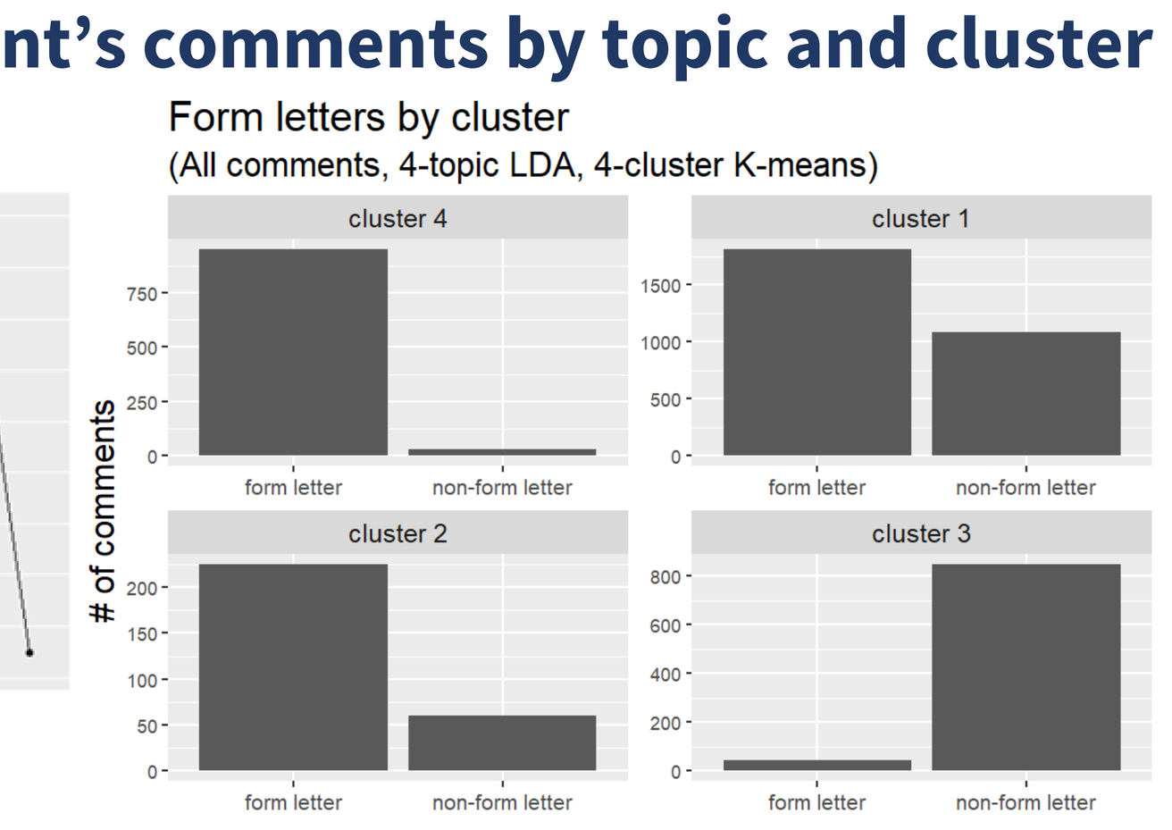


Topic vs. Cluster Ratio: Agency 1 / Agency 2

Topic	Cluster
7.6	2.4
9.5	**17.9**
2.6	1.6
1.7	**5.0**
2.1	**2.7**
1.2	**3.0**
1.1	**2.6**
1.6	**2.7**
2.8	**5.2**

Partitioning one document's comments by topic and cluster





CONCLUSIONS

This study shows that it is possible to build a model to accurately classify comments by whether they are part of a form letter campaign or not, even if the comment has been substantially rephrased. It was found that over a third of comments in the dataset were some form of form letter (exact duplicate, near duplicate, or prompt-driven duplicate). Because the number of comments has risen substantially in the last ten years, and the fraction of form letters has risen as well, this kind of model may be vital to regulators aiming to respond to public comments in a short timeframe.

In addition, it was found that topic models can be used to categorize comments into broad topics, and that traditional clustering methods, such as K-means, can be used on the comments' topic vectors to further improve the segmentation provided by the form letter model. However, topic models are most effective when there are many competing concerns and interest groups among the comments, and less effective when comments are largely focused on a single issue.

Future work may improve upon this proof-of-concept in several ways. Expanding the dataset to include more than 250 comments per document is a great place to start, as I found that topic models fit to so few comments were very unstable. Parsing the text of attachments is also worth considering, as they make up a large fraction of all comments on *Regulations.gov* and may be different from plain-text comments in some important ways (e.g., if there is a bias for companies to submit comments via attachment and individuals via plain-text entry). It would also be interesting, though time-consuming, to gather comments from some non-participating agencies' websites, such as the FCC. Finally, each document has metadata indicating the keywords used to search for a proposed regulation. A study could be conducted to explore the relationship between these topics and the topics among the document's public comments.