

Newsgroups NLP Classification

by M. Boyd-Vasiliou, A. Weber

Background

In order to stimulate discussion and increase activity in our newsgroups, we've decided to run a nightly import of short news stories. The process should be automated, so the stories are landed into the appropriate-topic newsgroup(s). In order to do that, we will analyze the current newsgroup content and train a machine-learning model to identify the topic of these stories and post them to the right newsgroup.

Proof-of-concept

Performance

- Evaluation system should classify an article into a maximum of 3 of the test newsgroups, with 'None' of the newsgroups being a commonly expected result. We are starting with a subset of 20 newsgroups, and even the full range is by no means comprehensive.
- Appropriateness will be measured by the user response:
 - ➔ Positive - Active discussion on the article
 - ➔ Neutral - Ignore article
 - ➔ Negative - Complain about receiving article

Data Capture

- System must log articles and classifications, 'read' % click-throughs for basic performance and model training

Objectives

- By providing fresh and targeted content, our newsgroups remain relevant and don't lose more footprint to social media feeds.
- Increase user base by 30%, add 10 new newsgroups and double traffic in 12 months time

Data Overview

The 20 newsgroups dataset comprises around 18000 newsgroups posts on 20 topics split in two subsets: one for training (or development) and the other one for testing.

More information and code for loading the data can be found here:

[5.6.2. The 20 newsgroups text dataset — scikit-learn 0.19.2 documentation](#)

- 18,846 records from 20 different newsgroups. About 11,300 records are labelled with their newsgroup (by number), the remainder are unlabelled for testing
- Keeping two columns of data: content of posting, and grouping (newsgroup) number
- In order to prevent meta-data from infecting the content analysis, the postings can (optionally, at load time) be obtained without headers, usernames, and quoted (copied) content which would otherwise keep reappearing in the data
- Grouping is a numeric with no matching to newsgroup
 - Only used post-hoc for comparing our topic modelling to actual results

Below is the full list of grouping labels and the corresponding actual newsgroup topic names:

| Group Label | Newsgroup Name |
|-------------|--------------------------|
| 0 | alt.atheism |
| 1 | comp.graphics |
| 2 | comp.os.ms-windows.misc |
| 3 | comp.sys.ibm.pc.hardware |
| 4 | comp.sys.mac.hardware |
| 5 | comp.windows.x |
| 6 | misc.forsale |
| 7 | rec.autos |

| | |
|----|------------------------|
| 8 | rec.motorcycles |
| 9 | rec.sport.baseball |
| 10 | rec.sport.hockey |
| 11 | sci.crypt |
| 12 | sci.electronics |
| 13 | sci.med |
| 14 | sci.space |
| 15 | soc.religion.christian |
| 16 | talk.politics.guns |
| 17 | talk.politics.mideast |
| 18 | talk.politics.misc |
| 19 | talk.religion.misc |

Data Exploration

Data dictionary “newsgroup posts”

| Column Name | Short Description | Notes for our application |
|-------------|---------------------------------|---------------------------|
| text | content of newsgroup posting | |
| label | numeric label of newsgroup name | |

Data Manipulation and Feature Extraction

The first thing we have to do is reduce and simplify the language that we feed into our algorithms. Pycaret does this automatically as part of the setup routine, it removes common language like “on” and “before” (prepositions), “a” and “the” (articles), and so on, as they are shared vocabulary that does not help distinguish topics from each other. These are known as Stop Words, and we are able to customize our word set by adding words to this list and thereby removing them from the data to be analyzed. It also removes punctuation and capitalization, and reduces words to their core component (lemmatization), so “studies” and “studied” become more instances of “study”.

(*What it also does, which was unfortunate when we were working on twitter messages, is it removes names and place names. Nouns are the most important component in topic grouping, and are usually defined as ‘a person, place or thing’. If you remove Person and Place from that, it severely limits your ability to characterize something as short and unstructured as a tweet. For instance if you thought Canadian Government was a potential topic, then removing the words Canada, Ottawa, and Trudeau from all text before sorting is obviously counter-productive.)

The second thing we do is represent the slimmed and trimmed language in a numeric table so the computer algorithms can analyze it. This is called vectorization. Two common methods are Bag of Words, and TF-IDF.

Bag of Words (BoW) is a plain count of the number of times a term(word) appears in the current document. From J. Brownlee, Machine Learning Mastery:

A bag-of-words is a representation of text that describes the occurrence of words within a document. It involves two things:

- 1. A vocabulary of known words.*
- 2. A measure of the presence of known words.*

TF-IDF The TF part (Term Frequency) is essentially like BoW above , a word count. However, [Karen Spärck Jones](#) (1972) conceived a statistical interpretation of term-specificity called Inverse Document Frequency (IDF), which became a cornerstone of term weighting.^[4] With TF-IDF, then, an IDF factor is incorporated which diminishes the weight of terms that occur very frequently in the document set and increases the weight of terms that occur rarely.

Modeling

We decided to use a Latent Dirichlet Allocation Model (LDA) as it is an excellent model for topic discovery in Natural Language Processing problems.

from Wikipedia:

One application of LDA in [machine learning](#) - specifically, [topic discovery](#), a subproblem in [natural language processing](#) - is to discover topics in a collection of documents, and then automatically classify any individual document within the collection in terms of how "relevant" it is to each of the discovered topics. A topic is considered to be a set of terms (i.e., individual words or phrases) that, taken together, suggest a shared theme.

Interestingly, while TF-IDF is a conceptual improvement over BoW, the LDA algorithm does not use linear algebra, but is rather a probability based model and it is designed to work on a 'plain-count' word representation. So our processing pipeline uses the standard BoW vectorizer.

Data Loading and Initial Model Generation

For our analysis we chose to use pycaret and LDA as the underlying technology and model. Pycaret is a great low-code package that provides easy model and chart generation, but there are very few parameters available to be modified for Natural Language Processing (NLP). Additionally, Pycaret NLP package does not yet provide any predict_model functionality, so our project will be focusing on analysis.

Of the initial dataset we choose a random sample of 942 (5%) of the dataset for our initial analysis. The 942 records spanned all 20 newsgroups. This size of data was chosen to simplify the analysis with a smaller initial vocabulary size, as well as allow the generation of the model and all associated charts in a reasonable time frame.

| Description | Value |
|------------------|-------|
| session_id | 123 |
| Documents | 942 |
| Vocab Size | 8915 |
| Custom Stopwords | False |

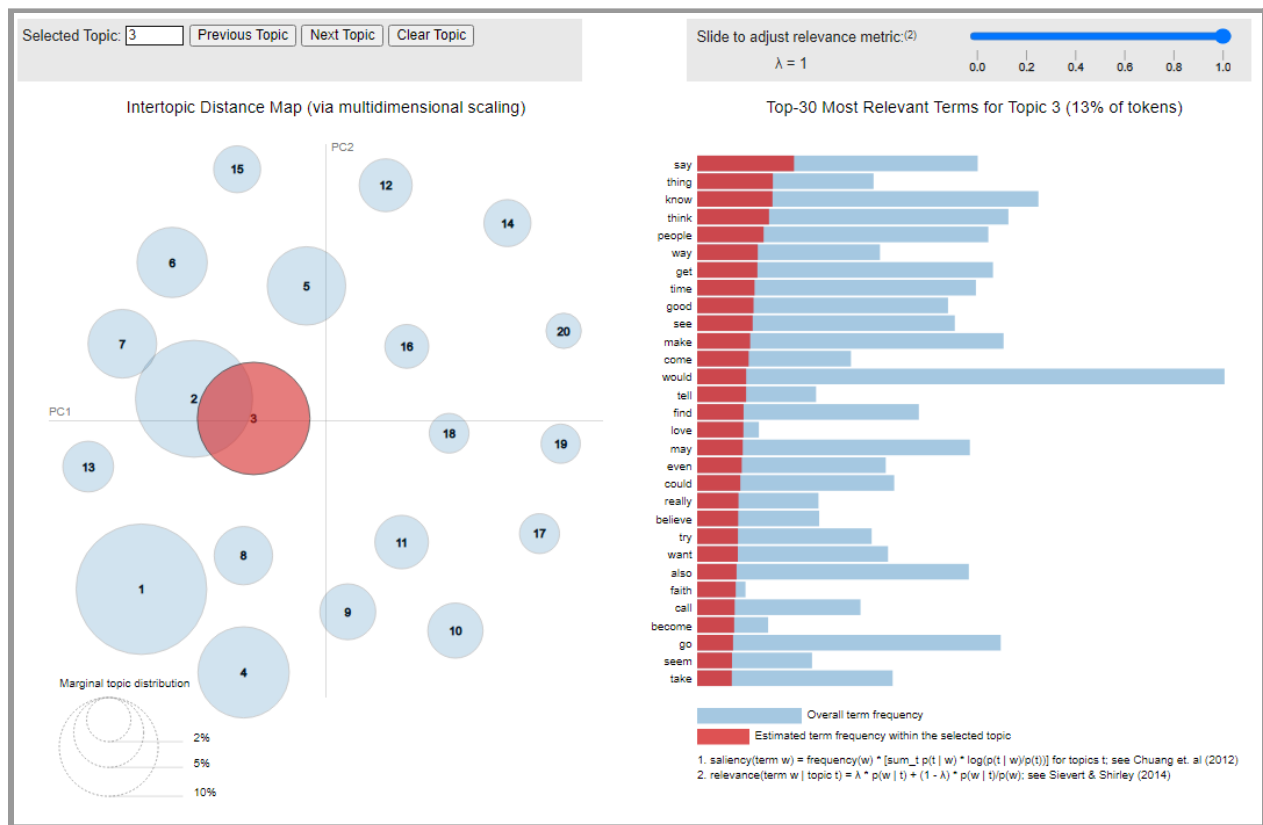
Note: additional data was included later in our experiments and is documented in our methodology below.

Topic Model Plot - Experiment 1

The topic model plot allows us to get a visual representation of how the data fits into our generated model. The plot is separated into two parts: left and right side.

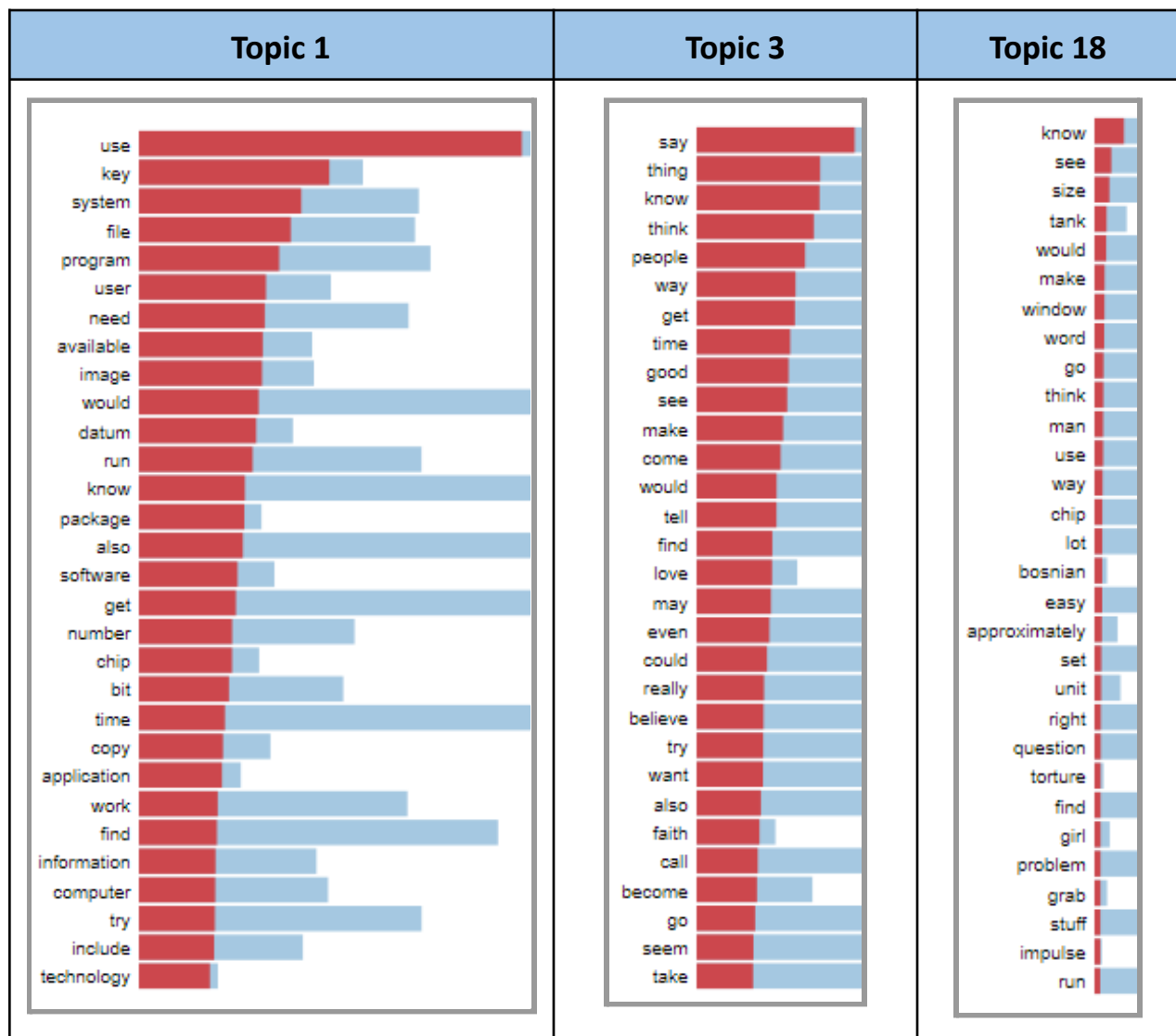
The left side of the plot, called an “intertopic distance map”, displays each topic as a circle, with the size of the circle representing the statistical weight of the topic. The separation of the topics is displayed by the separation of circles.

When selecting a topic, the right side of the plot will show the top words relevant to the chosen topic. The red bar shows the prevalence of that word in the chosen topic, with the blue component showing the prevalence of that word in the overall model



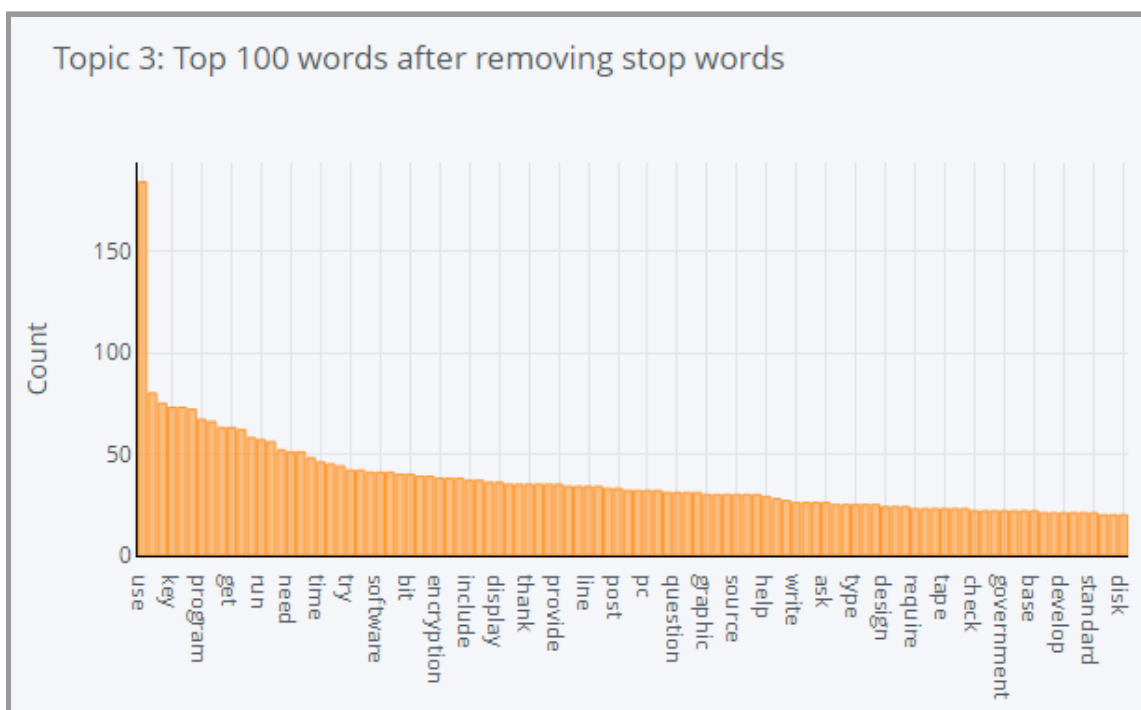
The initial model showed a few large topics, and a number of much smaller topics that were statistically far apart from one another. There were a few topics that were overlapping and some significantly so.

Examining the topics to determine word relevance showed that most topics were dominated by very common words such as “use”, “say”, “know”, “see”. These words had high prevalence in a number of topics across the model and were dominating the topics themselves. These words are not representative of any of the original topics and do not help us distinguish between them.



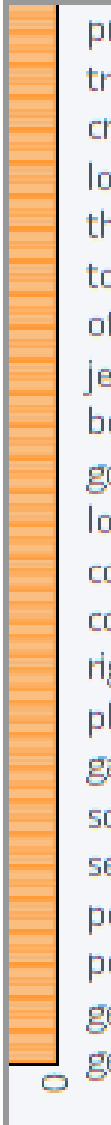


Word Frequency - Experiment 1

The word frequency plots shows us similar information to the previously mentioned topic plot but with a different visualization. Again we notice that topics are dominated by words that are unhelpful in distinguishing them from the rest of the topics. There are words further down the prevalence list such as “encryption”, “pc”, “bit”, that may help identify the topic but they are obfuscated by all the noise of the common words.



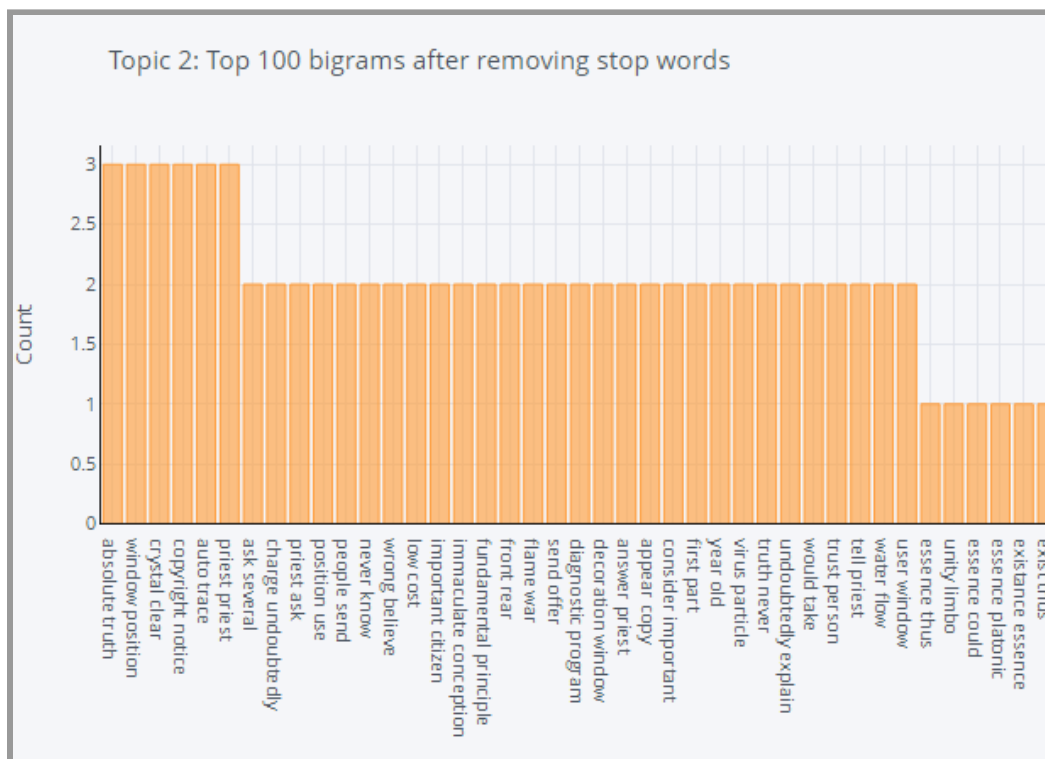
Examining all of the topics shows similar patterns of a number of words that would be helpful in identifying the topics (“hockey”, “player”, “jewish”, “protestant”) being drowned out by a plethora of noise from other less helpful words.

| Topic 3 | Topic 6 | Topic 17 |
|--|---|---|
|  <ul style="list-style-type: none"> come course buf collaborator demand people right kurdish last player also see game find win source go work seem may armenian percent tartar turkish point village know genocide say think get |  <ul style="list-style-type: none"> happen area feel post support always tell hockey truth state even delete show go get second people time think see use would |  <ul style="list-style-type: none"> protestant try crime look thank today official jewish book government long come collaborator right player game source seem percent point genocide get |




Bigrams - Experiment 1

The bigram plots show the frequency of occurrence of two words side by side, per topic. The adjacent words shown are those that remain together after all text processing and removal has been completed. The bigrams are considered n-grams where $n=2$.

After removal of all stopwords, the goal is for the bigrams to start displaying keywords that will help distinguish the topic as unique amongst others.

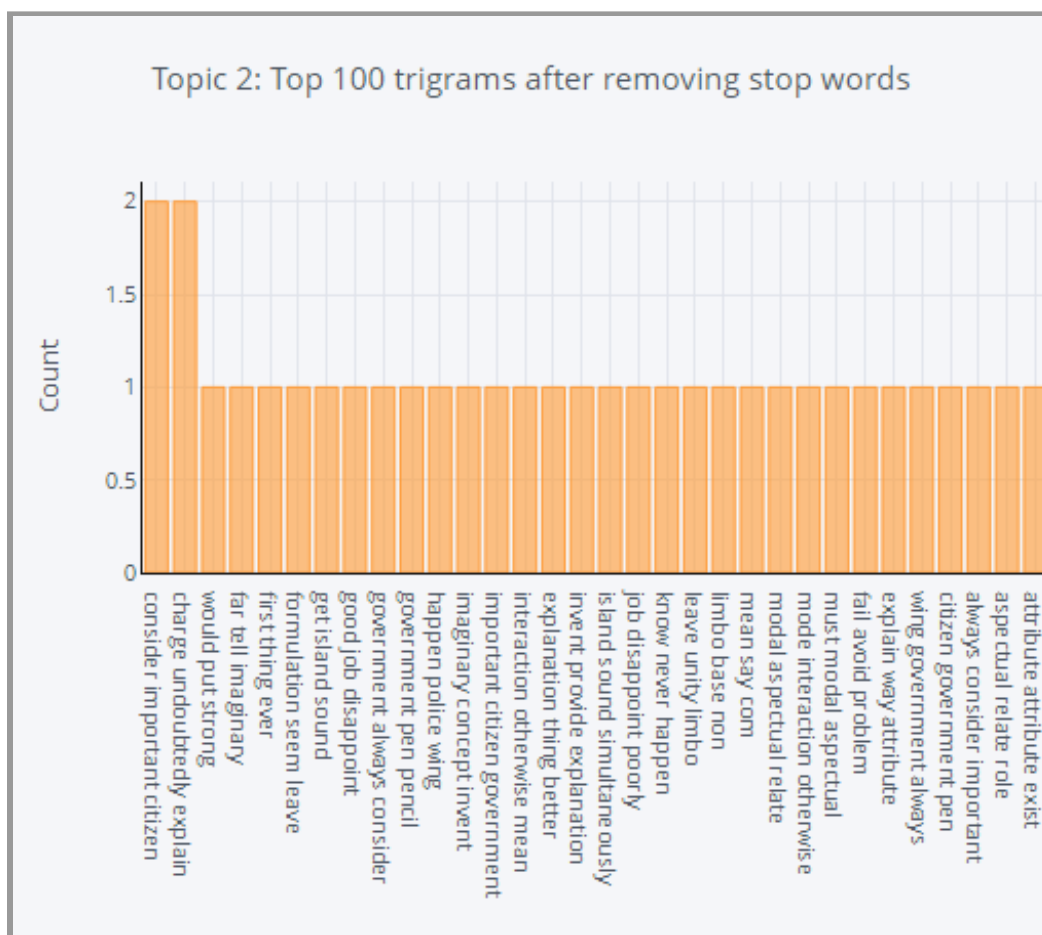


Examining the bigrams across all the topics revealed that there are some bigram phrases that appear to help distinguish the topics such as “involve firearm”, “firearm handgun”, “believe resurrection”, “anti semitic”. These distinguishing bigrams are significantly in the minority and are buried amongst very common and unhelpful bigrams, “fact know”, “would take”, “ring ring”, “fill certain”.




| Topic 14 | Topic 11 | Topic 19 |
|--|---|---|
|  <ul style="list-style-type: none"> call attention go wrong anti semitic send troop power sell assault homicide twice year take place next day nationally uniform firearm injury many people involve firearm may well last time would nice firearm handgun federal government homicide rate gun availability crime rate |  <ul style="list-style-type: none"> thermal observation take total morality dance death penalty come outside class people military consult believe resurrection ask military action take fact know federal officer hope much line duty low price give first would take ring ring passage time cooperative behaviour would make |  <ul style="list-style-type: none"> sort liver shim cost liver spot color indicate cause little call liver cage right brown spot arbitrary measure actual liver change size latter would hand call good_luck calculation death wrong may say yeast colonization bear wrong shim need fill certain space dock |

Trigrams - Experiment 1

Similar to bigrams, trigrams are n-grams with $n=3$, displaying the 3 most common words that appear beside each other after text processing.



Examining the trigrams across topics in our model again show that some of them assist in differentiating the various topics. Again though, some phrases are helpful in distinguishing the topics such as “decent home run”, “score still dodger”, and “issue pc magazine”, but the vast majority of trigrams are unhelpful in topic distinction.

| Topic 15 | Topic 13 | Topic 0 |
|--|--|--|
|  <p> stadium last year similar protectant son year fairly large get black plastic fairly large effect decent home run change name underneath car armor remove black plastic bumper apparently large still get look run go promote computer mile high stadium luck replace mold licensed change name last_year top year large effect homerun idea get look high stadium last score still dodger </p> |  <p> communiversity local family cause virus tumor benefit recurrent establish awareness rare disease appear local newspaper affect life enough continual surgery remove tumor grow block enough people inspire craft sale raise disease appear local daughter suffer disease undergone operation thus term fact arrest accuse ever make officially admit many long prison term crime accuse ever arrest crime accuse israeli citizen sentence </p> |  <p> day shall rise condemn death know betray know actual age know city death know would die know religious break bad habit good reason valid tomb body go week baby size evidence provide faith conclusion physical sphere even slight concept lie right side prince disarm subject percent economy grow ball live real grow percent economy issue pc magazine local bus slot </p> |

Wordcloud - Experiment 1

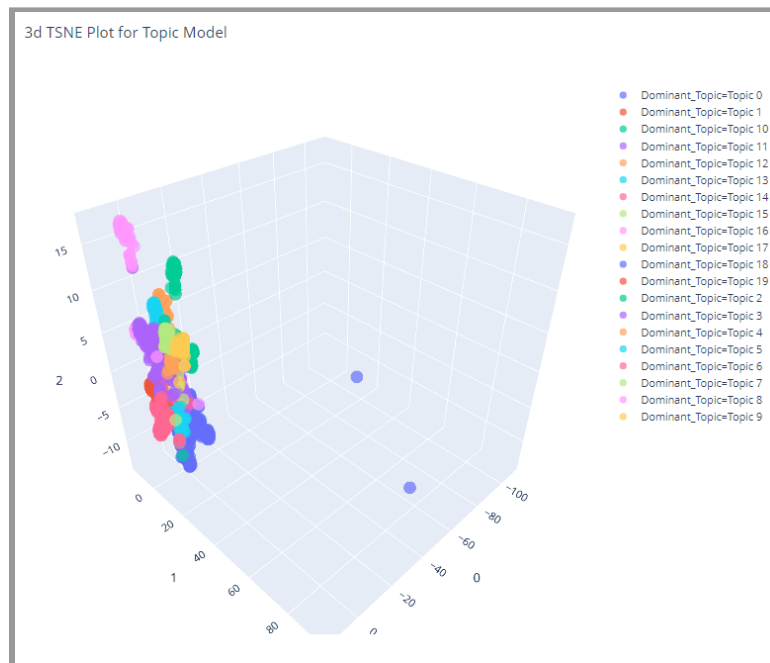
A wordcloud is a visual representation of the most common words in each of the topics. The larger and bolder the words are within the wordcloud the more often it appears within the topic.

Examining the wordclouds show that some have similar themed words that are present, such as “system”, “program”, “chip”, but it is not clear what the full theme is or what subgroup the particular topic may be in. As expected from the analysis completed in the topic model plot, word frequency, n-grams, the wordclouds display very common words as dominating each of the topics with no convergence that is easily able to be discerned.



TSNE Plot - Experiment 1

A t-distributed Stochastic Neighbor Embedding (t-SNE) plot allows a person to visualize complex associations in a reduced dimensionality. Our t-SNE plotting displays as a 3D map with different colour dots per topic. For strong topic clustering we would hope that our plot shows the topics clustered together, and separate from each other. Unsurprisingly, considering the common word overlap between the various topics, our t-SNE plot shows our topics bunched together with very little separation between them.



Experiment 1 - Summary

The data was successfully used to create a base model. Through our analysis we noticed that keywords that would help distinguish our newsgroups from each other are present but are drowned out by the noise of common and unhelpful words. There is strong potential in this model, but significant work needs to be completed.

Repeated Experiments

We completed 11 rounds of experiments on our data to extract topic information and try to elicit some noticeable differentiation between the various topics. We also completed an additional 4 rounds on a “forked” analysis of a subset of 10 newsgroups for comparison.

Analysis Methodology

Except as noted below, our analysis methodology for each experiment was:

1. Examined topic model plots for each topic
 - Keep lambda at 1 (except as noted later)
 - looked for words that were
 - non-topic specific (not indicative of any newsgroup)
 - high occurrence, in current and/or other topics
 - Look for signs of topic convergence in dominant words
2. Examined wordclouds for each topic
 - Look for words that are prevalent and non-defining
 - Look for signs of topic convergence with common
3. Examined bigrams and trigrams
 - Look for word combinations that are prevalent and non-defining
 - Look for signs of topic convergence with common
4. Examined t-SNE plot to ensure our analysis is moving in the right direction and toward topic convergence
5. Add custom stopwords to setup and re-run model and analysis

Experiments 1-5:

During experiments 1-5 we noted that topics were initially converging on common and unhelpful words (“see”, “go”, “use”) and as the initial common words were added to stopwords the topic models began to overlap and become less meaningful. Findings indicate that at the beginning topics were groupings on uncommon words showing a false level of separation between topics on the topic model plot. Initial removal of common words caused a bit of chaos in the model with additional overlaps between topics on the topic model.

There were no interesting details observed in either the Parts of Speech plots, or the sentiment plots during this time.

Experiment 6:

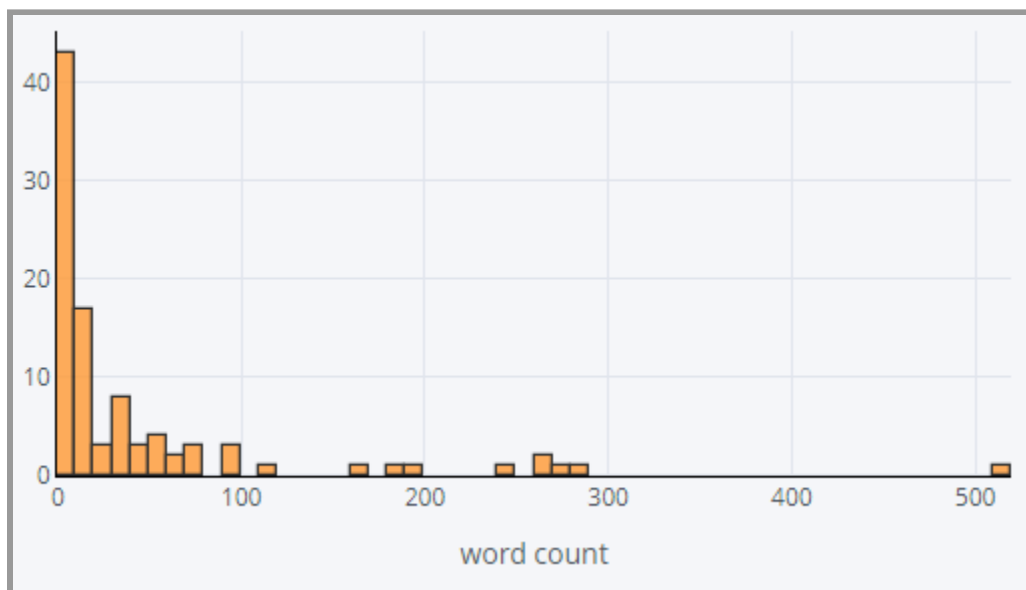
At this stage we altered our methodology and began looking at words for removal on the topic plot with Lambda set as both 1 and 0. We did this as the current methodology was yielding diminishing returns. By setting Lambda to 0 on the topic model plot it changed the weighting to show the words that are only dominant in the topic being examined. The goal was to remove not just common words across topics (Lambda 1), but words that are irrelevant and unique to the topic under consideration.

Topics that were beginning to converge with common themes provided the strongest opportunity to add Lambda 0 words to the stopwords list.

Experiment 7:

We noted that the topic convergence and changes in model were becoming less impactful. We examined the distribution plots of word count by documents in each topic and determined that after the text pre-processing of pycaret and our addition of extra stopwords that the number of documents with very few words remaining (0-9) was growing noticeably. At this stage (for Experiment 8) we added an additional 2000 records to our analysis. Understandably, this added in additional vocabulary and common words not seen before and initially created a bit more noise in the model.

| Description | Value |
|------------------|-------|
| session_id | 123 |
| Documents | 2827 |
| Vocab Size | 13799 |
| Custom Stopwords | True |

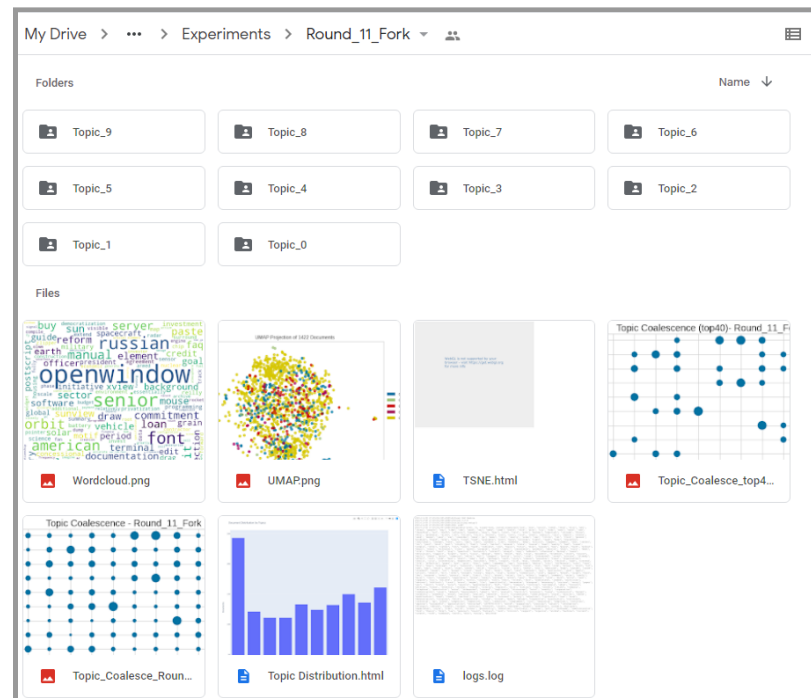
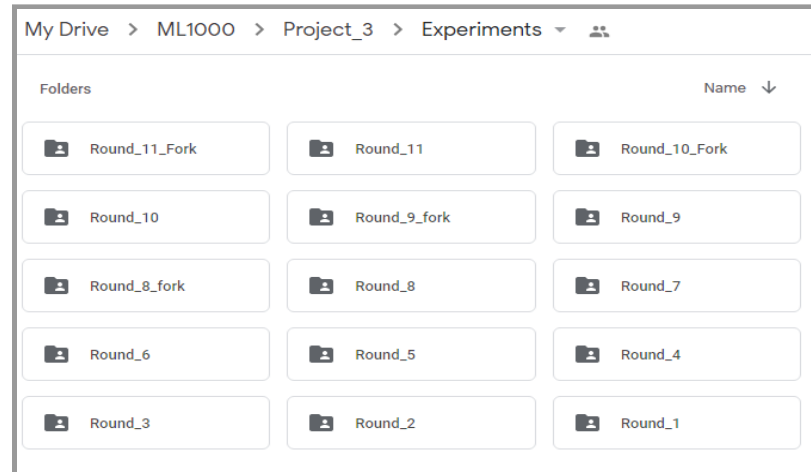


Experiment 8 and Experiment 8-Fork:

At this stage we created a fork in our model analysis. We continued with our current model, but created a concurrent “forked” model which only included 10 of the 20 newsgroups within our initial dataset. Topics selection was based on trying to include disparate categories to determine if topic separation may be more prevalent in a subset of the model

| | |
|--|---|
| <pre>cats = ['comp.graphics', '#comp.os.ms-windows.misc', '#comp.sys.ibm.pc.hardware', '#comp.sys.mac.hardware', 'comp.windows.x', '#rec.autos', 'rec.motorcycles', 'rec.sport.baseball', 'rec.sport.hockey', 'sci.crypt', '#sci.electronics', '#sci.med', 'sci.space', 'misc.forsale', 'talk.politics.misc', '#talk.politics.guns', '#talk.politics.mideast', '#talk.politics.war', 'sci.space', 'misc.forsale', 'talk.politics.misc', '#talk.politics.guns', '#talk.politics.mideast', '#talk.politics.war']</pre> | <p>Categories included in “fork”:</p> <ul style="list-style-type: none"> • comp.graphics • comp.windows.x • rec.motorcycles • rec.sport.baseball • rec.sport.hockey • sci.crypt • sci.space • misc.forsale • talk.politics.misc • alt.atheism |
|--|---|

We continued a forked model approach for our analysis purposes, but there was no significant differentiation between the forked and main model approach so we continued on our main path of analyzing all 20 newgroups. In total we completed and documented 15 separate experiments.

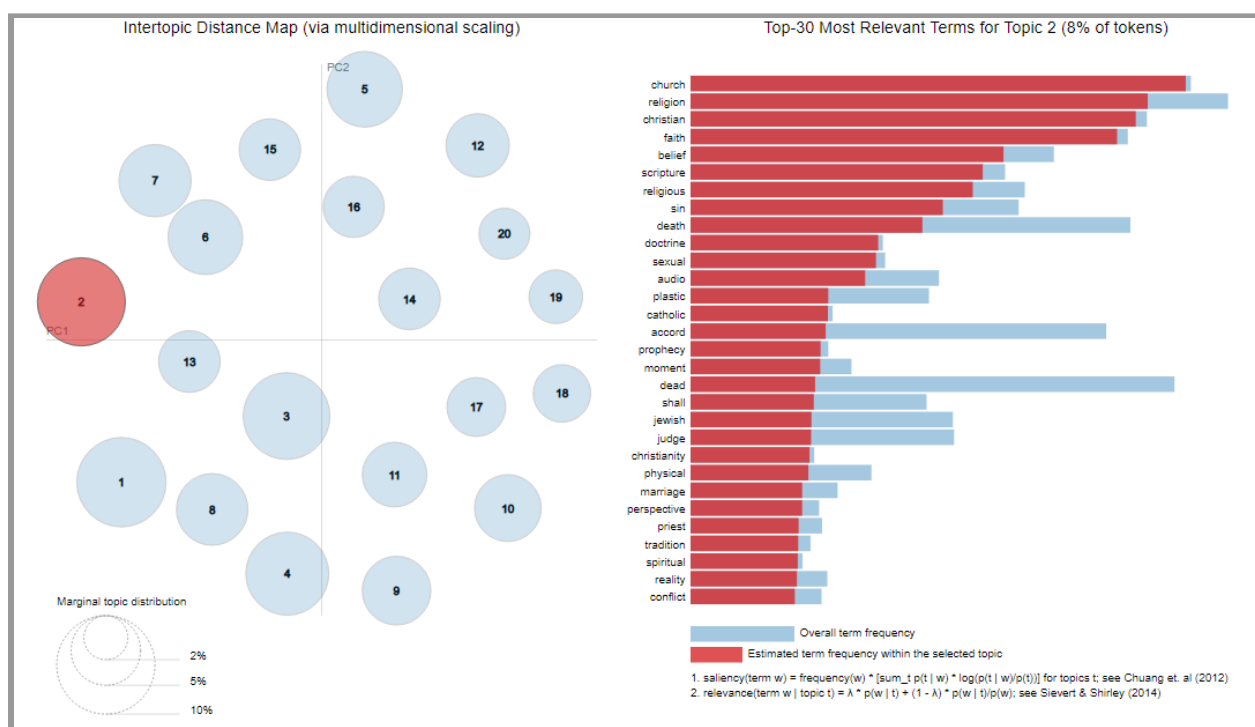


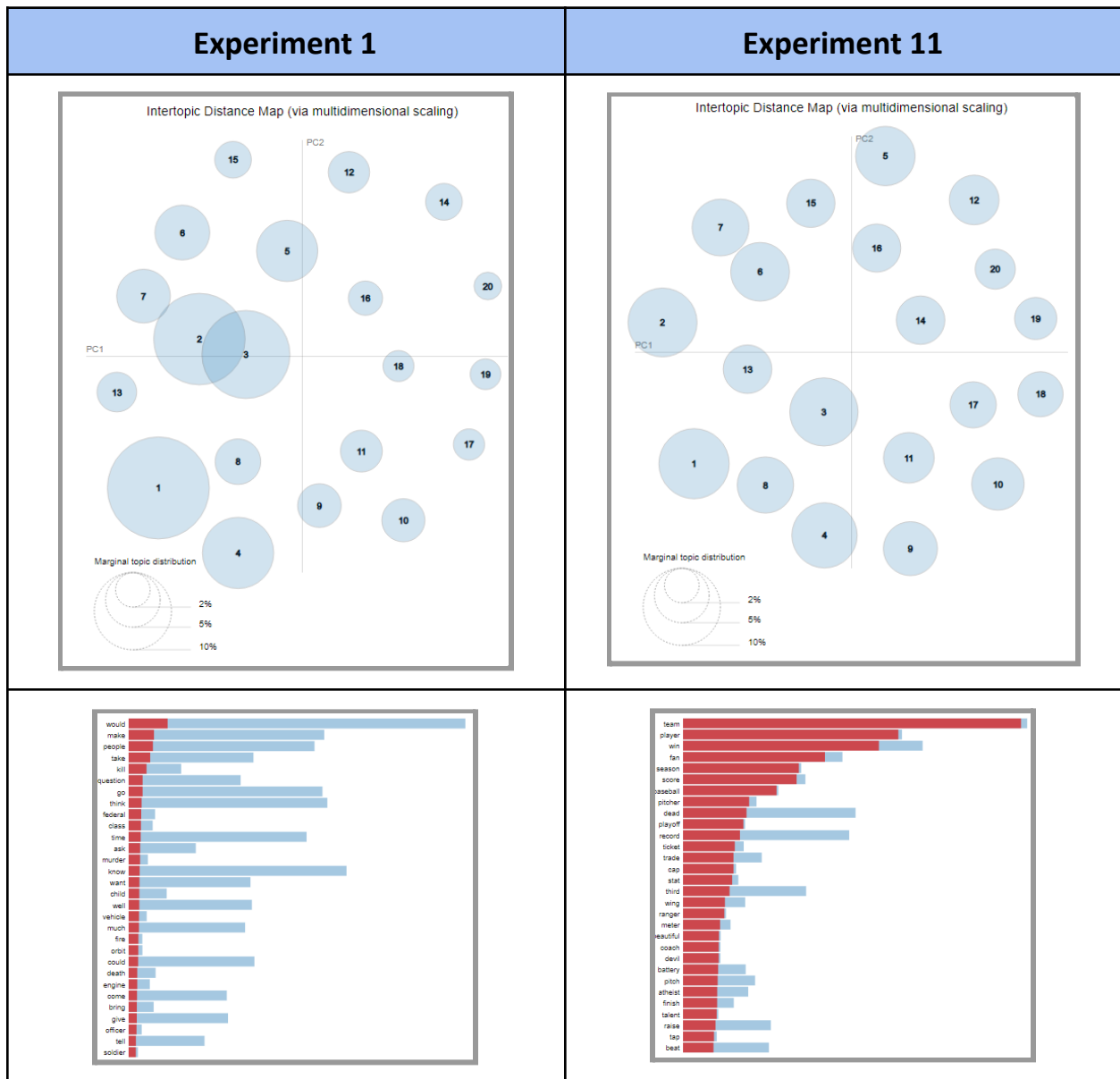
Experiment 11:

As of this experiment we had added a total of 1287 custom stopwords and were noticing topic convergence and common themes coming to the surface.

Topic Model Plot - Experiment 11

After Experiment 11, we are noticing significantly improved results from our initial model. Examining the topic model plot we can see that the topic circles shown in the “intertopic distance map” (left side of the plot) are significantly better separated with no overlap. Some topics such as 7,8 although with no overlap, are still fairly close.

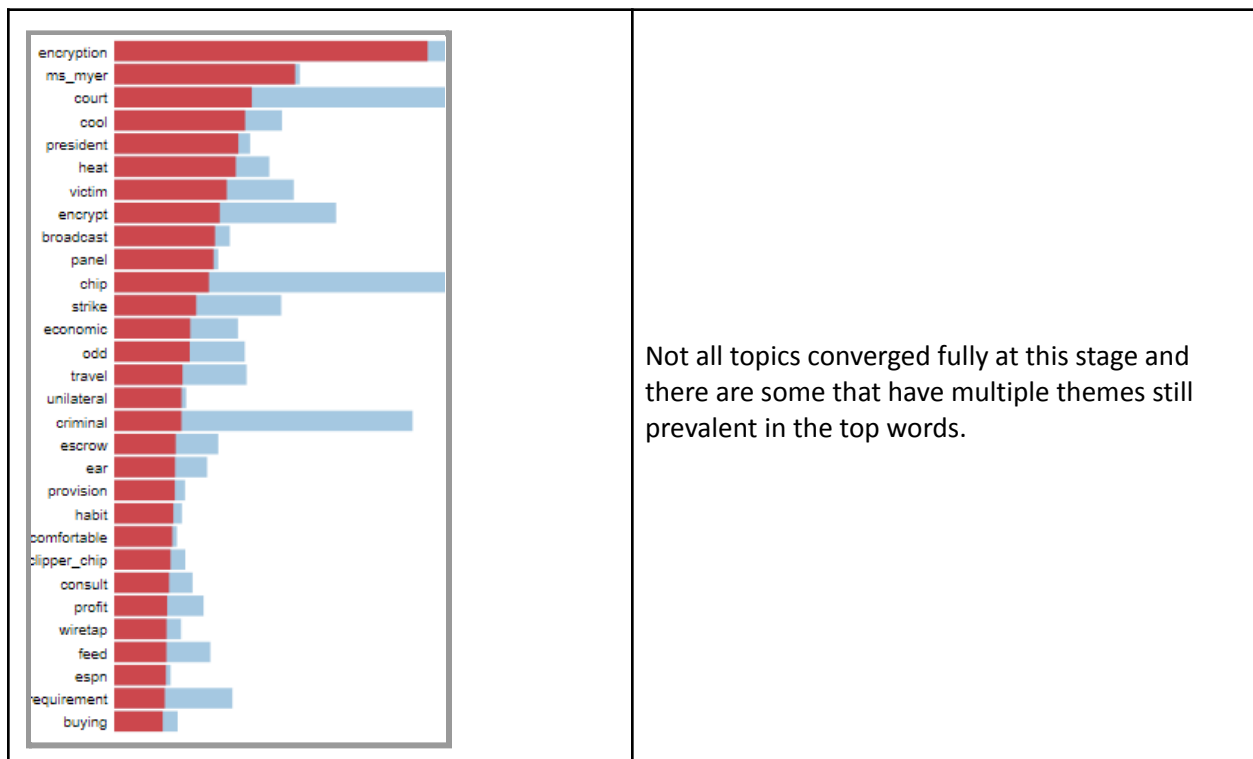
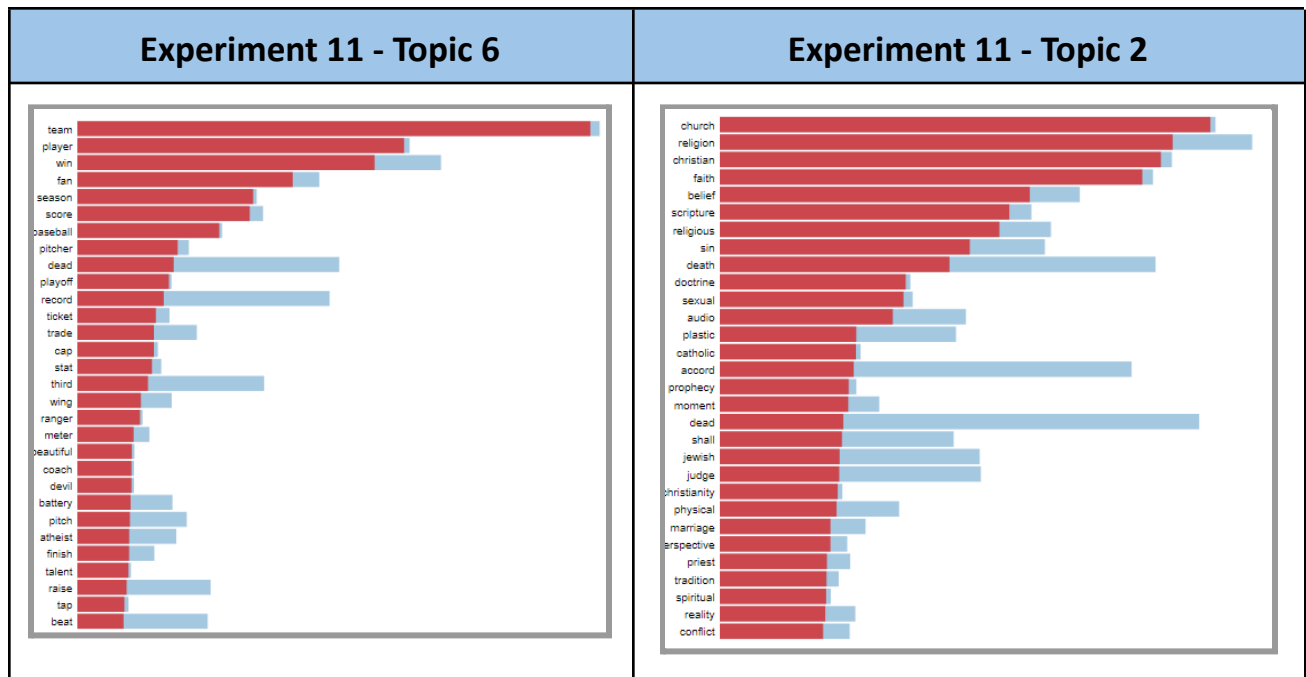




Comparing Experiment 1 with Experiment 11 results we notice the top 30 relevant terms are far more dominated by red (specific to the topic under consideration), and appear less frequently across the model (blue). This is a good indicator of topic convergence.




Additionally the top words for each topic are much stronger themed than in our initial models. For instance, Topic 2 has top words: church, religion, christian, faith, belief, scripture, religious, sin, death, doctrine. Words that thematically go together.

Compared to our initial model the most prevalent words for the topic are much less noticeable in other topics. All our bars are dominated by red with much less blue. Since our words are now becoming much more topic specific this is a positive direction for our modelling.



Word Frequency - Experiment 11




Examining the word frequency plots we obviously see the similar trend as shown in the topic model plot.

| Topic 15 | Topic 2 | Topic 14 |
|--|--|---|
|  <ul style="list-style-type: none">bugidealcryptographydataupgradeciphertextstreambytegeneratormegsaleencryptmanualhardwareinterfaceidebusmodemdiskscsi |  <ul style="list-style-type: none">electterroristslaughterparliamenttroopconvertibleindependentprofessionalmountainconstitutionalmurderinhabitantgreekofficerturkiyefirearmcitizenvillagemilitia |  <ul style="list-style-type: none">dotspiritpoorpuretargetserialminepeacehighwayconcludemphoilvacationsellmatchristianpriestsuspectrelationshippassage |

Bigrams - Experiment 11

We are noticing the topic themes emerging in a number of the topics. For example in Topic 9 the top most frequent words (beginning at the bottom) are “team win”, “team player”, “score team”, “fan fan”, “pitcher pitch”. We notice though that there is still some overlap in key bigrams. There is a clear sports theme but this topic includes both the baseball newsgroup (“pitcher pitch”) and the hockey newsgroup (“goal assist”).

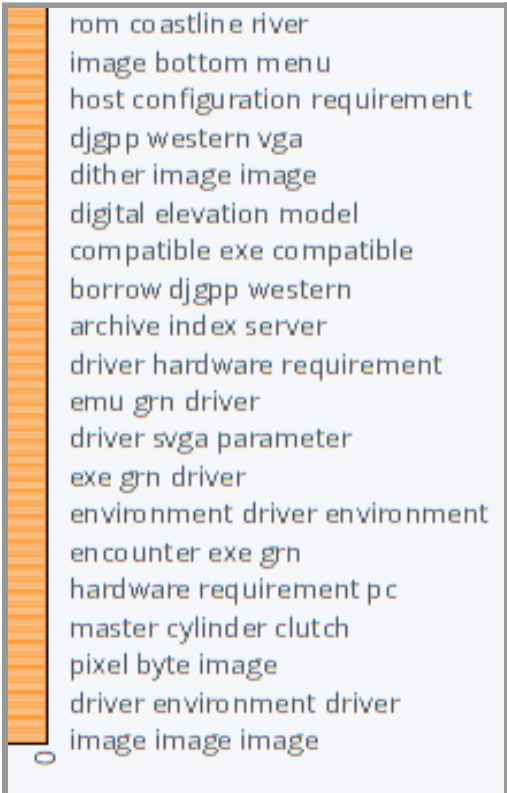
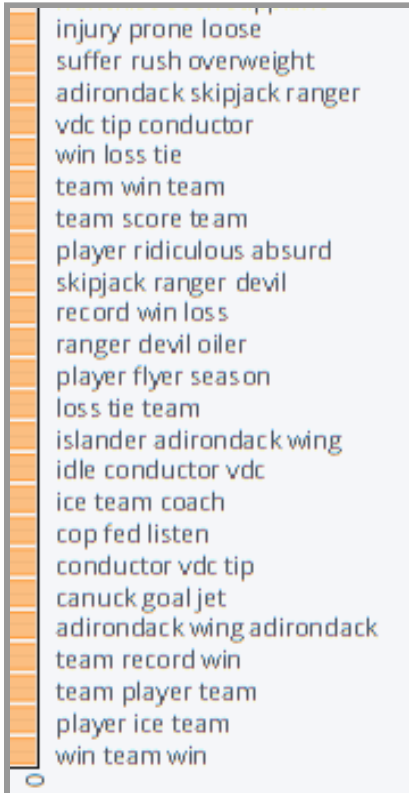
Topic 5 exhibits similar cohesion with a strong data theme “Image image”, “driver driver”, “gif image”, “byte image”. These are clearly technology themed, but appear to be coming from a number of different newsgroups, it also includes some that appear to belong to different newsgroups such as “scientific visualization”. There still appears to be a deal of topic overlap in here between various newsgroups dealing with technology, images, and science.


| Topic 14 | Topic 9 | Topic 5 |
|---|---|--|
|  <p>psalm dwell relationship female sexual relationship terrace foot wage burst abolish consent directory software definitely probe conclude strike buy potter bottom conclude adult relationship ledge rocky mangle death ethnic cleansing grievous sin financial backer guide arrest middle entrail passage passage</p> |  <p>loss tie division finish dead dead coach corner camera camera battery camera film cassette innocent murder ice stat hate team bike storage stat player goal assist score goal rod brind_amour pitcher pitch fan fan score team team player team win</p> |  <p>software image datum datum description image scientific visualization escape terminate palette image processing target image graphic borrow djgpp disk disk jpeg software datum image faq archive image pixel menu menu byte image shareware image gif image driver driver image image</p> |

Trigrams - Experiment 11

Similar to the bigrams, we are observing a convergence of keywords on a number of topics, but not all. Topic 9 for instance appears to have a clear sports theme and appears to be more dominated by hockey than the bigrams have potentially shown. There are some interesting trigrams in this Topic that don't initially appear to fit such as "cop fed listen".

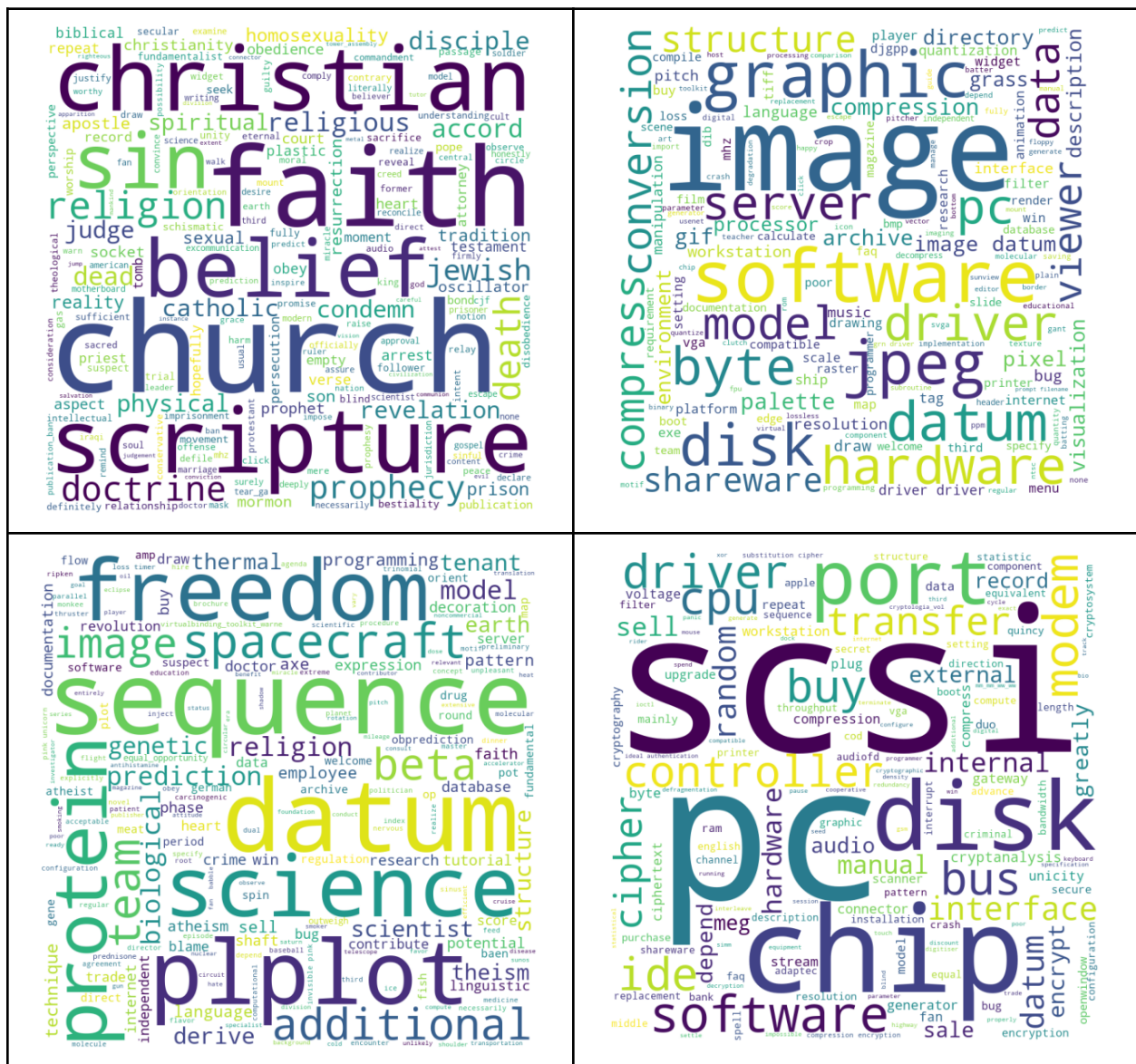
Topic 5 shows a strong theme of technology and images helping to identify where this topic may be converging. The 5th most prevalent trigram in this topic currently is "master cylinder clutch", which would be a more applicable phrase for the "rec.motorcycles" or "rec.auto" newsgroup than any of the technology ones.

| Topic 5 | Topic 9 |
|--|---|
|  <p>rom coastline river image bottom menu host configuration requirement djgpp western vga dither image image digital elevation model compatible exe compatible borrow djgpp western archive index server driver hardware requirement emu grn driver driver svga parameter exe grn driver environment driver environment encounter exe grn hardware requirement pc master cylinder clutch pixel byte image driver environment driver image image image</p> |  <p>injury prone loose suffer rush overweight adirondack skipjack ranger vdc tip conductor win loss tie team win team team score team player ridiculous absurd skipjack ranger devil record win loss ranger devil oiler player flyer season loss tie team islander adirondack wing idle conductor vdc ice team coach cop fed listen conductor vdc tip canuck goal jet adirondack wing adirondack team record win team player team player ice team win team win</p> |

| | |
|--|---|
|  <p> relevant datum science roll heart soul round round quiet science datum image theism theism favor touch possession obsession genetic algorithm genetic drag dirty hall danced zombie nervous coupon coupon coupon bam boom touch hall oate bam hooter nervous danced plant stir journey oate bam boom monkee monkee mediate journey raise suzanne identify relevant datum invisible pink unicorn </p> | <p>Not all topics showed convergence when examining the trigrams. Topic 7 showed a wide range of curious phrases. It's not immediately clear what newsgroup or theme is being invoked with "invisible pink unicorn" or "danced zombie nervous".</p> |
|--|---|

Newsgroup NLP Classification (ML1000-P3)

27

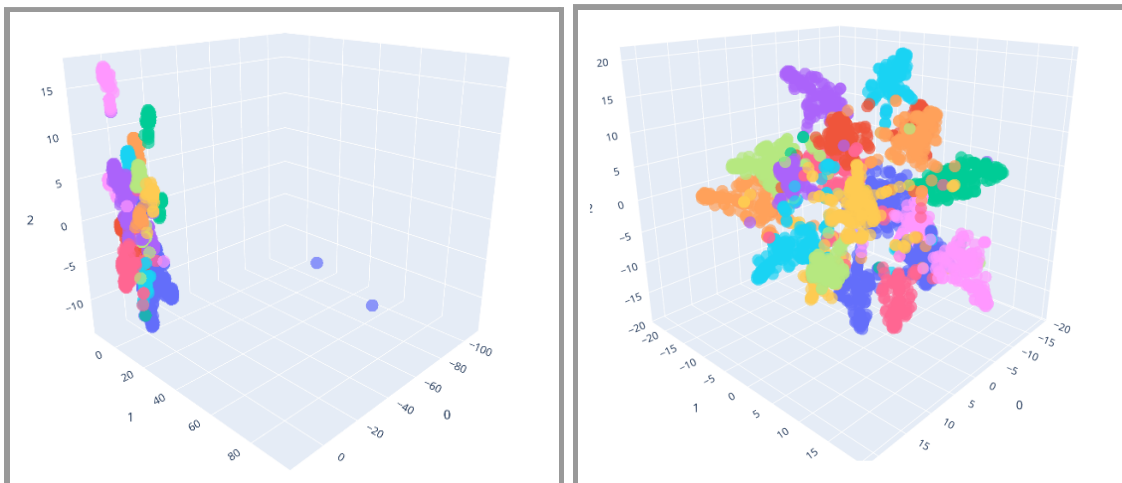


TSNE Plot - Experiment 11

Our t-distributed Stochastic Neighbor Embedding (t-SNE) plot showed significant improvement from our initial experiments. While initially the t-SNE (left) showed the topics bunched together with no differentiation, the t-SNE for experiment 11 (right) shows the topics are clustering and beginning to separate apart from each other.

Clusters such as the orange, or light blue show that while these topics are separating they haven't fully clustered as shown by their cluster separation on different sides of the "imaginary" centre of the plot.

Put another way, the new plot looks more like an exploding Death Star than a smear on a wall. This is good.



Ethical Considerations

The ethical considerations while creating semi-supervised models is rather significant. The choices one makes while choosing methodologies, subsets of data, pre-text processing, and stopwords has a significant impact on the model creation and the veracity of its conclusions.

We completed some experiments on different datasets and the default Pycaret NLP setup has a tendency to lemmatize/stem proper names. In an analysis we completed on a twitter dataset containing tweets from Donald Trump, the text processing had a tendency to get rid of words such as “Hillary”. This was a dominant topic which disappeared in the default Pycaret text processing. While this topic was unintentionally hidden by Pycaret, it would not be difficult to add “hillary” to a customized stopwords list and remove it from processing entirely. This would manually, and intentionally remove the topic from the dataset and analysis.

In our newsgroup analysis we encountered a wordcloud chart where some of the extremely dominant words were “jewish”, “israeli”, and “genocide”. While these are just the words that happen to bubble to the surface during our analysis, these are also words with a potentially heavy emotional context. It is certainly within the realm of possibility that were that wordcloud produced in a corporate setting, that its emotional weight and prevalence would be asked to be watered down.

The choices we make in our manual analysis and creation of the models come directly from our own insights, morality, and goals. While some people may believe that machine learning and AI are completely objective, this is not the case and it relies heavily on the person and direction in which the model is driven.

Conclusions

The newsgroup model has been improved dramatically by the implementation and subsequent tuning of the model through repeated analysis. We have seen some significant topic coherence in a number of areas but not all.

Additional experiments and changes to methodology to help tune and refine the model will greatly improve these results. There is still a great deal of opportunity for model gain by adding additional custom stopwords which would provide additional accuracy and a stronger representation in the t-SNE and other plots and charts.

Another option for improvement would be to introduce a noun-only approach for analysis (see Martin and Johnson) as that may provide an additional opportunity for an adjacent model to assist in classification. It would likely not be ideal to run a noun-only model through PyCaret, because as noted previously, there are not a lot of options to fine tune the text parsing and it has a tendency to remove important names.

Another area for exploration would be to combine the sentiment analysis as a supplementary component to see if there is any commonality in the sentiment compared to the topic. Using PyCaret's `tune_model` with a supervised component including sentiment analysis could help in this endeavour.

We were not able to complete our primary objective, which was to build a document classifier to import external articles into newsgroup discussion. The PyCaret NLP module doesn't provide a "predict_model" functionality that we could apply against fresh data for classification, but we could use the "assign-model" output to feed another classification process. That mechanism worked, but we didn't pursue it further, choosing instead to concentrate on exploring the topic modeling in more depth.

Bibliography

Prateek Baghel. Towards Data Science. NLP Text-Classification in Python: PyCaret Approach vs The Traditional Approach. URL: <https://towardsdatascience.com/nlp-classification-in-python-pycaret-approach-vs-the-traditional-approach-602d38d29f06>

pyLDAvis. UC Santa Barbara. URL: <https://we1s.ucsb.edu/research/we1s-tools-and-software/topic-model-observatory/tmo-guide/tmo-guide-pyldavis/>

S. Praphagaran. Machinelearningplus.com. URL: <https://www.machinelearningplus.com/nlp/topic-modeling-visualization-how-to-present-results-lda-models/>

Boostlabs.com. Word Clouds & the Value of Simple Visualizations. URL: <https://boostlabs.com/blog/what-are-word-clouds-value-simple-visualizations/>

Achinoam Soroker. Medium.com. T-SNE Explained Math and Intuition. URL: <https://medium.com/swlh/t-sne-explained-math-and-intuition-94599ab164cf>

Wikipedia. Bigram. URL: <https://en.wikipedia.org/wiki/Bigram>

Wikipedia. Latent Dirichlet Allocation (LDA). URL: https://en.wikipedia.org/wiki/Latent_Dirichlet_allocation

Priya Dwivedi. Towards Data Science. NLP: Extracting the main topics from your dataset using LDA in minutes. URL: <https://towardsdatascience.com/nlp-extracting-the-main-topics-from-your-dataset-using-lda-in-minutes-21486f5aa925>

Jason Brownlee. Machine Learning Mastery. A Gentle Introduction to the Bag-of-Words Model. URL: <https://machinelearningmastery.com/gentle-introduction-bag-words-model/>

F Martin, M. Johnson. More Efficient Topic Modelling Through a Noun Only Approach. URL: <https://aclanthology.org/U15-1013.pdf>

Kaggle. Newsgroup dataset. URL: <https://www.kaggle.com/crawford/20-newsgroups>

Appendix A: Custom Stopwords

Note the spelling errors below are intentional as there were multiple instances of badly spelled words that were impacting the model:

```
stopwords_R1 = ['use', 'say', 'would', 'need', 'know', 'also', 'get', 'thing',  
                'people', 'may', 'time', 'see', 'good', 'way', 'make', 'come',  
                'could', 'find', 'give', 'try', 'take', 'want', 'much', 'new'  
                ]
```

```
stopwords_R2 = ['therefore', 'set', 'month', 'actually', 'example', 'easy',  
                'hear', 'look', 'tell', 'even', 'become', 'important', 'include',  
                'think', 'well', 'probably', 'allow', 'low', 'follow', 'still',  
                'second', 'talk', 'cost', 'true', 'least', 'public', 'call',  
                'report', 'reason', 'send', 'sense', 'year', 'run', 'consider',  
                'base'  
                ]
```

```
stopwords_R3 = ['go', 'mean', 'list', 'really', 'order', 'new', 'little',  
                'lot', 'first', 'person', 'problem', 'thank', 'help', 'case',  
                'long', 'high', 'available', 'work', 'many', 'provide', 'right',  
                'read', 'home', 'therefore', 'individual', 'change', 'note',  
                'put', 'day', 'last', 'present', 'line', 'however', 'keep',  
                'article', 'happen', 'number', 'end', 'rather', 'lot', 'post',  
                'mind', 'part', 'slightly', 'fast', 'size', 'better', 'several',  
                'state', 'course', 'lead', 'easily', 'bad', 'issue', 'simply',  
                'level', 'rather', 'non', 'seem', 'hand', 'indicate', 'one',  
                'basis', 'total', 'else', 'assume', 'key', 'place', 'back',  
                'almost', 'begin', 'event', 'different', 'front', 'quite',  
                'program', 'stuff', 'small', 'reasonable', 'produce', 'mention',  
                'show', 'start', 'week', 'live'  
                ]
```

```
stopwords_R4 = ['throw', 'must', 'play', 'question', 'certainly', 'contact',  
                'always', 'copy', 'address', 'otherwise', 'belong', 'leave',  
                'definition', 'support', 'forget', 'save', 'local', 'design',  
                'meet', 'whole', 'box', 'stay', 'document', 'evidence', 'store',  
                'source', 'early', 'check', 'point', 'numerous', 'white',  
                'mail', 'suggest', 'notice', 'color', 'room', 'interested',  
                'sound', 'picture', 'already', 'effort', 'old', 'side', 'best',  
                'short', 'idea', 'red', 'perhaps', 'card', 'story', 'remove',
```

```

'possible', 'final', 'organization', 'family', 'city',
'interesting', 'expensive', 'business', 'ever', 'couple',
'file', 'big', 'choose', 'point', 'version', 'type', 'instead',
'life', 'effective', 'power', 'ask', 'gain', 'thus', 'name',
'remember', 'age', 'increase', 'device', 'city', 'difference',
'rate', 'less', 'fall', 'news', 'average', 'site', 'resource',
'subject', 'great', 'cheap', 'claim', 'cause', 'lie', 'man',
'side', 'actual', 'word', 'let', 'add', 'pay', 'whole', 'sure',
'stop', 'phone', 'child', 'control', 'claim', 'anti', 'never',
'tool', 'describe', 'hope', 'system', 'cause', 'standard',
'dream']
stopwords_R5 = ['cut', 'accept', 'clock', 'book', 'far', 'move', 'tape', 'care',
'express', 'guy', 'particular', 'yank', 'apply', 'action',
'object', 'bit', 'refer', 'enough', 'wrong', 'email', 'sign',
'answer', 'commercial', 'lose', 'appear', 'break', 'large',
'folk', 'fee', 'purpose', 'cover', 'pretty', 'wide', 'limit',
'full', 'piece', 'value', 'far', 'effect', 'supply', 'comment',
'exist', 'useful', 'major', 'real', 'argument', 'especially',
'group', 'money', 'package', 'watch', 'away', 'discussion',
'wish', 'kind', 'write', 'real', 'suppose', 'fail', 'believe',
'please', 'sorry', 'unknown', 'late', 'feel', 'friend', 'fact',
'understand', 'indeed', 'particular', 'obvious', 'difficult',
'truth', 'error', 'maybe', 'fix', 'specifically', 'determine',
'disagree', 'necessary', 'curious', 'worry', 'paper', 'eat',
'learn', 'risky', 'occur', 'quality', 'approximately', 'kid',
'test', 'compare', 'continue', 'extremely', 'able', 'concern',
'perform', 'today', 'recently', 'strong', 'regard', 'clearly',
'explain', 'nice', 'similar', 'com', 'answer', 'relate',
'company', 'telephone']
stopwords_R6 = ['speculation', 'seize', 'widespread', 'conceal', 'legitimate',
'discourage', 'hotel', 'readily', 'cooperation', 'communication',
'asset', 'demand', 'measurement', 'adopt', 'etc', 'analyst',
'facility', 'progress', 'tense', 'culmination', 'danger',
'wonderful', 'somewhat', 'ensue', 'excited', 'realise', 'deserve',
'thin', 'circumstance', 'unbelievable', 'scare', 'barrier',
'smoke', 'modification', 'shed', 'freely', 'earlier', 'require',
'press', 'interest', 'purely', 'adjective', 'typically', 'town',
'norm', 'demonstration', 'occasion', 'drink', 'imperative',
'invoke', 'joined', 'bene', 'decade', 'respective',
'association', 'convinced', 'confusing', 'ap', 'plead', 'composite',
'expensive', 'buf', 'reduction', 'complement', 'flat', 'beginning',
'compliant', 'controler', 'approve', 'successful', 'appreciative',
'distorted', 'pffffffttttt', 'world', 'suggestion', 'previous',

```

'select', 'information', 'strongly', 'define']

```
stopwords_R7 = ['simple', 'window', 'posting', 'receive', 'class', 'message',  
    'wonder', 'reference', 'normal', 'multiple', 'experience',  
    'style', 'item', 'build', 'closer', 'personality', 'advertise',  
    'routinely', 'espouse', 'grill', 'unfounded', 'administration',  
    'form', 'theory', 'drop', 'law', 'general', 'board', 'fantasy',  
    'term', 'involve', 'respect', 'organize', 'window', 'machine',  
    'current', 'create', 'view', 'command', 'worth', 'shit', 'bring',  
    'believable', 'tend', 'later', 'face', 'fear', 'sample',  
    'avoid', 'process', 'namely', 'correct', 'guess', 'page',  
    'release', 'analysis', 'matter', 'willingness', 'return',  
    'surprising', 'girl', 'open', 'stand', 'contain', 'return',  
    'straight', 'context', 'beg', 'brand', 'absolute', 'act', 'cry',  
    'opinion', 'reading', 'ok', 'view', 'quote', 'yesterday',  
    'reply', 'clear', 'refute', 'meaning', 'insult', 'statement',  
    'solution', 'correctly', 'likely', 'nature', 'human',  
    'population', 'male', 'interpret', 'recognize', 'together',  
    'laugh', 'like', 'excellent', 'ca', 'expert', 'usage', 'reduce',  
    'condition', 'speak', 'experience', 'junk', 'plan', 'human',  
    'speak', 'distinction', 'needless', 'minute', 'contain', 'half',  
    'attempt', 'frame', 'light', 'receive', 'theory', 'reference',  
    'obsessive', 'assemble', 'count', 'frequently', 'sake', 'attempt',  
    'knowledge', 'elsewhere', 'ax', 'max', 'gv_gv_gv_gv', 'net']
```

```
stopwords_R8 = ['info', 'screen', 'install', 'study', 'author', 'miss', 'chance',  
    'publish', 'result', 'fund', 'block', 'material', 'letter',  
    'attack', 'account', 'party', 'text', 'approach', 'measure',  
    'represent', 'turn', 'station', 'speed', 'dog', 'head', 'close',  
    'charge', 'lock', 'fairly', 'single', 'chair', 'agree',  
    'objective', 'often', 'hang', 'observation', 'delete',  
    'practice', 'personal', 'false', 'explanation', 'situation',  
    'reject', 'oppose', 'head', 'argue', 'woman', 'land', 'member',  
    'mother', 'body', 'remain', 'history', 'shout', 'burn', 'entire',  
    'access', 'service', 'obsolete', 'attend', 'miss', 'school',  
    'detail', 'improvement', 'register', 'electronic', 'hard',  
    'eye', 'pick', 'night', 'marry', 'rule', 'sit', 'building', 'fine',  
    'career', 'close', 'morning', 'soon', 'tv', 'apparently', 'primary',  
    'protect', 'self', 'prove', 'sort', 'option', 'insist', 'obtain',  
    'deny', 'scheme', 'neighbor', 'yet', 'maintain', 'rule',  
    'apartment', 'door', 'volume', 'position', 'intend', 'title',  
    'next', 'top', 'willing', 'share', 'job', 'associate', 'hold',  
    'teaching', 'future', 'rule', 'thought', 'response', 'particularly',
```

'location', 'obviously', 'grant', 'free', 'fit', 'typical',
 'colour', 'boy', 'factor', 'speaker', 'original', 'unit', 'price',
 'extension', 'decision', 'private', 'stage', 'assumption',
 'usually', 'push', 'date', 'assertion', 'prevent', 'coverage',
 'house', 'mix', 'love', 'pressure', 'ground', 'hour', 'hard',
 'fire', 'hot', 'water', 'die', 'student', 'prize', 'advantage',
 'bathroom', 'food', 'pound', 'exchange', 'ago', 'connection',
 'rock', 'explicit', 'hold', 'gift', 'trouble', 'damn', 'leaf',
 'deliberately', 'gift', 'investigate', 'result', 'certain',
 'figure', 'significant', 'inflamm', 'shape', 'inform', 'reserve',
 'height', 'batf', 'currently', 'access', 'associate', 'completely',
 'offer', 'turn', 'job', 'decision', 'discuss', 'fair', 'mirror',
 'space', 'observation', 'hole', 'black', 'enter', 'conference',
 'application', 'majority', 'security', 'perfectly', 'community',
 'risk', 'truly', 'self', 'special', 'prefer', 'perfect', 'common',
 'century', 'library', 'waste', 'rare', 'expose', 'bear', 'lay',
 'shortly', 'hideous', 'television', 'sat', 'practically',
 'ownership'
]

stopwords_R9 = ['dragon', 'evaluate', 'pain', 'principle', 'deep', 'sometimes',
 'subjective', 'generally', 'treat', 'arithmetic', 'blood',
 'phenomenon', 'discover', 'stress', 'interpretation', 'mission',
 'launch', 'hearing', 'around', 'experiment', 'degree', 'legal',
 'rest', 'collect', 'develop', 'display', 'convert', 'directly',
 'input', 'feature', 'government', 'user', 'voice', 'patent',
 'technical', 'industry', 'project', 'hide', 'request', 'topic',
 'daughter', 'proposal', 'fellow', 'table', 'slow', 'instal',
 'appreciate', 'wait', 'product', 'path', 'kill', 'force',
 'commit', 'serve', 'arm', 'debate', 'funny', 'imply', 'merely',
 'animal', 'demonstrate', 'anymore', 'prepare', 'defense', 'review',
 'copyright', 'rational', 'restriction', 'trip', 'expect',
 'exhaust', 'decide', 'brother', 'afraid', 'pass', 'admit',
 'tree', 'wife', 'doubt', 'refuse', 'longer', 'suddenly', 'hardly',
 'coat', 'grow', 'mark', 'phrase', 'longer', 'reader', 'addition',
 'separate', 'ability', 'introduction', 'parent', 'distance', 'deal',
 'poster', 'anyway', 'replace', 'risk', 'market', 'advice', 'alone',
 'society', 'rise', 'stone', 'brave', 'decide', 'spec', 'pad',
 'blow', 'confuse', 'around', 'develop', 'game', 'fault', 'ahead',
 'achieve', 'hence', 'method', 'mile', 'registration', 'practical',
 'absolutely', 'pre', 'manner', 'catalog', 'lift', 'lean', 'warm',
 'instruction', 'drive', 'remote', 'faster', 'function', 'string',
 'exit', 'band', 'eventually', 'license', 'safety', 'search',

```

'delay', 'load', 'flame', 'switch', 'smooth', 'reach', 'reverse',
'intrinsic', 'client', 'technology', 'area', 'monitor',
'appropriate', 'development', 'subsidize', 'various', 'generally',
'field', 'specific', 'carry', 'compromise', 'network', 'energy',
'evening', 'double', 'inflatable', 'assert', 'pass'
]

```

```

stopwords_R10 = ['pen', 'policy', 'finger', 'node', 'tone', 'performance', 'pack',
'shift', 'dial', 'disable', 'complete', 'character', 'manager',
'bar', 'speech', 'code', 'handle', 'format', 'previously', 'domain',
'authority', 'conclusion', 'misuse', 'serious', 'uncomfortable',
'normally', 'prior', 'prepared', 'sentence', 'dome', 'novelty',
'clarify', 'aid', 'recommend', 'seriously', 'ad', 'laughter',
'country', 'conversation', 'controllable', 'assume', 'output',
'computer', 'connect', 'management', 'entry', 'damage', 'capable',
'immediately', 'inne', 'owner', 'deficit', 'choice', 'hill',
'schism', 'recognition', 'respond', 'rightful', 'bean', 'main',
'traveler', 'relative', 'range', 'exam', 'kit', 'million',
'percent', 'improve', 'minimum', 'weekend', 'selective', 'due',
'mostly', 'extra', 'pull', 'multi', 'postage', 'capability',
'limited', 'abolished', 'complexity', 'superior', 'regardless',
'sex', 'trace', 'outstanding', 'hundred', 'analogy', 'resemble',
'gp', 'wac', 'spot', 'sad', 'traditional', 'nonsense', 'inner',
'basically', 'compulsive', 'dangerous', 'wear', 'fi', 'balcony',
'echo', 'movie', 'precision', 'mumble', 'determination', 'fancy',
'crush', 'appreciated', 'unfortunately', 'memory',
'considerably', 'fluid', 'link', 'manager', 'signature',
'agency', 'join', 'assistance', 'button', 'role', 'operation',
'activity', 'initial', 'unlimited', 'update', 'propose',
'personally', 'default', 'implement', 'inference', 'destroy',
'core', 'depth', 'basically', 'summer', 'task', 'laughter',
'announce', 'heavy', 'blue', 'anonymous', 'fallacy', 'conclusion',
'premise', 'valid', 'null', 'remark', 'proof', 'logic',
'stupid', 'hit', 'amount', 'complete', 'distribution', 'mode',
'expense', 'minority', 'basic', 'somewhere', 'opposite', 'property',
'abstract', 'deletion', 'distribute', 'plausible', 'due', 'ton',
'rear', 'teach', 'surface', 'luck', 'young', 'imagine', 'stock',
'joke', 'thereby', 'tough', 'recent', 'bother', 'possibly',
'reliable', 'link', 'establish', 'pin', 'thief', 'slice', 'leather',
'enforcement', 'amount', 'operate', 'step', 'though', 'tounge',
'pant', 'craft', 'recent', 'sky', 'comprehensive', 'steal',
'publicly', 'vendor', 'normally', 'inquiry', 'exactly', 'branch',
'medium', 'section', 'cook', 'export', 'sentence', 'ariable',

```

'instrument', 'branch', 'pop', 'insert', 'contradiction', 'uccxkvb',
'folder', 'destroy', 'operation', 'iniquity', 'hurt', 'fat',
'mov', 'retail', 'environmental', 'anywhere', 'car', 'vote',
'automatic', 'choice', 'pub', 'routine', 'consist', 'utility',
'distribute', 'modify', 'award', 'air', 'video', 'excuse',
'sleep', 'rich', 'plenty', 'bounce', 'violation', 'rough',
'finally', 'outside', 'constraint', 'flesh', 'rip', 'executive',
'consist', 'attention', 'characteristic', 'commonly', 'entropy',
'operate', 'nearly', 'connect', 'guarantee', 'relative', 'prior',
'helpful', 'correlation', 'trust', 'equally', 'lack', 'dollar',
'center', 'step', 'bill', 'cable', 'minority', 'decent', 'guard',
'proceed', 'generalization', 'flaw', 'emotional', 'accident',
'social', 'partner', 'proceed', 'aware', 'yell', 'calm', 'official',
'broad', 'air', 'mamma', 'professor', 'ignorance', 'variety',
'print', 'main', 'rectangle', 'impact', 'radio', 'section',
'excessive', 'finally', 'club', 'setup', 'lack', 'alternative',
'ring', 'observable', 'schism', 'existence', 'polygon',
'nearly'
]