Project Report

Objective

My goal was to apply exploratory text analytics to original scientific research papers and find out whether we can observe the change of contents, topics, or even sentimental values of papers following recent extraordinary pandemic events.

Corpus

For this project, I decided to apply ETA to study recent research papers about COVID-19. I collected my corpus from the website: https://www.ncbi.nlm.nih.gov/pmc/about/covid-19/

On this website, I selected the category as "Coronaviruses broadly (historical and current iterature) (all)" and "Author Manuscripts Only". There were 1471 papers as of July 11, 2020. Since I will download these pdf papers one by one manually, I initially only limited myself to papers published in the last two years. However, limiting my corpus to only two years will lead to the loss of chronological information on this particular topic. I decided to limit my selection to papers in one particular month. I was able to narrow down my corpus to 93 papers published in the month of May from year 2010 to year 2020.

Methods

The corpus, consisting of a total of 93 papers, was downloaded from the aforementioned weblink one by one as PDF documents. This F0 form of text data are under the directory of /data. I have tried different methods to convert the PDF format into plain text, and finally I settled on the package tika.

In the text pre-processing, I removed all titles, authors, abstracts from the start, and removed all supplementary, appendix, references, conclusions from the end of each PDF converted text. In addition, I have to remove all formatting characters and notations in all pages of PDF documents. I quickly realized that I cannot preserve a bag of paragraph, since some papers use single newline character (\n), but other papers use double newline characters (\n\n) to signify the break for paragraphs. Instead, I was able to preserve the page information by search for terms containing "Author manuscript". In the end, I was able to convert PDF documents to a DOC table with three OHCO levels: paper, page, and line. Importantly, I was able to extract "year", i.e., the year when the original research paper first published in scientific journals, in my LIB table. This piece of date turned out to be important for me to extract chronological trend of my data analytics.

Based on DOC and LIB tables, I applied tokenization, annotation, and stemming with the NLTK library to form my VOCAB and TOKEN tables. The next step in my ETA is to generate TFIDF table using two different bags, either paper or page. In order to control the number of features, I also created a SIGS table, which only keeps top 4000 words by their ranking in tfidf_page_sum values. Normally we exclude proper nouns in this SIGS table when dealing with corpus from

novels. However, names of people are not common in scientific papers, they are unlikely to dominate the tfidf table. In addition, in the hope of keeping some important biological terms, such as COVID, I decided to not drop proper nouns.

TFIDF tables presented me with vector space representations of my papers. I applied various distance measurement methods to generate pair distances for all papers, and created hierarchical plot of all these novels accordingly. In order to further reduce the feature space into its most informative dimensions, principle component analysis was employed to reveal axes of maximum variances. Loadings of words as well as distributions of first few PCAs were plotted for visualization.

Topic modeling was performed with Latent Dirichlet Allocation procedure, using page as my bag. I experimented to also include NNP in my TOKENS table for analysis since many disease names are proper nouns, however the result is not satisfactory because many useless terms were also remained. For my final result, I only kept NN, or NNS words in my analysis. Other notable choices of parameters are number of topics at 10, max_features of 4000, and stop_words as "English".

Word Embedding was generated with Gensim's library, in particular the word2vec function. Similarly, I used page as my bag, and excluded proper nouns in my corpus. To generate word vectors, I used the length of window of continuous bag-of-words as 5, the length of word vector at 200, and limited to only include words with a minimal count of 50. In the end, I obtained word vectors for a total of 1027 words, and subsequently I applied t-SNE to convert these high dimension vectors into 2-dimensions for visualization. In addition, I used semantic algebra to obverse analogies among words.

Sentiment Analysis was performed with the lexicon file salex_nrc.csv. With this lexicon, I labeled my TOKEN table with sentimental values, and observed the change of sentiment values in my corpus at both the line and page levels. I also preserved the information of years when these papers were published in my TOKENS table, and subsequently applied VADER analysis to all lines since this is a bag that is closely matched with sentences.

Results and Discussions

All results and figures are in the jupyter notebook. Most of results are self-explanatory, however I would like to point out a number of important and interesting results here.

1. COVID is the top vocabulary in my tfidf table when I use paper as bag (Figure 1). COVID is new word, coined by United Nation World Health Organization in Februrary 2020 following the latest outbreak of coronavirus pandemic. This outbreak was first identified in December 2019 in Wuhan, China, and the actual coined word is COVID-19. In my text processing, I removed all trailing numbers since trailing numbers are a common practice in scientific papers when citing references. In my corpus, there are actually only 4 papers published after this outbreak, i.e., four papers published in May, 2020. This result indicated that the

word COVID was heavily used in these four papers, and this high specificity lead to a very high value of tfidf. We can inspect the output VOCAB table ranked by tfidf_paper_sum, COVID only showed up a total of 86 times, the 2nd smallest number in these top 20 terms. Other terms with high tfidf include hbov and osteonecrosis, names of two diseases which are not prevalent in our corpus. It is also interesting to note that other diseases, such as influenza, SARS, MERS, which are common in our corpus of coronaviruses, are not among the highest tfidf values since they are present in many papers.

2. PCA analysis and Loading. For PCA, I only focused on significant 4000 terms with highest tfidf_page_sum values. In particular, I included NNP terms in order to keep important disease terms such as COVID. Indeed, in the PCA loadings, the positive loading of PC0 included three important diseases: covid, ebola, and influenza (Figure 2). This term also included words such as health, global, china, patients, indicating these words are highly important in explaining the variance of our data.
3. Topic modeling is an unsupervised algorithm to assign labels to the words in documents. I limited the number of topics to 10 since my corpus only had 93 papers. As I had already mentioned in the method section, for this analysis I stuck to customary approach of keeping only NN or NNS terms. LDA method was employed to create THETA and PHI tables. To inspect these topics, top 10 terms of each topic were displayed and their contributions to documents were plotted. In aggregation, the topic 3 carried the highest weight, and it consisted of terms such as activity, protein, binding, cell, receptor, domain, etc. In order to observe whether there was a chronological correlation among topics, I embedded "year" information into the DOCTOPIC table, and plotted the ranking of topics in relation to year. There was an anti-correlation from year 2011 to 2020 (Figure 3). The most important topic in year 2020 was "studies study effect patients virus children data infection size influenza", however this exact topic ranked as the least important in year 2011. On the other hand, the least important topic in year 2020, "detection antibody dna aptamers nan surface analysis protein figure pathogen", ranked as the most important in year 2011. This change of importance in topics also appeared to be gradual, as the importance of one topic gradually shifted upwards, whereas the other one gradually shifted downwards. The most important topic in year 2020 "studies study effect patients virus children data infection size influenza" appeared to be more patient-centric, whereas the most important topic in year 2011 "detection antibody dna aptamers nan surface analysis protein figure pathogen" was more basic science-centric. This shift of topics in my corpus probably signified the recent shift in coronavirus studies, i.e., the urgency of treating patients was taking precedence over understanding the molecular interactions between viruses and other proteins or antibodies.
4. Word embedding as visualized by tSNE is quite interesting. I identified a number of interesting features in this plot. For example, around coordinate (50,2) (Figure 4, top), there is a cluster of words, min, nm, ul, mm, h, buffer, concentration, temperature, room, ph, samples, all related to experimental conditions. This is probably a unique feature for my corpus of research papers. At coordinate around (15,10) (Figure 4, bottom), a cluster of words: structures, crystal, backbone, xray, cysteine, hydrophobic, bond. These words are

very common words in structural biology, a scientific discipline studies protein structures and interactions. At coordinate (18,-8.5), two words dementia and diabetes. These two diseases are pathologically very different, one is a neurodegenerative disorder, and one is a metabolic disorder. However, both diseases are usually manifest at an old age. It is likely that they both were mentioned in the context of old age in my corpus. I have also played a number of scenarios in semantic algebra. One interesting example is that protein to enzyme is analogous to viruses to influenza, i.e., the latter is a type of the former. I listed a number of examples in my jupyter notebook.

5. Sentimental analysis. It is a common practice, as well as an unwritten rule, that scientific paper should be written in a neutral tone. However scientists do not live in vacuum, either consciously or subconsciously, scientists usually end up using words with sentimental values in their papers. When I started this project, I was worried that there would not be enough words with sentimental values for me to perform meaningful analysis. In the end, I was pleasantly surprised by my findings. Firstly, my result showed that "trust" has the highest weight in this corpus, followed by fear and sadness (Figure 5). This result is understandable since scientific research is usually build on top of other people's research, a "trust" between scientists is an important feature to augment, repeat, or improve other people's work. "Fear" and "sadness" are also understandable since many of these papers are dealing with diseases and pandemic events. The most interesting, perhaps also the most important finding of my ETA project is the chronological trend of sentiment values when I use VADER approach to analyze sentiment by lines. In this VADER plot of "compound" value (Figure 6), I observed a dip centered on year 2013, and then recovered by year 2016, and subsequently had some temporary ups-and-downs, and finally ended with a likely starting downturn at 2020. My corpus is "Coronaviruses broadly (historical and current iterature) (all)", and this dip in VADER plot is in concert with another global outbreak of coronavirus, MERS, Middle East respiratory syndrome-related coronavirus. MERS is first identified in September 2012, and by year 2015, MERS-CoV cases had been reported in over 21 countries. It will be an interesting extension of my project to examine the data all the way back to year 2000, in order to include the event of another coronavirus outbreak, SARS, which was first identified in November 2002. By the same logic, it will also be interesting to revisit the same analysis a few years down the road in order to map out a similar change in sentiment for our current COVID-19 pandemic.

Summary and outlook

In my project, I used various ETA tools to analyze a collection of research papers of coronavirus from year 2010 to year 2020. COVID has the highest tfidf values when "paper" is used as bag, siginifying its high specificity. In topic modeling, I observed a chronological reversal of topic importance. In earlier years, topics are more basic science-centric, whereas in year 2020, the most important topic is more patient-centric. In sentiment analysis, the change in compound sentiment value is likely in concert with the outbreak of coronavirus pandemic events. It will be

interesting to extend this project to include additional coronavirus pandemic events, or apply this analysis to corpus of other infectious diseases.

References

1. https://www.ncbi.nlm.nih.gov/pmc/about/covid-19/
2. https://en.wikipedia.org/wiki/2002%E2%80%932004_SARS_outbreak
3. https://en.wikipedia.org/wiki/Middle_East_respiratory_syndrome
4. https://en.wikipedia.org/wiki/Coronavirus_disease_2019

term_id	term_str	n	pos_max	num	stop	p_stem	tfidf_paper_sum
5598	covid	86	NNP	0	0	covid	0.089514
9928	hbov	222	NNP	0	0	hbov	0.086890
7966	et	1840	CC	0	0	et	0.078284
4166	cells	2248	NNS	0	0	cell	0.073709
1485	al	1836	RB	0	0	al	0.072979
1054	ace	222	NNP	0	0	ace	0.068110
2352	ask	115	NNP	0	0	ask	0.066521
4049	cd	689	NNP	0	0	cd	0.065696
15850	osteonecrosis	82	NN	0	0	osteonecrosi	0.059695
11771	irf	254	NNP	0	0	irf	0.059353
17355	pparγ	180	NNP	0	0	pparγ	0.057878
10781	igf	111	NNP	0	0	igf	0.055303
15346	ns	354	NNP	0	0	ns	0.055263
14377	ms	369	NNP	0	0	ms	0.054986
23731	wnv	174	NNP	0	0	wnv	0.054535
2697	axl	202	NNP	0	0	axl	0.054418
1662	am	151	NNP	0	1	am	0.053392
8577	fig	478	NNP	0	0	fig	0.052099
9977	health	622	NN	0	0	health	0.049710
1826	ang	118	NNP	0	0	ang	0.049438

Figure 1. VOCAB table ranked by tfidf_paper_sum

```
Books PC0+ health covid global china public ebola patients cases risk influenza
Books PC0- al et cells t cd ms ns irf expression ace
Books PC1+ ace adam et al ang covid glutamate patients at1r chf
Books PC1- antibody antibodies hiv gp env binding dna s bnabs protease
```

Figure 2. Postive and negative PCA loading for 1st and 2nd PCAs.

year	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	topterms
topic_id												
2	0.170457	0.050480	0.009776	0.120254	0.014049	0.019999	0.197195	0.017001	0.139421	0.105904	0.231455	studies study effect patients virus children data infection size influenza
6	0.006112	0.033083	0.018281	0.013386	0.058887	0.009930	0.018488	0.104714	0.077771	0.001702	0.218729	patients chalcones mabs nan chalcone reaction lesions injury liver catalyst
5	0.074233	0.035563	0.025100	0.042518	0.001907	0.280217	0.141476	0.084928	0.079092	0.296838	0.155144	nan cases antibodies disease antibody studies strains number virus gp
8	0.149977	0.097160	0.145674	0.090692	0.066904	0.042341	0.053961	0.076584	0.202637	0.145668	0.110523	virus cells nan infection treatment replication days animals titers dogs
4	0.283363	0.116577	0.046226	0.129702	0.271629	0.019738	0.202962	0.032098	0.086770	0.131241	0.108951	cells cell infection virus response mice responses expression immunity epitope
0	0.038368	0.126979	0.229819	0.062840	0.215281	0.106597	0.070163	0.129308	0.029362	0.004519	0.050252	health blood risk disease information group research symptoms population countries
1	0.025926	0.100740	0.143459	0.148454	0.166402	0.020910	0.045072	0.059076	0.085369	0.116830	0.046461	cells expression cell treatment mm nan mice macrophages differentiation control
7	0.091285	0.140245	0.034231	0.089603	0.016897	0.191115	0.049092	0.118949	0.032527	0.031370	0.032126	model figure time temperature rate data nan school hmtrnap simulation
3	0.152839	0.110494	0.194479	0.118687	0.130527	0.251171	0.201357	0.355686	0.137531	0.138836	0.025451	activity protein binding nan proteins cell site receptor domain receptors
9	0.007440	0.188679	0.152955	0.183864	0.057518	0.057982	0.020235	0.021656	0.129521	0.027091	0.020908	detection antibody dna aptamers nan surface analysis protein figure pathogen

Figure 3. Topic weights of each year as ranked by importance in year 2020

Figure 4. Word embedding as visualized with TSNE projection. Top: around (50, 2), Bottom: around (15,10)

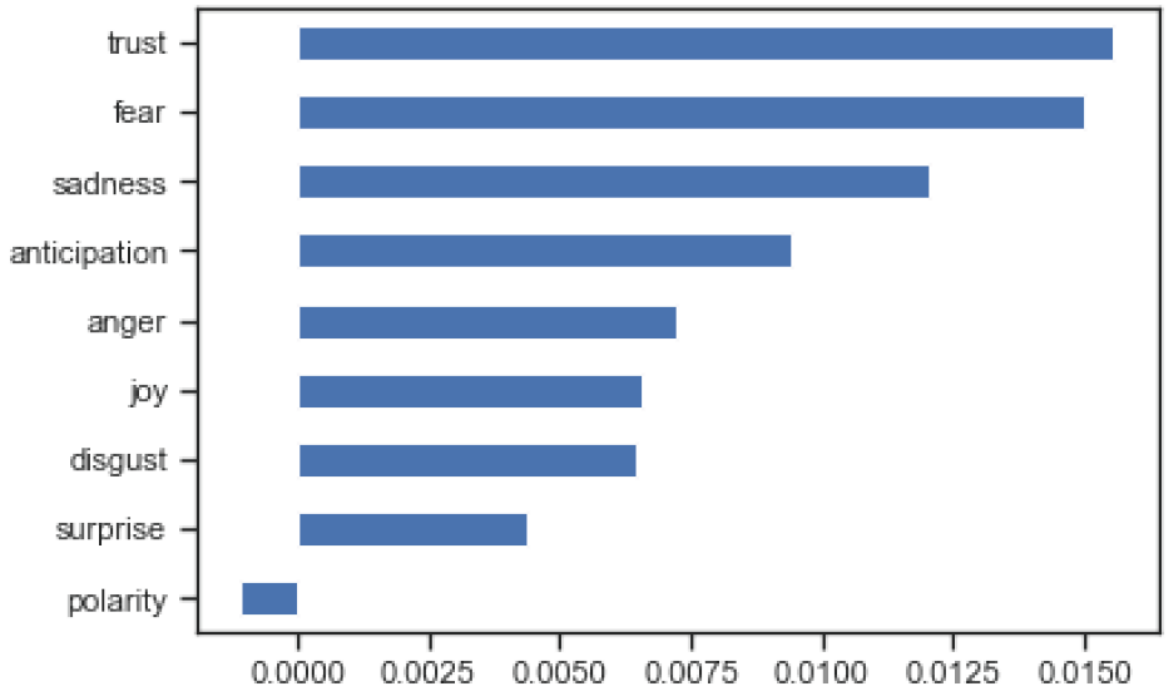

Figure 5. Sentiment Analysis: weights of each emotions in all corpus

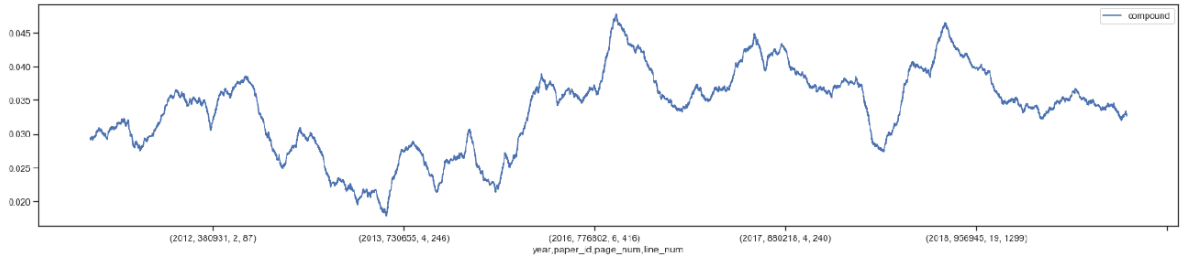

Figure 6. VADER plot of “compound” values in relations to the year of publication.

Appendix: A list of all tables generated in the project

Name	**OHCO / shared index**	**Note**
DOC.csv	['paper_id','page_num','line_num']	
LIB.csv	['paper_id']	
VOCAB.csv	['term_id']	
TOKEN.csv	['paper_id', 'page_num', 'line_num', 'token_num']	
VOCAB_withtfidf.csv	['term_id']	VOCAB table annotated with tfidf
SIGS.csv	['term_id']	VOCAB table with only top 4000 terms
LOADINGS.csv	['term_id']	Loadings of 4000 terms in top 10 PCs
DCM.csv	['paper_id']	Document-component table
THETA.csv	['paper_id','page_num']	
THETA_year.csv	['year', 'paper_id','page_num']	THETA table with additional index ‘year’
PHI.csv	['term_str']	
coords.csv	The column ‘label’ can be matched with ‘term_str’ in SIGS, or VOCAB table	Word2vec coordinate table with TSNE
TOKEN_SA.csv	['year','paper_id', 'page_num', 'line_num']	TOKEN table with SA embedded
VADER.csv	['year','paper_id', 'page_num', 'line_num']	VADER value table, can be matched with TOKEN table at the line_num level