MEng Project Report Fall 2019
FCC rulemaking Comment Clustering

Team members - Kaveesha Shah (ks2379), Larissa Pereira (lp445)
Advisor - Sara Frug (Cornell Law School)

1. Description of the original problem

The U.S. Federal Communications Commission (FCC) considered regulations regarding net neutrality – the principle that the Internet Service Providers must treat all data the same, regardless of the origin or purpose of that data.Most FCC rules are adopted by a process known as "notice and comment" rulemaking. Under that process, the FCC gives the public notice that it is considering adopting or modifying rules on a particular subject and seeks the public's comment. As per the statistics from April 27 to Aug. 30, 2017, 21.7 million comments were submitted electronically and posted online for review. The FCC system could not handle this large volume of comments and it crashed.

Analyses of these 21.7 million comments submitted to the Federal Communications Commission (FCC) regarding its proposed repeal of net neutrality regulations are indicative of spam comments.The most well-known form of spamming is the fake negative comment generation by bots employed by agencies that don't want a particular rule to be in existence or too many support comments in case they want those rules to come into existence. This makes it very difficult for the FCC as they have to identify which of those comments are legitimate and which ones are spam.

2. Solution

We took the aid of the techniques used in the domain of Natural Language processing to help us solve this problem. We do not not have any sort of labelled trained data readily available that has pre classified certain comments as spam and not spam that can be learned by a classifier to appropriately predict the burst of new incoming comments and label them accurately. Hence this problem cannot be solved using the supervised Machine Learning techniques. We need to use unsupervised modelling techniques to analyze this data. We have used clustering techniques to group similar comments together.

We took inspiration from Jeff Kao's work on the same. We had a dataset of the comments available publicly (https://www.kaggle.com/jeffkao/proc_17_108_unique_comments_text_dupe_count) for the project.

The primary techniques implemented in this project are elucidated below --

1. Word to vector conversion

    The Machine Learning and Natural Language processing techniques are incapable of processing data in raw text format and deriving knowledge from them for processing it further. We therefore need a numerical representation of each individual word so that it can be used to extract important information.The mathematical representation of a word is expressed as a vector. We need to have a sense of similarity and difference between the words so that it can be used as an input to our clustering algorithm . We can exploit concept of vectors and distances between them (Cosine, Euclidean ,Manhattan etc.) to find similarities and differences between words.

    To implement this we have used Spacy's vector attribute which will convert a word into its vector form. Spacy uses the Glove 300 dimension embedding for encoding.

2. Clustering HDBScan

HDBScan stands for Hierarchical Density based Scan. We have used this extension of DBScan for the purpose of this project. It extends DBSCAN by converting it into a hierarchical clustering algorithm, and then using a technique to extract a flat clustering. The first step of this algorithm is to transform the space according to the density using the maximum reachability distance. Now that we have a mutual reachability metric on the data, we create a minimum spanning tree using Prim's algorithm on that data. Given the minimal spanning tree, the next step is to convert that into the hierarchy of connected components. This is most easily done in the reverse order: sort the edges of the tree by distance (in increasing order) and then iterate through, creating a new merged cluster for each edge.The first step in cluster extraction is condensing down the large and complicated cluster hierarchy into a smaller tree with a little more data attached to each node. Once we have a value for minimum cluster size we can now walk through the hierarchy and at each split ask if one of the new clusters created by the split has fewer points than the minimum cluster size. Here our data points will be the comments that we are supposed to cluster and they will represented in the feature space as the average of the word embeddings of the words in the comment. Spacy library handles all the preprocessing part like tokenization, lemmatization, removing stop words etc. We have supplied the minimum clusters parameter as 5. For our project, we got 79 clusters. We judge the effectiveness of the clusters using silhouette scores. The Silhouette Coefficient is calculated using the mean

intra-cluster distance (a) and the mean nearest-cluster distance (b) for each sample. The Silhouette Coefficient for a sample is (b - a) / max(a, b). To clarify, b is the distance between a sample and the nearest cluster that the sample is not a part of. Note that Silhouette Coefficient is only defined if number of labels is 2 <= n_labels <= n_samples - 1. The best value is 1 and the worst value is -1. Values near 0 indicate overlapping clusters. Negative values generally indicate that a sample has been assigned to the wrong cluster, as a different cluster is more similar.

3. Google Collaboratory

Because of the infrastructure challenges, we decided to switch to a cloud platform with GPU functionality available and Google Collaboratory fulfilled all the requirements and hence we chose that. We used the Google File Stream Transfer to ensure faster retrieval of the huge data file to be read. We were able to reduce the reading time from more than 5 min on local machine to 45 seconds via Google Collaboratory.

3. Experiments

1. We tried to use Spacy[Cuda] which is the GPU version of the Spacy library
2. We used the below Word to vector conversion methods-
    a. First vectorise each word in a sentence and then subsequently find the mean for the entire sentence
    b. Directly use the sentence.vector method to implicitly calculate mean of the sentence
3. Tried to implement multi-threading for parallel processing of encode_doc_vectors method that was consuming the most amount of time.
4. In order to improve performance we experimented with storing the word vectors in a python dictionary but that caused an overhead on the memory due to the large number of unique words.

. 4. Benchmarks

The legacy system took approximately 5 hours to process 10000 comments. This made the system very inefficient for it to be used on real data and especially ineffective to handle the large volume of comments that came in. We have achieved the below benchmarks while working on this project-

Size of Dataset	Time taken
1000	32.4 sec
5000	2mins 43secs
10000	4mins 35secs
50000	22mins 4secs
100000	43mins 19secs
200000	2 hrs 24 mins 57 secs
500000	6 hrs 48 mins 36 secs

From the above benchmarks we see that we were able to process 10000 comments in 4 mins 35 seconds which is much better in terms of performance compared to the legacy system. We further ran the code by iteratively increasing the number of comments to observe the time taken and we achieved a significant performance of 43 mins 19 secs for 100000 comments.

5. Results

Below are the first few clusters that we have observed after running the clustering algorithm on the provided dataset.

Cluster 1 -

1	197626	I STRONGLY support net neutrality backed by Title 2 oversight of ISP's!!!!!!!! We must maintain net neu
1	271111	Keep the ISPs under title II FCC !!!!!!!!!!!!!!!
1	421732	I support STRONG net neutrality backed by TITLE II oversight!!!!!
1	762511	Preserve net neutrality and title 2!!!!!!!!!!!!!!!!!!!!
1	1084740	PRESERVE NET NEUTRALITY AND TITLE TWOOOOOOO!!!!!!!!!!!!
1	1408460	Save net neutrality!!!!!
1	7065117	Protect net nutrality!!!!!
1	7571553	I DO NOT want net neutrality rules repealed!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!I think it is a bad idea to repeal t
1	8462720	Ending Net Neutrality will lead to MASSIVE CENSORSHIP via the VERY SLOW LANE !!!!!!!!!!!!!
1	2.2E+07	Leave our internet alone !!!!!!!!

In the first cluster we observed unpromising results because the algorithm has clusters these comments into the first cluster based the commonality in the

comments which is that of the number of exclamation marks used in all of these comments.

Cluster 2-

2	2560645	I am in favor of strong net neutrality under Title II of the Telecommunications Act.
2	2560829	I am in favor of strong net neutrality under Title II of the Telecommunications Act.Sincerely,Kristine Pegg
2	2562706	I am in favor of strong net neutrality under Title II of the Telecommunications Act.Sincerely,Phillip Garcia
2	2566727	I am in favor of strong net neutrality under Title II of the Telecommunications Act.Sincerely,Christina Case
2	2566933	I am in favor of strong net neutrality under Title II of the Telecommunications Act.Sincerely,Timothy Rhine
2	2578229	I am in favor of strong net neutrality under Title II of the Telecommunications Act.Sincerely,Joseph Olsen
2	2578531	I am in favor of strong net neutrality under Title II of the Telecommunications Act.Sincerely,William Silverman
2	2578684	I am in favor of strong net neutrality under Title II of the Telecommunications Act.Sincerely,Laverne Campbell
2	2578870	I am in favor of strong net neutrality under Title II of the Telecommunications Act.Sincerely,Laura Maule
2	2578933	I am in favor of strong net neutrality under Title II of the Telecommunications Act.Sincerely,Jose Baumer
2	2579528	I am in favor of strong net neutrality under Title II of the Telecommunications Act.Sincerely,Thelma Michalak
2	2580075	I am in favor of strong net neutrality under Title II of the Telecommunications Act.Sincerely,Willie Holton
2	2580394	I am in favor of strong net neutrality under Title II of the Telecommunications Act.Sincerely,Robert Fyfe
2	2580673	I am in favor of strong net neutrality under Title II of the Telecommunications Act.Sincerely,David Petrovich
2	2581080	I am in favor of strong net neutrality under Title II of the Telecommunications Act.Sincerely,Billy Kim
2	2581193	I am in favor of strong net neutrality under Title II of the Telecommunications Act.Sincerely,Audrey Griffith
2	2582733	I am in favor of strong net neutrality under Title II of the Telecommunications Act.Sincerely,Carla Ramirez
2	2583203	I am in favor of strong net neutrality under Title II of the Telecommunications Act.Sincerely,Eugene Saville
2	4400504	I am in favor of strong net neutrality under Title II of the Telecommunications Act.Sincerely,Kelly Christensen
2	8923384	I am in favor of strong net neutrality under Title II of the Telecommunications Act.Sincerely,Amanda Koss
2	9147480	I am in favor of strong net neutrality under Title II of the Telecommunications Act.Sincerely,Paul Folden
2	9148785	I am in favor of strong net neutrality under Title II of the Telecommunications Act.Sincerely,Brenda McBryde
2	9151748	I am in favor of strong net neutrality under Title II of the Telecommunications Act.Sincerely,Erin Hutchins
2	9372877	I am in favor of strong net neutrality under Title II of the Telecommunications Act.Sincerely,Rosalie Jefferson
2	9383830	I am in favor of strong net neutrality under Title II of the Telecommunications Act.Sincerely,Patricia Bryan
2	9400991	I am in favor of strong net neutrality under Title II of the Telecommunications Act.Sincerely,Virginia Lee
2	9463473	I am in favor of strong net neutrality under Title II of the Telecommunications Act.Sincerely,Sharon Cobos
2	9516727	I am in favor of strong net neutrality under Title II of the Telecommunications Act.Sincerely,Dorothy Calle
2	9656591	I am in favor of strong net neutrality under Title II of the Telecommunications Act.Sincerely,Willie Johnston

As observed in the results of the cluster 2 the algorithm seems to have appropriately identified all the comments with similar content and these differ only by the name in the signature. It is very likely that these comments were generated by a non- human bot that copied the same content and only changed the names in the signature section of the comment. These kind of clusters are very useful to identify spam and will be able to provide useful information to further analyse the number of such comments.

Cluster 9-

9	4704322	Chairman Pai: My comments re: NET NEUTRALITY. I would like to encourage the FCC to repeal President Obama's power grab to control Internet access. Individuals, rather than the FCC, should be empowered to buy the services they
9	4706306	To the FCC: I would like to comment on the FCC's Open Internet order. I would like to suggest the FCC to reverse Obama's power grab to regulate Internet access. Citizens, as opposed to unelected bureaucrats, should be able to purch
9	4706694	To the FCC: In reference to Net neutrality. I want to advocate the commission to undo President Obama's order to take over the web. Internet users, not unelected bureaucrats, should be empowered to enjoy the applications they w
9	4707709	FCC commissioners, In the matter of net neutrality. I would like to ask the government to undo President Obama's scheme to regulate the Internet. Individual Americans, not unelected bureaucrats, should be able to enjoy whatever p
9	4708632	FCC: In the matter of internet regulations. I strongly implore you to rescind The Obama/Wheeler power grab to take over the web. Citizens, not the FCC Enforcement Bureau, should be able to select whichever products they choose. T
9	4711082	Dear Mr. Pai, I'm very concerned about Title 2 and net neutrality. I would like to urge the Federal Communications Commission to undo Obama's decision to control Internet access. Citizens, not Washington bureaucrats, should be en
9	4713092	Dear FCC, I want to give my opinion on the FCC regulations on the Internet. I would like to demand the FCC to reverse Obama's power grab to control broadband. People like me, not the FCC, should select whichever products we pref
9	4716546	Chairman Pai: I'm concerned about the FCC rules on the Internet. I recommend Chairman Pai to repeal President Obama's policy to control Internet access. Americans, rather than the FCC Enforcement Bureau, should be empowered
9	4717513	To whom it may concern: I have concerns about Title II rules. I strongly implore the FCC to undo Obama's plan to control Internet access. Internet users, as opposed to Washington bureaucrats, should buy the products we desire. Oba
9	4717728	To whom it may concern: I am a voter worried about the FCC's so-called Open Internet order. I encourage the Federal Communications Commission to reverse The previous administration's plan to regulate the web. Americans, as op
9	4718405	Dear Commissioners: I'd like to share my thoughts on net neutrality. I would like to urge the commission to undo President Obama's plan to regulate broadband. Internet users, as opposed to Washington bureaucrats, deserve to purch
9	4719167	Dear Mr. Pai, Hi, I'd like to comment on Title II rules. I strongly suggest Chairman Pai to repeal Tom Wheeler's order to take over Internet access. Individual Americans, not unelected bureaucrats, ought to enjoy whatever applications
9	4722050	I have concerns about Network Neutrality. I strongly encourage the government to repeal The previous administration's plan to regulate the Internet. Citizens, not unelected bureaucrats, should purchase whichever products we prefer
9	4723494	Chairman Pai: I am concerned about internet regulations. I would like to suggest you to overturn The previous administration's power grab to regulate Internet access. People like me, not Washington bureaucrats, should be free to pu
9	4724752	Chairman Pai: My comments re: Internet freedom. I'd like to demand Ajit Pai to overturn Barack Obama's power grab to regulate Internet access. Individual citizens, as opposed to big government, should be able to use whichever pro
9	4728164	Dear Chairman Pai, I would like to comment on the FCC rules on the Internet. I request the commissioners to overturn Obama's policy to control broadband. Citizens, not so-called experts, should be empowered to select which produ
9	4731914	My comments re: an open Internet. I'd like to demand Ajit Pai to repeal Tom Wheeler's scheme to regulate the Internet. Individual citizens, rather than Washington, ought to purchase whatever applications they desire. Tom Wheeler's

The results produced in cluster 9 by the clustering algorithm are indeed promising where the algorithm has not only identified comments based on the similarity of punctuation or the content but by the semantic meaning of the different words used in the comments and has grouped the comments into this cluster by taking into account the context in which different words are used and the similarity between them.

6. Challenges

a. We initially started with using the small version of the en_core_web_md module provided by spacy for word to vector conversion but we encountered a problem where every word was represented with a zero vector since the module didn't have the representation for certain words seen in the corpus of the large number of comments. This was resolved by using the medium version of the same en_core_web_md module.
b. Large volume of the dataset cause difficulty with file loading and its subsequent processing.
c. The method "encode_docs_vec" that converts words to their corresponding vectors takes a long time to complete running.
d. Computational capacity of our local infrastructure was one of the greatest challenges we faced. We therefore considered running the code on Google Collaboratory to overcome the RAM and Disk limitations of our local computers.
e. Due to the memory restrictions it was difficult to store the vector conversions in a python dictionary as the large number of unique words

would require a lot of memory to store the conversions even though it would mean ease of access and could have reduced the runtime.

7. Future integration

This project is a small module of the actual framework imagined by Sunlight foundation. The proposed system has a framework wherein a structured view of all comments is available and one can find the related comments to that in the same section. It proposes other features like highlighting words which actually cause the similarity, find which words are frequently occurring and so on. We can do this using integration with Elastic search to ensure faster retrieval of information.

8. Future Scope

a. We can further try to incorporate context specific word embeddings that can further help with clustering similar comments based on the context in which different words are used
b. Since the large volume of comments is a bottleneck to the performance of the algorithm, we could preprocess the comments to eliminate comments which are evidently possess characteristics of spam.
c. Multi-threading could be implemented to parallelly encode the words to vectors thereby reducing the time required for the conversion.
d. AWS & other parallel processing platforms could be used to overcome the limitations and restrictions of running the algorithm on a local infrastructure.

9. References

- A general overview of the FCC comment-spamming: https://www.pewinternet.org/2017/11/29/public-comments-to-the-federal-communications-commission-about-net-neutrality-contain-many-inaccuracies-and-duplicates/ .
- Article: https://hackernoon.com/more-than-a-million-pro-repeal-net-neutrality-comments-were-likely-faked-e9f0e3ed36a6
- An article with a useful table (“Figure 1”) aligning the elements of each quasi-duplicate comment: https://web.archive.org/web/20171126032711/https://fiscalnote.com/2017/11/13/human-like-bots-infilitrate-u-s-lawmaking-process/