Assignment 3

Yuan Zhou

Abstract

In this research, we applied clustering for both docs and terms. We generated a meaningful topic clustering model and an ontology to describe critical terms in technology domains related to COVID-19 and the relationships between the terms.

1. Introduction

Continuously crawling and analyzing the reports and news covering COVID-19, an unprecedented public health crisis for decades, provides significant value for many purposes. Specifically, our research aims to find trending technologies that help people get through the difficulties. However, given the massive volume of articles reporting COVID-19 every day, it is impossible to review all news and manually summarize critical information. Here Natural Language Processing comes to the rescue. Applying appropriate clustering and topic modeling for documents and terms, we can summarize the corpus efficiently, cluster identifiable groups, and build our ontology to deepen the understanding of this domain.

2. Literature Reviews

Traditionally, many document clustering algorithms rely on the off-line clustering of the entire document collection. But for web document clustering, Oren Zamir el ta. introduced Suffix Tree Clustering (STC) - a novel, incremental, O(n)=1 time algorithm that does not treat a document as a set of words but rather as a string, making use of proximity information between words. [1] For hierarchical clustering, Akira Ushioda provided a way that can easily convert the history of the merging process to a tree-structured representation of the vocabulary. [2] Other ML systems, such as the ASIUM, which learns subcategorization frames of verbs and ontologies from syntactic parsing of technical texts in natural language, also saved us a considerable amount of time compared with tedious human labor. [3]

3. Methodology

3.1 Data preprocessing and visualizations

From the previous research, we concluded the ideal number of features is 500 for the corpus, given its highest F-1 score in the classification task. We saved each document's weights of all 500 features vectorized by three approaches to individual files to generate the matrices we need in this research. Before each clustering task, we used MinMaxScaler to normalize the inputs.

3.2 Document clustering

We used the K-Means clustering method when doing document clustering. We have iterated K from 2 to 10 and use the t-SNE to rescale the documents in a 2D diagram, which helped us make a call for best K for each of the approaches. As for each cluster's summary, we calculated their most important terms and compared them to find each cluster's characteristics.

3.3 Term clustering

Similar to document clustering, as for terms, we used t-SNE for multidimensional scaling to 2D and visualized the terms distribution as a scatter diagram. As to determine the most appropriate number of clusters for terms, we used the hierarchical cluster analysis for Approach 1 and 2, and visualized them as dendrograms individually. Inspired by the clustering results, we created an ontology regarding the tech-related notions in this corpus.

3.3 Topic Modeling and Biclustering

We have applied topic modeling via Latent Dirichlet Allocation and biclustering via Spectral Coclustering. The topics count was derived from the analysis in 3.1; these two models generated summary lists of each topic's most important terms. We could relate to the document summary we created in 3.1 to support clustering results.

4. Results & Evaluation

4.1 Document clustering

In Appx-Fig-1, 2, and 3, we illustrated the distribution of documents in a 2D space after applying t-SNE for multidimensional scaling. It is intuitive to declare that K=5 is the most suitable conclusion driven by the distribution's characteristics of Analyst Judgment and TF-IDF. The K-Means clustering results proved that: there were clear boundaries among 5 clusters in each diagram, and there were few mislabeled docs from the coloring. But the distribution of documents vectorized by Doc2Vec did not show clear borders between clustered docs; thus, we concluded it was not efficient for clustering.

When K=5, we collected the top 10 terms within each cluster's docs identified by Analyst Judgment and TF-IDF and created Appx-Table-1 to compare/contrast them. From the table, we could tell that both approaches identified clusters with very similar top terms, and even the dissimilarity between clusters was almost the same. We colored some keywords that we believe could help define the theme of each cluster.

4.2 Terms clustering and Ontology

For Analyst Judgment and TF-IDF, we used t-SNE to generate a distribution map of terms for each approach before clustering them, as is shown in Appx-Fig-4, and 5. From both diagrams, it was not obvious to separate any cluster from others as there are no clear boundaries within the distribution. Our theory is that the terms are not special enough to become domain-specific because the corpus is scrapped mostly from tech news articles facing mass readers.

To get the best clusters to count, we have illustrated the result of hierarchical cluster analysis as two dendrograms in Appx-Fig-6 and Appx-Fig-7. From both dendrograms, we could derive the ideal cluster count is 6 for each approach. After applying K-Means onto t-SNE distribution, we got Appx-Fig-8 and Appx-Fig-9. We can tell that both Analyst Judgment and TF-IDF behaved badly in the K-Means clustering, as clusters overlapped heavily. Such results concur with our theory that most of the terms extracted from vectorization are common words instead of domain-specific jargon. But that doesn't mean we could not generate useful ontology. We could use the terms shown in Appx-Table-1 and Appx-Table-2, plus our understandings of the fact to inspire us the ontology structure. Backed by the reasonings above, we created an ontology illustrated in Appx-Fig-10.

4.3 Topic Modeling and Biclustering

As shown in 4.1, the best number of topics is 5 for docs clustering. Using that, we trained LDA and Spectral Co-clustering and collected the top terms of each topic generated by them in Appx-Table-2. By comparing Appx-Table-1 and Appx-Table-2, we found the most critical terms of all approaches shared substantial commonality. The result supported our conclusion on the topic clustering with terms and our ontology.

5. Conclusion

The research illustrated that meaningful clusters exist in the corpus, and we built a useful clustering model. In the future, we could use this model to classify the crawled documents based on the clusters and generate summaries of keywords appearing in each cluster, which could help us find out the technologies getting more prevalent during and post this COVID-19 affected era.

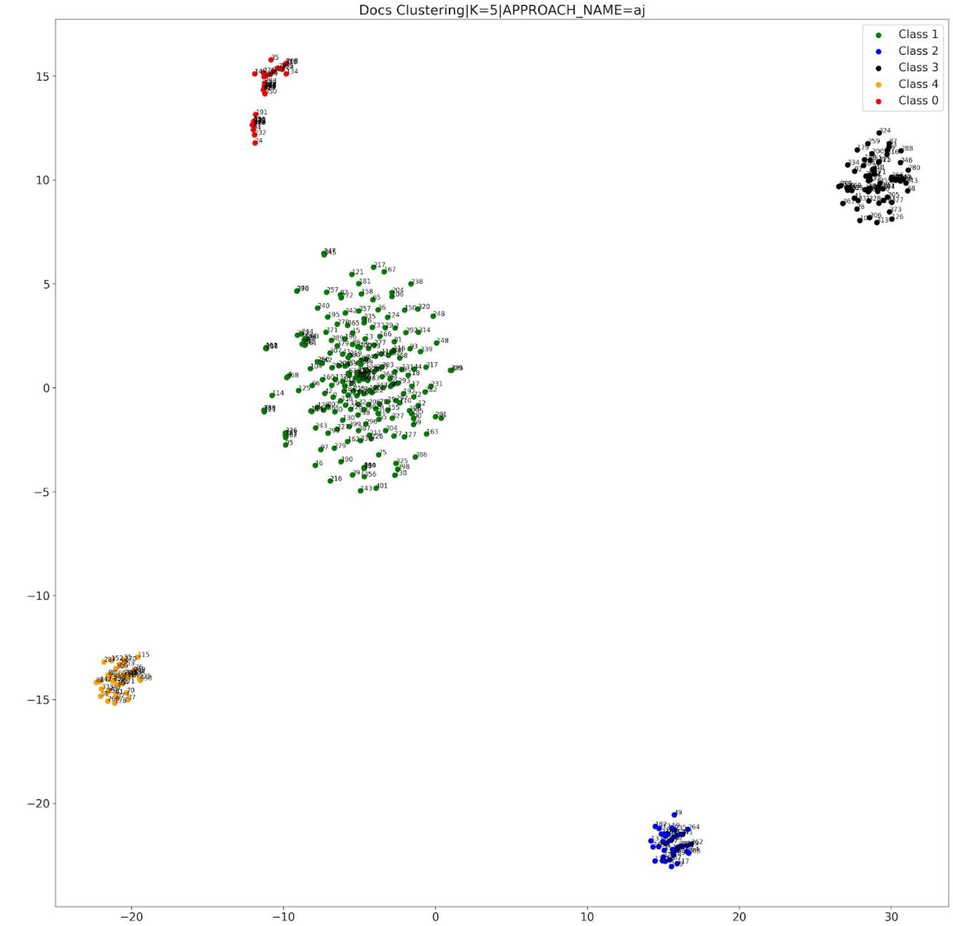
References

[1] Zamir, Oren, and Oren Etzioni. "Web document clustering: A feasibility demonstration." In Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, pp. 46-54. 1998.

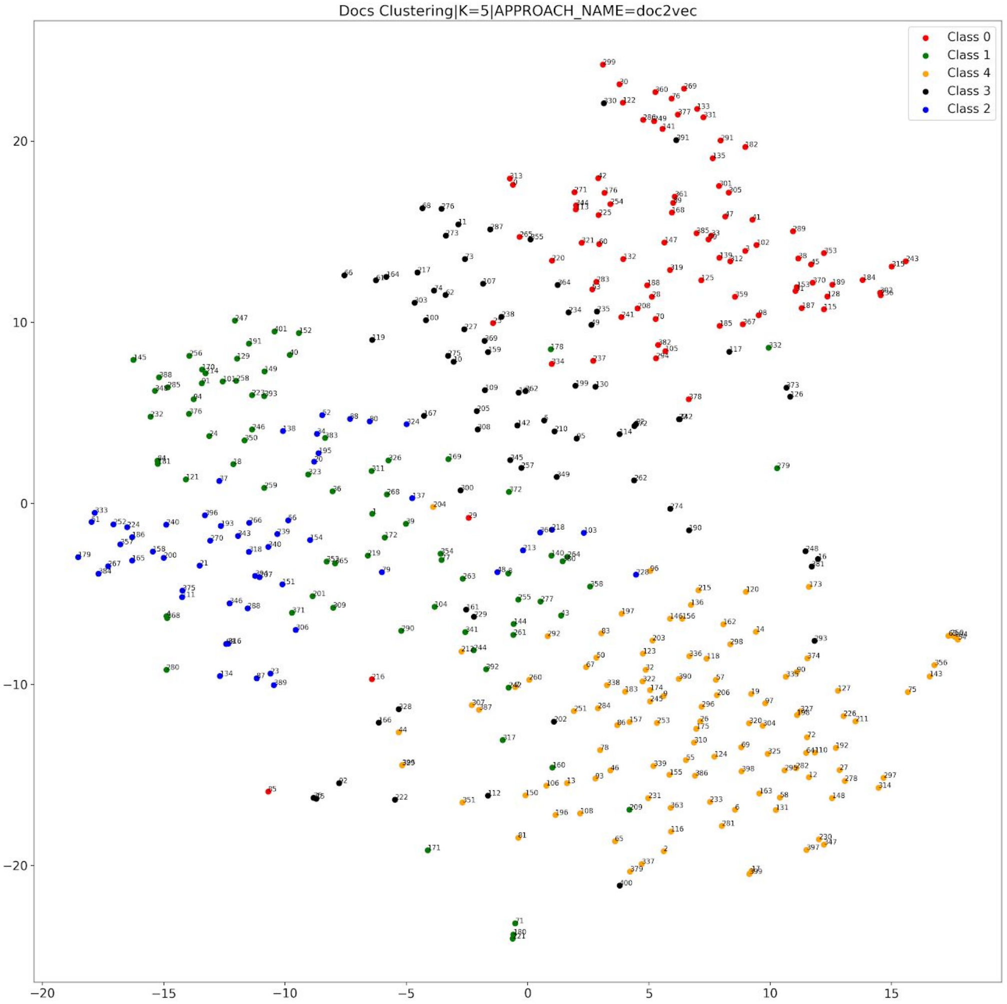
[2] Ushioda, Akira. "Hierarchical clustering of words and application to NLP tasks." In Fourth Workshop on Very Large Corpora. 1996.

[3] Faure, David, and Claire Nédellec. "A corpus-based conceptual clustering method for verb frames and ontology acquisition." In LREC workshop on adapting lexical and corpus resources to sublanguages and applications, vol. 707, no. 728, p. 30. 1998.

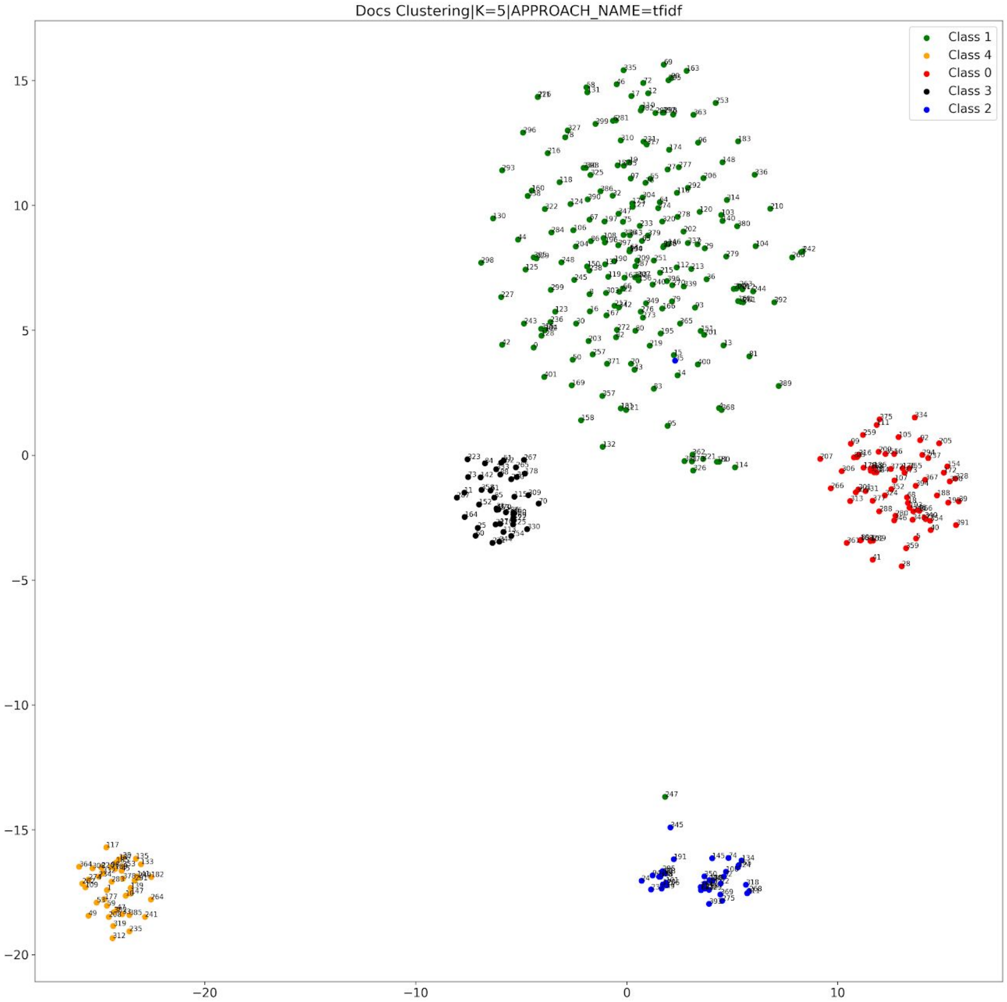
Appendix



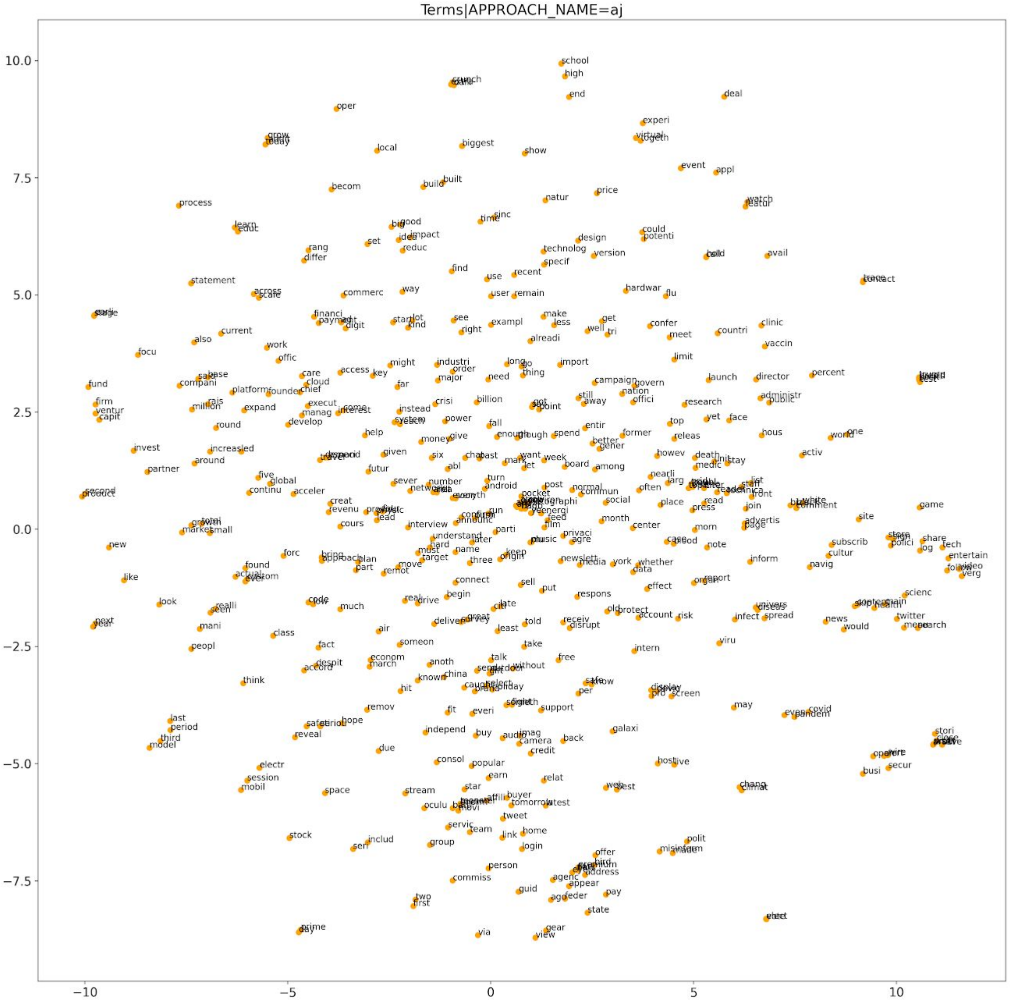
Appx-Fig-1 Analyst Judgment K=5 t-SNE documents distribution



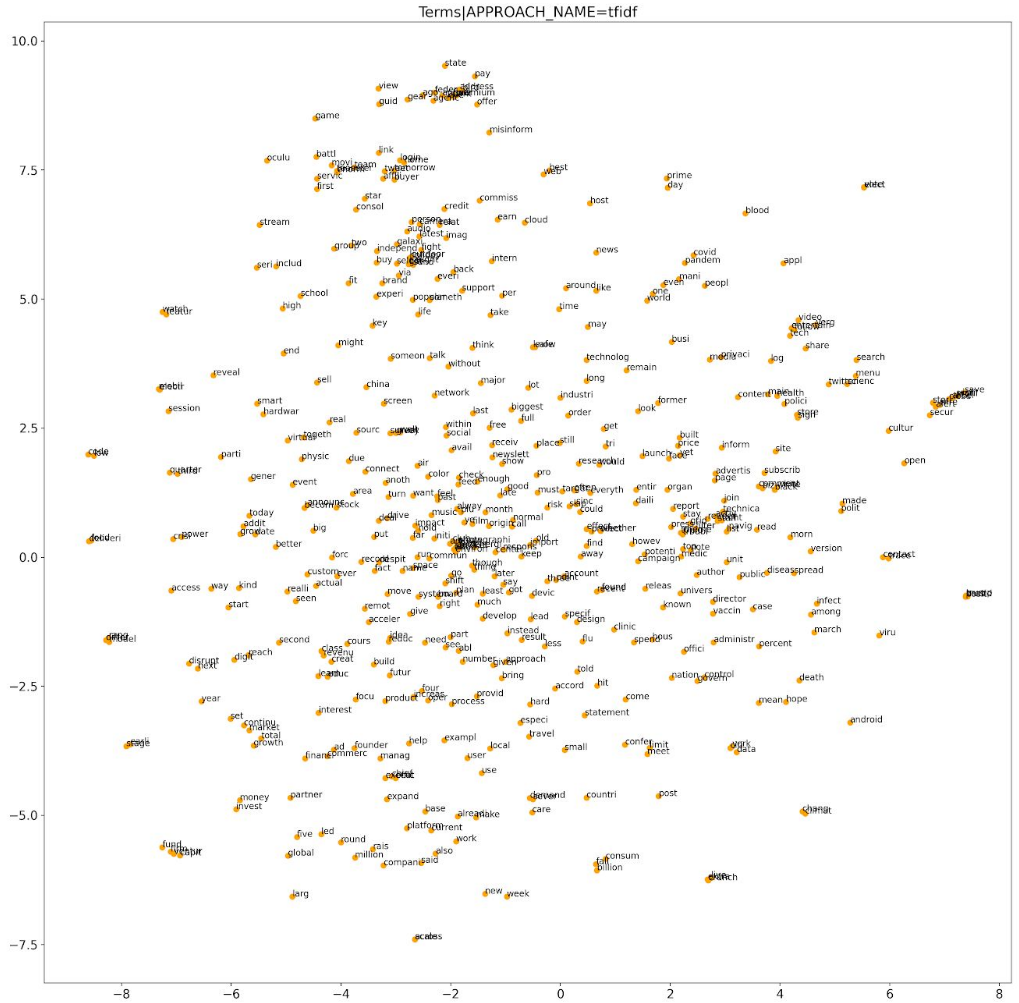
Appx-Fig-2 Doc2Vec K=5 t-SNE documents distribution



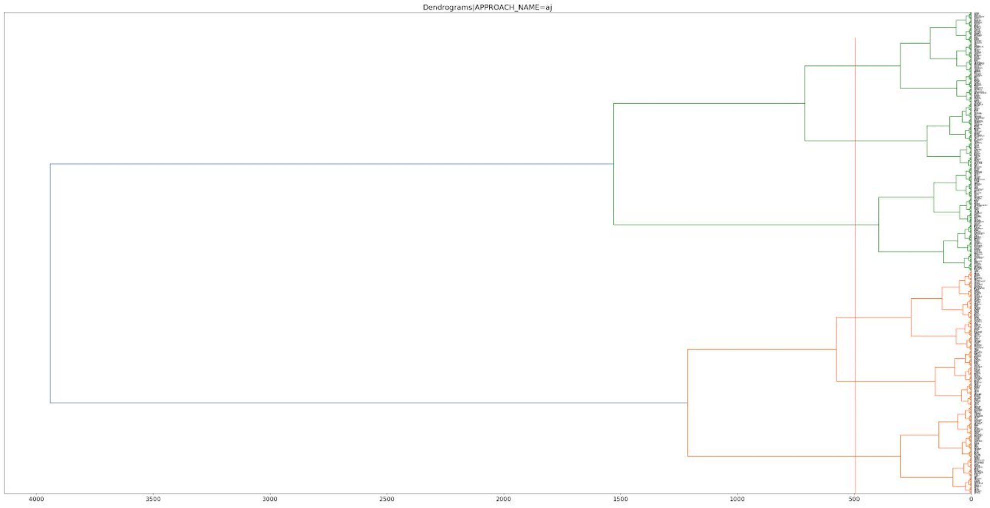
Appx-Fig-3 TF-IDF K=5 t-SNE documents distribution



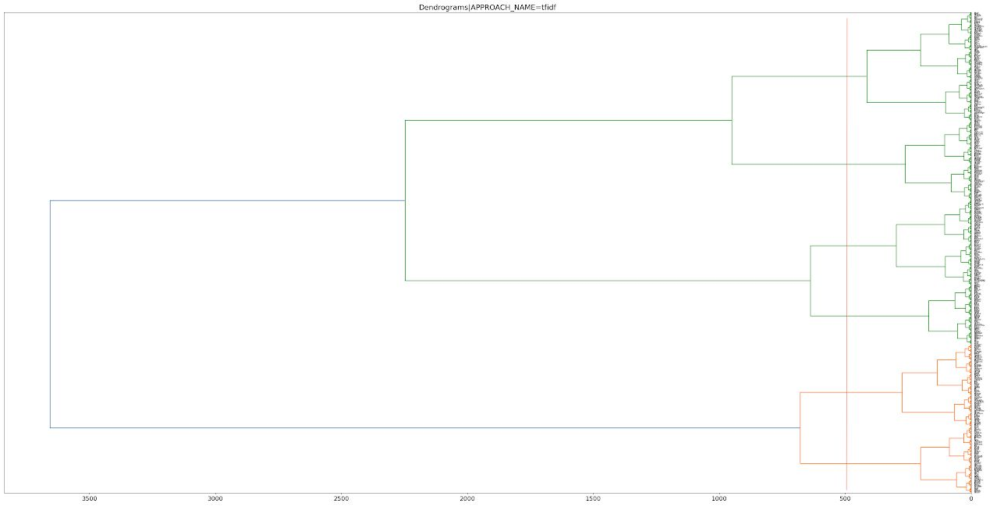
Appx-Fig-4 t-SNE terms distribution for Analyst Judgment



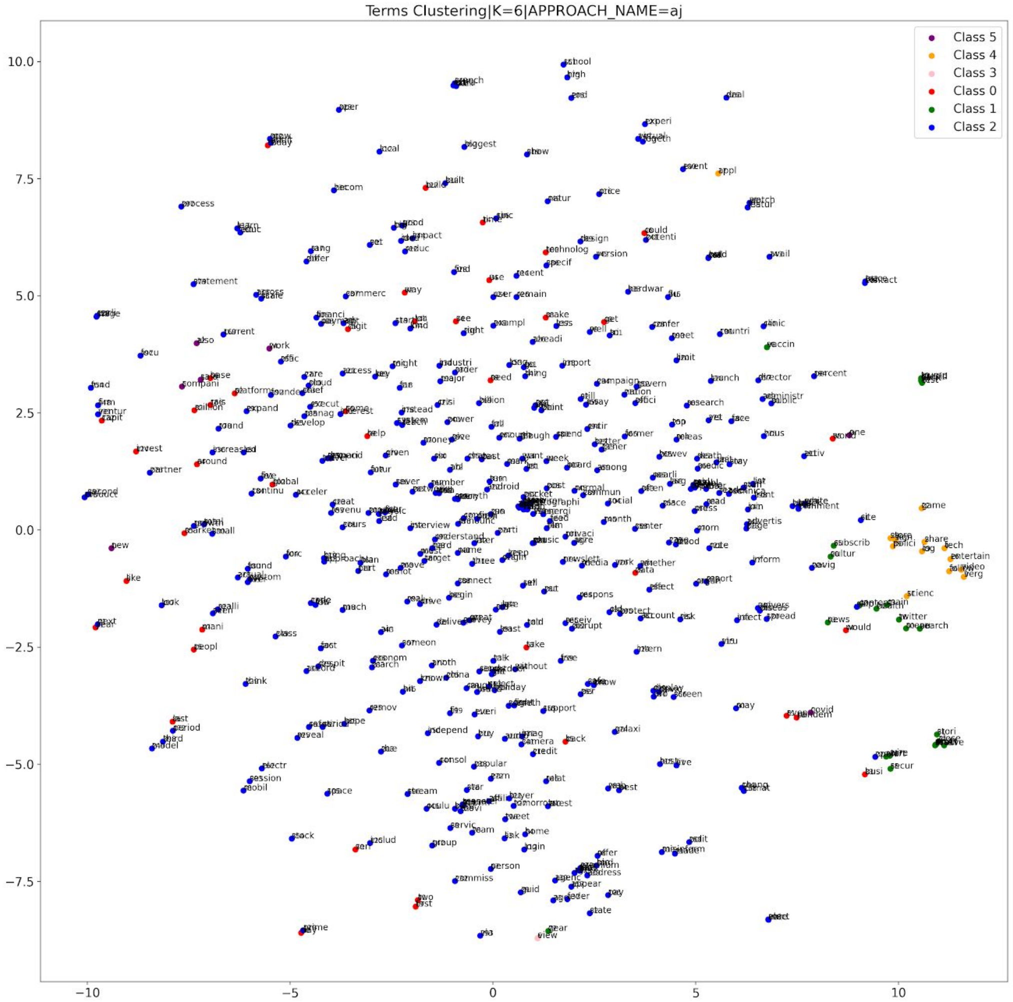
Appx-Fig-5 t-SNE terms distribution for TF-IDF



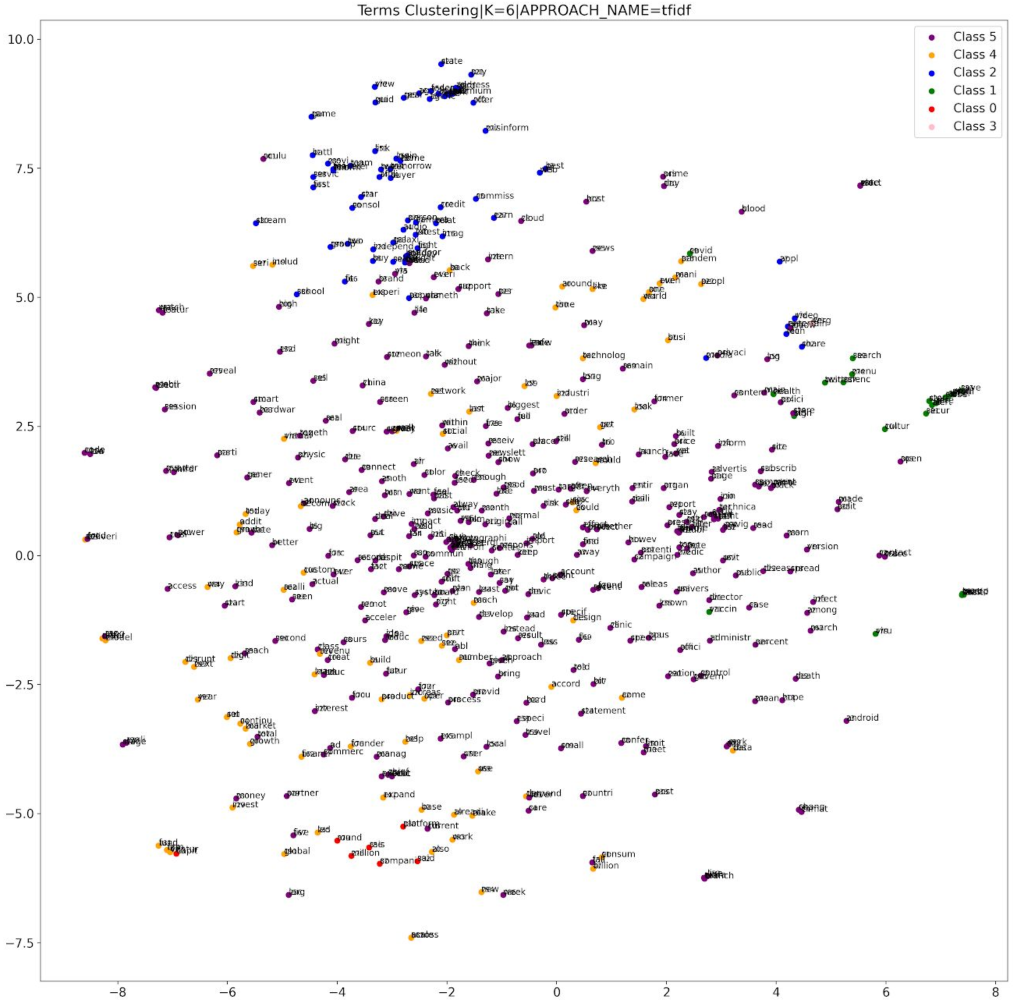
Appx-Fig-6 Dendrogram of terms for Analyst Judgment



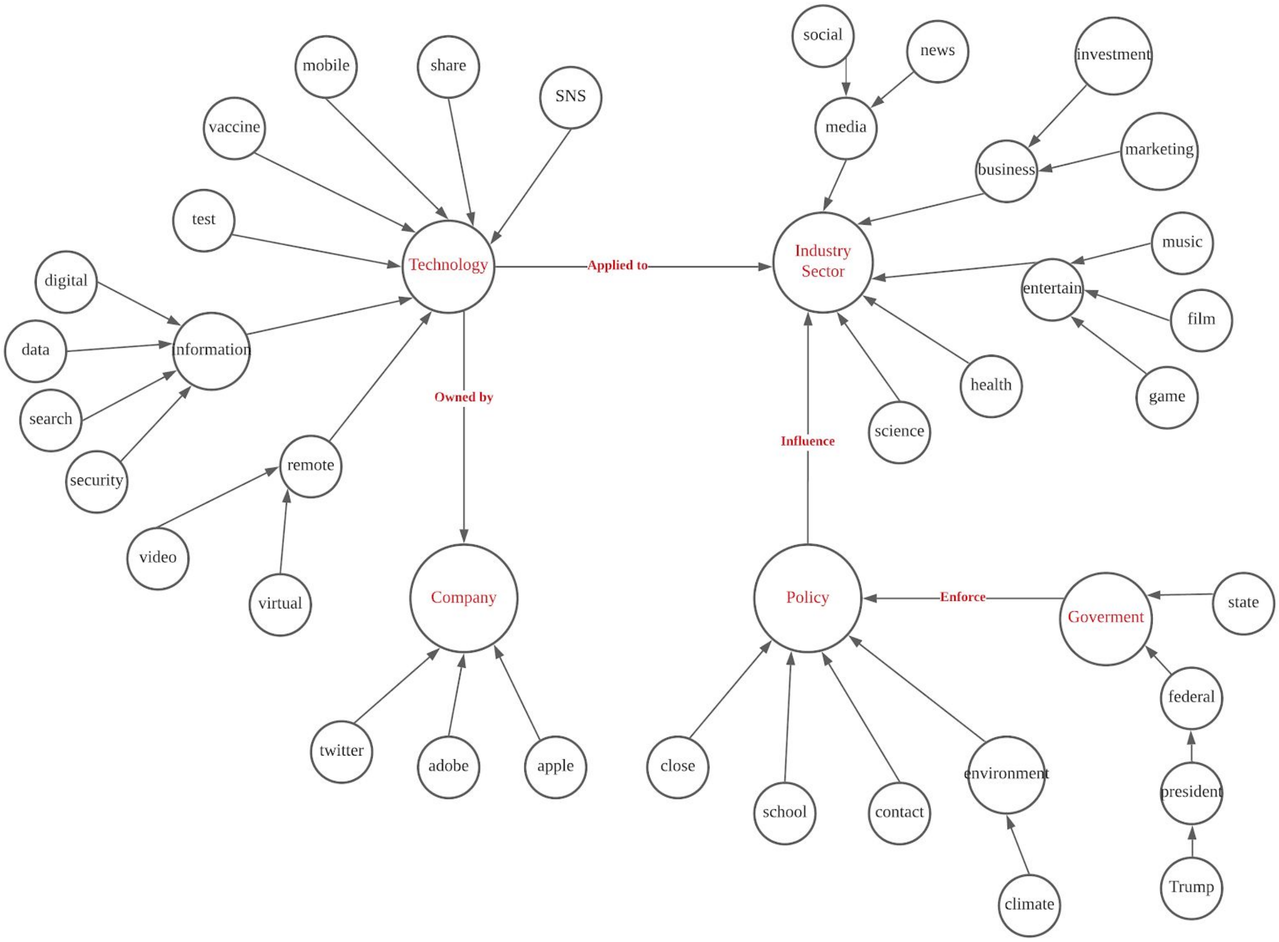
Appx-Fig-7 Dendrogram of terms for TF-IDF



Appx-Fig-8 Analyst Judgment K=6 t-SNE documents distribution



Appx-Fig-9 TF-IDF K=6 t-SNE documents distribution



Appx-Fig-10 Ontology of Technology domain related to COVID-19

clusters	Analyst Judgment	TF-IDF
0	gift, ago, view, buyer, affili, login, park, entertain, eye, gear	view, buyer, ago, affili, gift, gear, login, entertain, park, eye
1	also, like, work, said, see, look, compani, even, covid, million	also, like, said, one, work, compani, covid, million, new, see
2	skip, menu, articl, alert, profil, save, cultur, stori, close, visit	articl, alert, menu, profil, save, stori, skip, cultur, visit, gear
3	skip, pocket, menu, log, art, arrow, adob, follow, verg, main	menu, adob, verg, skip, pocket, arrow, follow, share, entertain, log
4	skip, biz, subscrib, cultur, reader, share, comment, troubl, topic, theme	biz, topic, skip, sign, subscrib, comment, troubl, share, cultur, theme

Appx-Table-1 Top 10 terms in docs of each cluster when K=5

clusters	Topic Modeling	Biclustering
0	skip, gift, login, log, biz, share, subscrib, affili, view, ago	gift, buyer, affili, park, outdoor, eye, ago, teaser, monster, flaw
1	also, like, work, said, look, see, even, covid, one, much	million, capit, rais, invest, round, base, growth, ventur, compani, technolog
2	disrupt, oculu, flu, film, blood, campaign, star, hold, consol, club	stock, safeti, hold, pro, vote, reveal, seriou, display, remov, govern
3	alert, articl, skip, menu, profil, save, cultur, stori, visit, close	cultur, biz, alert, articl, profil, save, reader, subscrib, troubl, theme
4	pocket, arrow, skip, art, menu, log, adob, follow, main, verg	adob, arrow, art, pocket, verg, photographi, follow, bag, ye, environ

Appx-Table-2 Top 10 terms in each topic when topics number is 5