Semester 2:

Week Beginning 14th Mar 22 - Pre-Meeting Status Report

- Most of my time this week has been spent tying up loose ends.
- I have been systematically working through each of the notebooks to ensure that they can be run by anyone and that the datasets required are accessible on Github.
- I have been testing the CoronaCentral search API to ensure that anyone can run it and the accompanying front end application.
- I have updated the repo so that the readme and manual are up-to-date and provide detailed instructions for anybody who has not seen it before.
- Do I write a manual in the appendices?
- I have been working on my presentation - seems to be going ok.
- I have been regularly reading through the dissertation and I am mostly at this point making stylistic and presentation changes as opposed to content. I personally think that I have provided a substantial amount of content - do you agree that there is no more content required and I mostly need to focus on the style?

Week Beginning 28 Feb 22 - Pre-Meeting Status Report

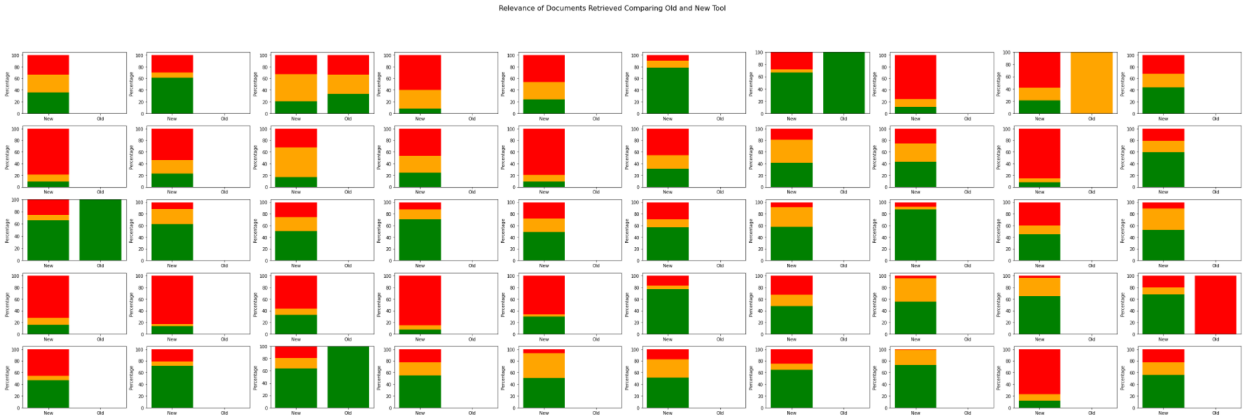
- Most of my time this week has been focussed on finishing up the main chapters of my dissertation. I've spent a bit of time reading it all through to check whether it all flows well. This has raised a few areas of improvement for me:
    - I think some parts of the implementation and evaluation chapters could be more concisely and eloquently expressed so I am going to take some time to work through this.
    - I'm not quite sold on the final section discussing the CoronaCentral search API. I think it is important to put in there, but I'm not quite sure it flows perfectly yet. There is perhaps a bit of waffling which could be diluted down. Perhaps discussing the real-life queries in the CoronaCentral logs is a bit too much? Instead of going into depth with complex figures and descriptions, I could just briefly mention it? Seems like it might be going off on a bit too much of a tangent.
    - Also, a bit of a daft question - do I need to reference tools I have used such as Google Colab, Sklearn etc? Things that don't have papers to summarise them.
- With regards to the presentation - I'd just like to clarify the requirements. My understanding is it should be like the conclusion to a paper - briefly discuss the background and context of the problem, what I aimed to achieve, what I actually did and what lessons have been learned from this.

Week 4 - Pre-Meeting Status Report

- I have spent some time getting the experiments running, placing emphasis on only changing one variable at a time. This has given me a set of results that we can talk through on Wednesday.
- I have also been working on the dissertation, just to keep myself sane really so that I am not writing the entire paper in a few days. I have a first draft of my introduction, literature review and requirements, and I am currently working on the design chapter. To do that, I am trying to make up some diagrams, showing ways of visualising the IR methods I have designed, as opposed to just showing code.
- In the experiments, when I use tools such as CoronaBERT, PubMed Bert, DeepCT etc, where should I reference them? Should this be in the notebooks or in the dissertation itself?

Week 2 - Pre-Meeting Status Report

- I have created the second version of the API - it has some noticeable changes. Firstly, it has a new file for indexing. The .json.gz file simply needs to be added to the directory and then when this file is executed, an index is created. This indexing is a slight change from the original version - most importantly there is no DeepCT. This took a significant amount of time running locally, even on just a subset of the dataset. Afterwards, the API can be restarted very quickly, and there should be no downtime whenever the dataset is updated.
- I have mainly worked on my dissertation this week. My plan is that I am spending a lot of time planning it out and building a narrative, so that when I begin writing, I have all the information I need to hit the ground running. This has identified some gaps in my research, so I have mostly spent time plugging these, running new experiments and visualisations etc.
- One gap I have noticed in my research is a solid way of evaluating the new tool compared to the old one. I have calculated the precision, recall and F1 of the new and old tool and there is a significant improvement. Another way I have evaluated it is by creating visualisations, showing the number of errors they make, shown below. Finally, I'm hoping that we can gain some sort of insight from the CoronaCentral logs to compare them in a real-life environment.


Relevance of Documents Retrieved Comparing Old and New Tool

Things to Discuss:

- Carrying on from last week, just to discuss whether you have had any time to access the logs?
- Is the current API version going to be compatible with CoronaCentral? Is there any other way I need to adapt it?

Week 1 - Pre-Meeting Status Report

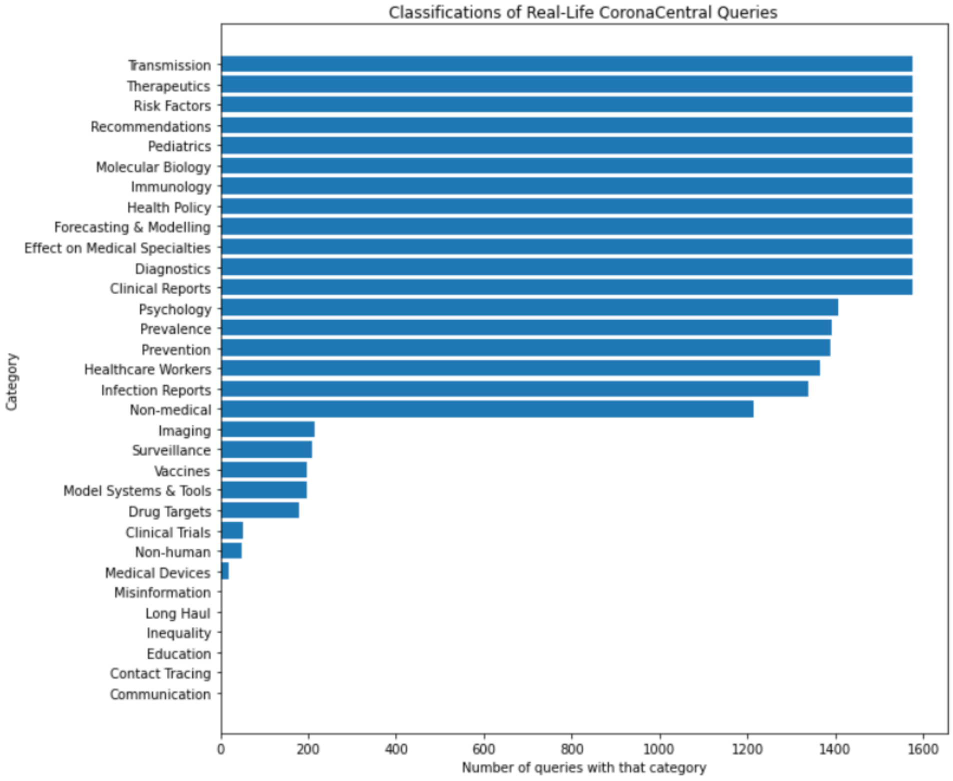
- I have come up with a tentative plan for the dissertation, and I have accompanying notebooks for each section which provide diagrams, results etc.
- I have spent some time collating all the research and experiments that I have carried out into some clear notebooks, giving me a clear train of thought and also diagrams for the dissertation. I think it would be a good next step for me to take all the figures I want to give a concrete idea of the flow of the dissertation, and identify any areas which might need more research.

Things to Discuss:

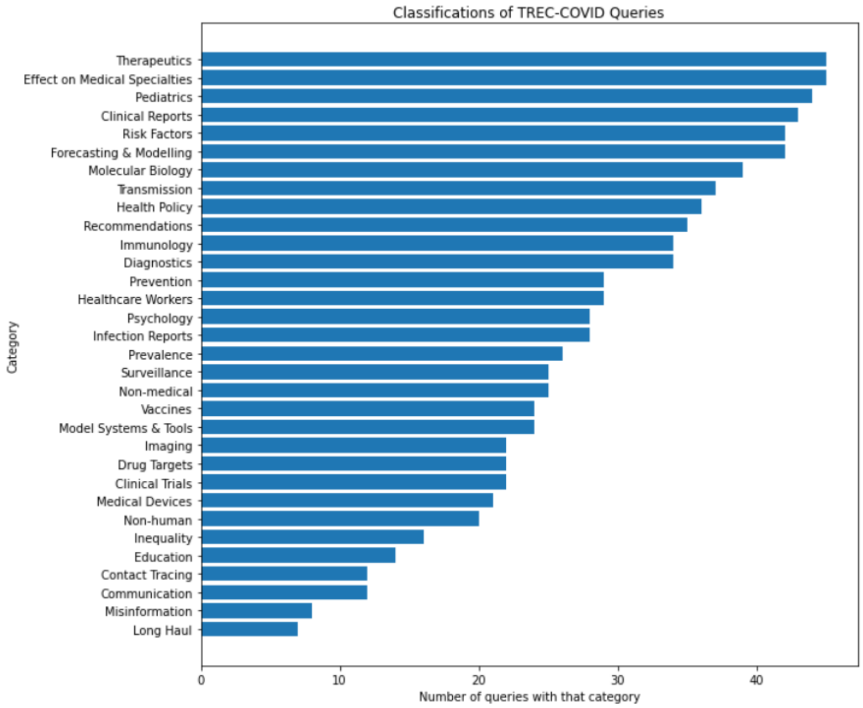
- Deadline for practical work ending and dissertation writing beginning
- As I have already analysed some real-life queries from the CoronaCentral website and we intend to do more of this, will this require ethics approval? Not sure if searches would be classed as personal data?
- Can we have a look at some of the logs in more depth to see how the current search tool is performing for users such as whether they are actually reading the documents retrieved for them? I think it would be good to come up with some sort of metric or strategy for evaluating the tool in a realistic setting.
- What strategy is going to be best to integrate the REST API I have created into the website. I'm guessing I'll need to get it hosted somewhere, like an AWS instance or something similar?

Semester 1:
Week 11 - Pre-Meeting Status Report

- As we discussed at the last meeting, I needed a way of classifying the queries obtained from the CoronaCentral logs to see what kind of topics people are mostly searching for. Obviously, with a lack of training data, this was not very easy but after much experimentation, I think I have come up with a decent strategy. First I took the annotations which were performed for the TREC-COVID competition. For each query in TREC, I picked the relevant documents and got all the topics for those documents. So for each query, I had a list of topics which were my training labels. I created a pipeline which vectorised the documents with TF_IDF and then used logistic regression for multilabel classification. From this, I made a set of predictions for the unseen real-life queries from the CoronaCentral logs. I apologise - I know that I haven't described it very well but hopefully the graph demonstrates what I mean.



I also plotted the training data below, to show the TREC queries and the categories associated with them. While the datasets are vastly different sizes, I thought it would be good to look at the distribution. So it seems actually that the order of importance for these categories is pretty similar. However, we can see in the TREC queries that they have tried to cover an array of topics, while in the real world many of those topics are not actually sought after.

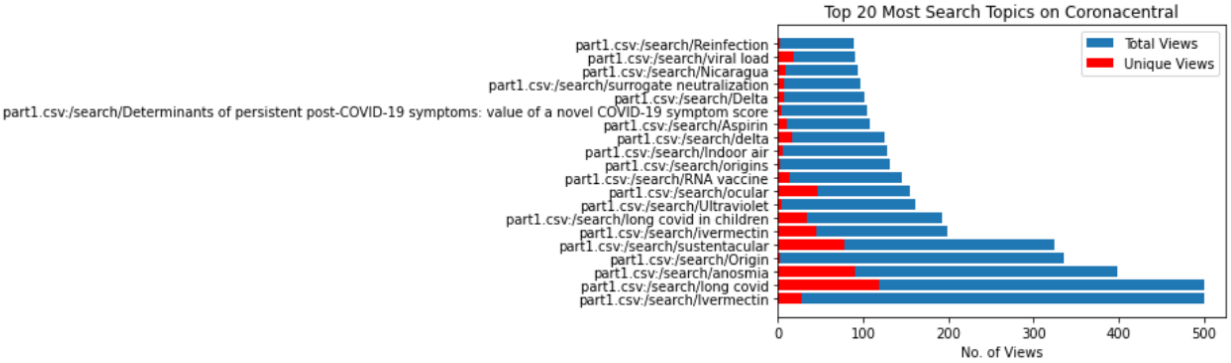


- Thankfully I have solved the problems I was having getting the search API up and running after finding a mistake in the configuration of my virtual environment, so I am now able to run it locally from the command line. For the query, it returns the title, abstract and altmetric scores of the documents although I can obviously change this to whatever information is needed. I'm looking into making a very basic front-end for this as you mentioned so I just need to figure out which framework will be best. I'm going off on a tangent here, but it's actually got me thinking about how it would be cool to have an API gateway providing front-end access to all these kinds of academic research tools - not sure if anything like this exists at the moment.

- I also had a look into the binary searches which were present in coronacentral, like you mentioned. While some did actually cleverly pick up on combined papers, like "mrna vaccine and children", most of the queries were just returning papers about one topic. For example "eye and vaccine" would return papers about the eye and papers about the vaccine, but not papers about both.

Week 10 - Pre-Meeting Status Report

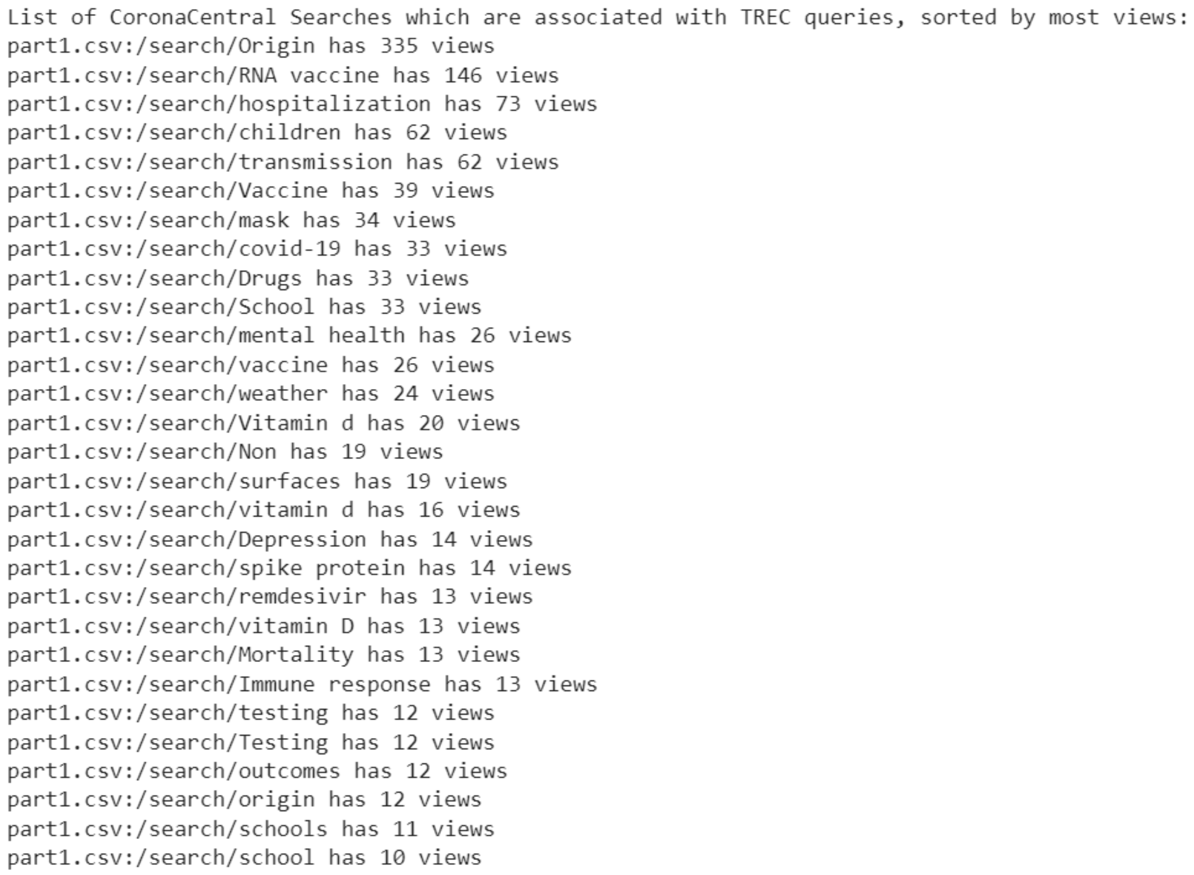
- I have gone through the logs that you sent over. Here is a visualisation of the top 20 most visited topics.



A few observations here.

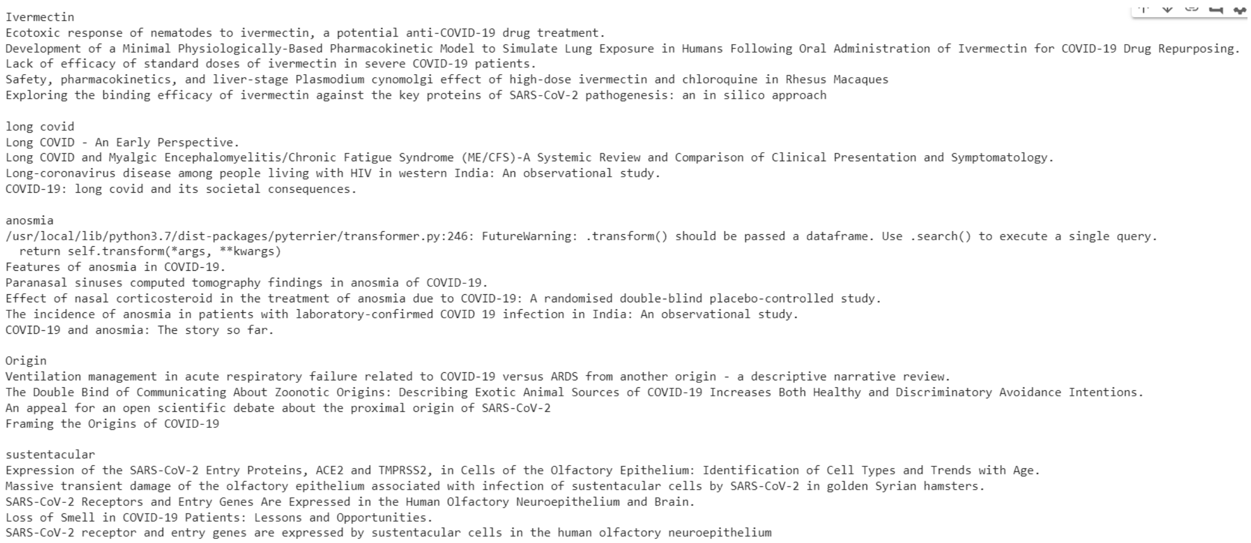
First, I think it is probably best to look at the unique views - for example ivermectin has many total views, but only by a small number of unique visitors.
Looks like long covid is in quite high demand - possibly a bit troubling since there aren't too many papers in CoronaCentral about long covid.
What I wanted to find was essentially what kind of queries were being made. Are they specific, scientific based queries or are they vague like "covid outside the body"? It seems they are specific and scientific which is good news, because my tool works best on these types of queries as discussed last week.

- On the next page I have given a screenshot of the coronacentral pages logged which are associated with TREC queries. I have ordered them by pageviews, to try and get a sense of the TREC queries which are most important, which should help when I am evaluating the tool.



```
List of CoronaCentral Searches which are associated with TREC queries, sorted by most views:
part1.csv:/search/Origin has 335 views
part1.csv:/search/RNA vaccine has 146 views
part1.csv:/search/hospitalization has 73 views
part1.csv:/search/children has 62 views
part1.csv:/search/transmission has 62 views
part1.csv:/search/Vaccine has 39 views
part1.csv:/search/mask has 34 views
part1.csv:/search/covid-19 has 33 views
part1.csv:/search/Drugs has 33 views
part1.csv:/search/School has 33 views
part1.csv:/search/mental health has 26 views
part1.csv:/search/vaccine has 26 views
part1.csv:/search/weather has 24 views
part1.csv:/search/Vitamin d has 20 views
part1.csv:/search/Non has 19 views
part1.csv:/search/surfaces has 19 views
part1.csv:/search/vitamin d has 16 views
part1.csv:/search/Depression has 14 views
part1.csv:/search/spike protein has 14 views
part1.csv:/search/remdesivir has 13 views
part1.csv:/search/vitamin D has 13 views
part1.csv:/search/Mortality has 13 views
part1.csv:/search/Immune response has 13 views
part1.csv:/search/testing has 12 views
part1.csv:/search/Testing has 12 views
part1.csv:/search/outcomes has 12 views
part1.csv:/search/origin has 12 views
part1.csv:/search/schools has 11 views
part1.csv:/search/school has 10 views
```

- I then thought it would be a good idea to look at the top few queries logged in coronacentral and actually test my tool on them, to see how well it performs on real life queries, as opposed to the ones we've been considering from TREC. Below are the results for the top 5 most viewed pages on CoronaCentral. Obviously there is no way of evaluating them with metrics but at a glance, they seem very appropriate. What I really like about the sustentacular searches is that when I look at more results, I can see that it is looking at anosmia as well, so it is learning that there is a connection between the two areas when it comes to coronavirus.



```
Ivermectin
Ecotoxic response of nematodes to ivermectin, a potential anti-COVID-19 drug treatment.
Development of a Minimal Physiologically-Based Pharmacokinetic Model to Simulate Lung Exposure in Humans Following Oral Administration of Ivermectin for COVID-19 Drug Repurposing.
Lack of efficacy of standard doses of ivermectin in severe COVID-19 patients.
Safety, pharmacokinetics, and liver-stage Plasmodium cynomolgi effect of high-dose ivermectin and chloroquine in Rhesus Macaques
Exploring the binding efficacy of ivermectin against the key proteins of SARS-CoV-2 pathogenesis: an in silico approach

long covid
Long COVID - An Early Perspective.
Long COVID and Myalgic Encephalomyelitis/Chronic Fatigue Syndrome (ME/CFS)-A Systemic Review and Comparison of Clinical Presentation and Symptomatology.
Long-coronavirus disease among people living with HIV in western India: An observational study.
COVID-19: long covid and its societal consequences.

anosmia
/usr/local/lib/python3.7/dist-packages/pyterrier/transformer.py:246: FutureWarning: .transform() should be passed a dataframe. Use .search() to execute a single query.
  return self.transform(*args, **kwargs)
Features of anosmia in COVID-19.
Paranasal sinuses computed tomography findings in anosmia of COVID-19.
Effect of nasal corticosteroid in the treatment of anosmia due to COVID-19: A randomised double-blind placebo-controlled study.
The incidence of anosmia in patients with laboratory-confirmed COVID 19 infection in India: An observational study.
COVID-19 and anosmia: The story so far.

Origin
Ventilation management in acute respiratory failure related to COVID-19 versus ARDS from another origin - a descriptive narrative review.
The Double Bind of Communicating About Zoonotic Origins: Describing Exotic Animal Sources of COVID-19 Increases Both Healthy and Discriminatory Avoidance Intentions.
An appeal for an open scientific debate about the proximal origin of SARS-CoV-2
Framing the Origins of COVID-19

sustentacular
Expression of the SARS-CoV-2 Entry Proteins, ACE2 and TMPRSS2, in Cells of the Olfactory Epithelium: Identification of Cell Types and Trends with Age.
Massive transient damage of the olfactory epithelium associated with infection of sustentacular cells by SARS-CoV-2 in golden Syrian hamsters.
SARS-CoV-2 Receptors and Entry Genes Are Expressed in the Human Olfactory Neuroepithelium and Brain.
Loss of Smell in COVID-19 Patients: Lessons and Opportunities.
SARS-CoV-2 receptor and entry genes are expressed by sustentacular cells in the human olfactory neuroepithelium
```

- Here's what TREC says about the judgement of documents:

  Manual judgment of IR results is a time- and resource-intensive process but essential for a gold-standard test collection. It is estimated that it takes approximately 1 minute to judge a single article for a topic, and the goal is to assess several hundred results per topic, requiring hundreds of hours of assessment over the course of the task.

  Due to the pace of the evaluation, the growth of the collection (which doubled within a month), and limited availability of qualified annotators, the depth of the above-described judgment pools is fairly shallow, and some relevant documents will remain unjudged and therefore be considered not relevant. The second limitation is the nature of the collection that combines peer-reviewed and preprint work that is judged solely for topical relevance, which might lead to some less rigorous and potentially erroneous publications judged as relevant.

- Finally, I am working on making the REST API which will act as the search tool. Not quite finished yet, but I might have something to show you at the meeting.

Week 9 - Pre-Meeting Status Report

- I am still on the hunt for the best strategy yet. So far the best tactic is:

  Filter out documents which don’t have Altmetric score
  Index documents by title and abstract
  Create TF_IDF embeddings

  About as simple as it gets, but adding more complexity seems to degrade performance.

- I used this strategy on the queries and analysed some of the mistakes which are made:

  https://colab.research.google.com/drive/1GTNbNtkXHOGZYkHrMOaWGNA-ntomxnIq?authuser=1

- What did I learn? I would say several of the annotations are incorrect. Also, it is very strange to see that a high percentage of the ones which we don’t pick up in the retrieval but were judged as relevant are about molecular biology. I have no idea why this would be the case or how we can resolve it.

- I analysed this strategy applied to each of the 50 TREC queries, and plotted the NDCG score for each one. Notebook linked below:

  https://colab.research.google.com/drive/1ufHD0CmKs6TtOU7bXsEurdjJGq-oy3Jc?authuser=1#scrollTo=CAQ3gOfSmRZh

  Very roughly, the ones which are more specific, which mention lots of technical, medical jargon perform better. More social topics aren’t as easy to search for.

- I think it would be good to have a look at the CoronaCentral logging to see what kind of things people are searching for so that we can tailor the tool to that. For example, are they just searching for keywords? Are they searching in more of a question format? Are they searching for specific scientific issues or less scientific, more social issues?

- I thought it would be beneficial to look into the current performance of the search tool for CoronaCentral and compare it to the results which my tool retrieves.

  https://colab.research.google.com/drive/1bneH3U8k6XZdBtSMkVifqGyfu2CZcYvg?usp=sharing

  Only 12 of the TREC queries actually retrieve any documents. The others simply return no documents. Of the ones that do return documents, it is mostly picking relevant ones, but it might be, for example, only one relevant query. There are hundreds, if not thousands, which are classed as relevant which are not picked up. Obviously, I know I need to be careful about relying on the annotations as they might not be 100% accurate, but still it is clear that the current search tool is missing out a lot of information and my current tool is definitely an improvement.

- Finally, I made some comparison graphs for you to show how the characteristics of documents which are relevant vs irrelevant vs the entire population.

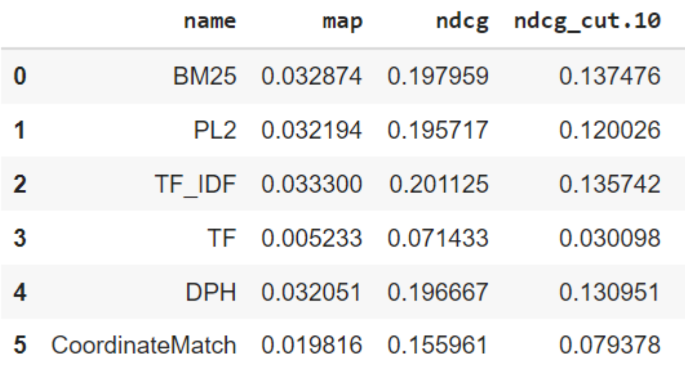
  https://docs.google.com/document/d/1zT3NqJ1mgKEFX3P5vXXQo1qdYOI-g1sI8qdI7xXtxZo/edit?usp=sharing

Week 8 - Pre-Meeting Status Report

- Filtering the dataset by altmetric scores:

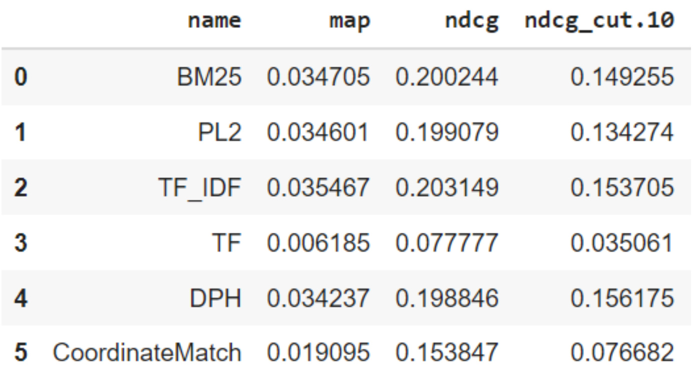
  I filtered the dataset down to remove the documents which don't have a score. This was a pretty massive reduction from around 223000 documents to around 157000. It gave an improvement in results with just a basic search. It seems to make the biggest difference for ndcg cut to the first 10 results.

  Full Dataset:



	name	map	ndcg	ndcg_cut.10
0	BM25	0.032874	0.197959	0.137476
1	PL2	0.032194	0.195717	0.120026
2	TF_IDF	0.033300	0.201125	0.135742
3	TF	0.005233	0.071433	0.030098
4	DPH	0.032051	0.196667	0.130951
5	CoordinateMatch	0.019816	0.155961	0.079378

  Dataset Filtered by Occurence of Altmetric:



	name	map	ndcg	ndcg_cut.10
0	BM25	0.034705	0.200244	0.149255
1	PL2	0.034601	0.199079	0.134274
2	TF_IDF	0.035467	0.203149	0.153705
3	TF	0.006185	0.077777	0.035061
4	DPH	0.034237	0.198846	0.156175
5	CoordinateMatch	0.019095	0.153847	0.076682

I then removed the documents which had scores of less than 10, reducing the size from 157000 to 59000. This made things marginally worse. So I then tried taking all the documents out which are prior to 2019 which also made things marginally worse. So, the only solution I have found which improves so far is taking out documents which don't have an altmetric score at all.

- I tried ignoring abstracts and simply searching through the topics assigned - this degraded performance dramatically.
- I then tried concatenating the topics listed and the abstract when indexing. This had mixed results. It improved performance very slightly for the BM25 representation but not TF_IDF.
- Filtering by DOI makes things very slightly worse.
- Only indexing the titles not the abstracts had no positive effect
- Indexing the titles and the abstracts gives the most positive results yet, giving an NDCG of 0.23. Strangely enough, putting a date filter onto this actually degraded performance.
- I think that perhaps it would be a good idea to have a TF_IDF representation of the documents. We filter out the ones which don't have an altmetric score. We then index them by their title and their abstract concatenated together, perhaps concatenating the topics as well. Then perhaps we then perhaps re rank them according to date or altmetric?

- I have also created a notebook to look at the ways we can filter the dataset - the link is provided below. I think there are lots of useful visualisations in there that show us the trends for the most relevant topics, and hopefully this means we can try and filter the dataset prior to any intense computations. https://colab.research.google.com/drive/1HxQz1Qh8Wakpux0QgNjE5dL8DaM7OZdc?usp=sharing

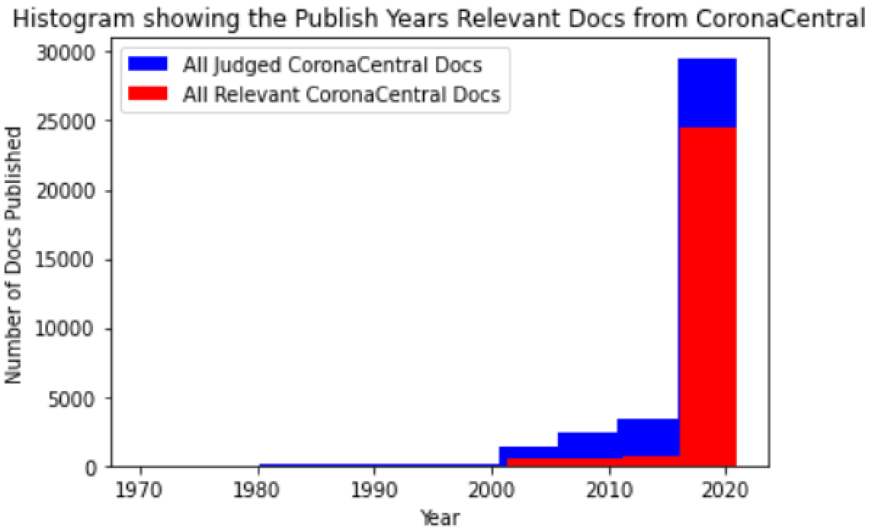
- My main lesson learned this week - altmetrics are important. Removing the documents which don't have an altmetric score can be beneficial for retrieval. However, when we look at the relevant documents, we can see a high percentage of them have a low altmetric score. So, I don't think having some sort of filter on the actual score is better, just a filter on whether the score exists in the first place.
- I also think article types are important - we can see that some article types are rarely classed as relevant, with the majority being research papers.
- DOI is definitely something to consider - the majority of relevant documents have one.
- Unsurprisingly, most of the relevant documents were peer reviewed, meaning they will have more credibility and be more relevant.
- We can see that there are lots of journals which are more relevant.
- Publication year is definitely one to look at - pretty much all of the relevant documents are from 2020.
- Pubmed ID was a bit more ambiguous, more of the documents had one but not necessarily that important.
- Finally it is clear that more topics are more relevant than others, but this could be down to the queries that TREC provides.

- I think it is important to incorporate all these elements into the retrieval, but perhaps in different ways. Some might be used to filter the dataset down, and then others might be trained as features in the pipeline.

- I have also finished writing up a draft for my literature review. I can send you it if you'd like to have a look but it is quite lengthy so maybe not!

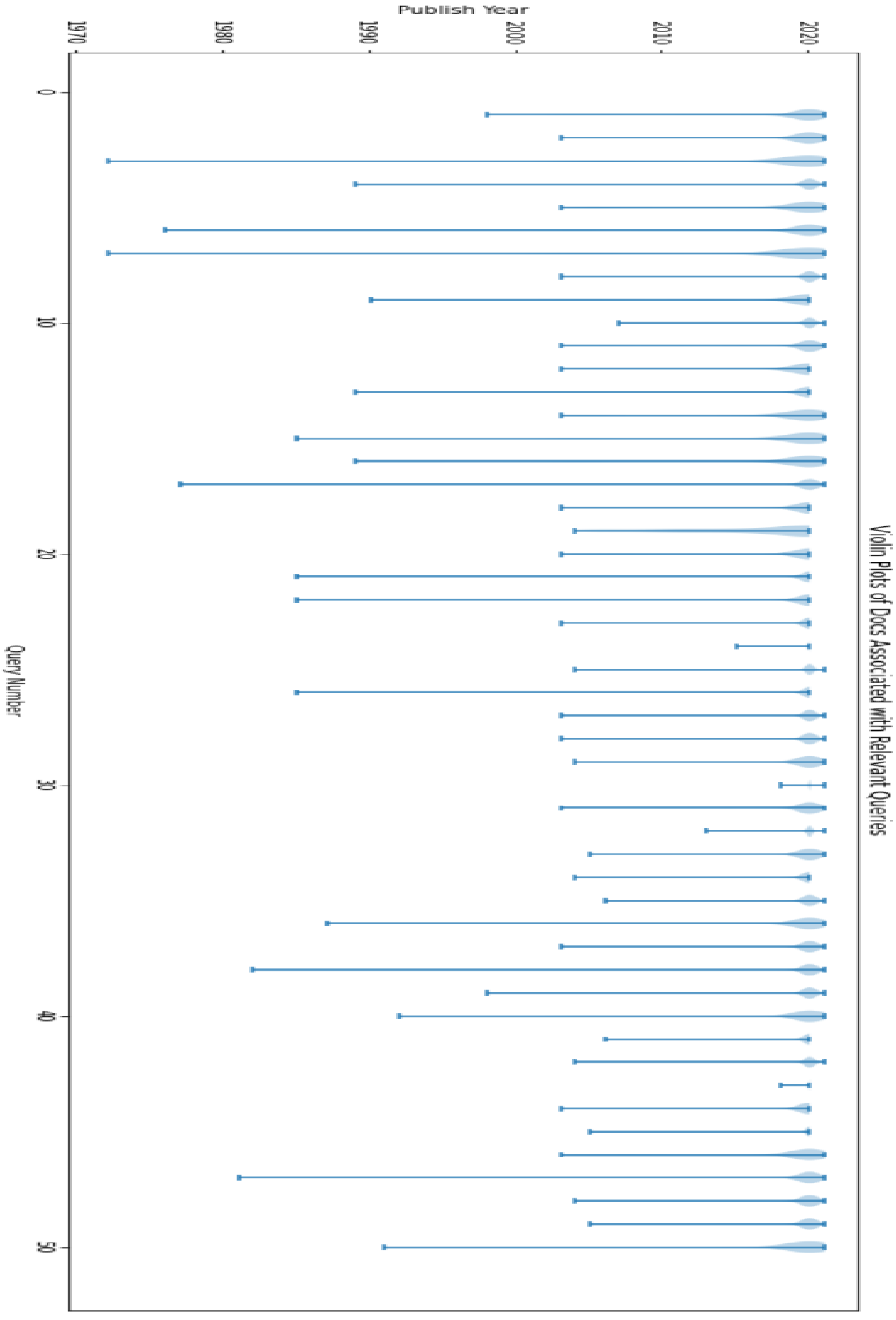
Week 7 - Pre-Meeting Status Report

- I dug further into the publication dates of relevant documents as suggested. I have to admit, there was a line of code missing in my diagram from last week, which makes a rather big difference explaining when the relevant documents were published! Sorry for giving you rubbish last week! Here is the correct diagram, which makes a lot more sense for CoronaCentral. We can see that they generally become relevant after around 2002 when SARS first appeared. The CORD-19 diagram from 2 weeks ago is still correct.



- On the next page is the collection of violin plots, showing what years the relevant documents were published. Doesn't show too much. Seems to be a bigger spread of years for the first round of queries. Maybe the judging changed for rounds 2 to 5? At the first round, the judges would have only had a few new documents to look at, since COVID-19 was quite new then. As a result they might have looked more into the older papers and classed them as relevant. As the rounds went on, however, newer papers became available and they were more relevant to the queries. Just a thought - this could be a rationale behind making sure the test queries are the latest ones. The older queries were judged when there wasn't as much literature available.

- On that note, I split the data up the same way that is suggested in the pyterrier tutorials by splitting up the queries. So, the first 30 queries are for training. The next 5 are for validating and the final 15 are for testing. This works well with my previous thought that we want to test on queries which have been judged more recently.


Violin Plots of Docs Associated with Relevant Queries
Publish Year
1970
1980
1990
2000
2010
2020
Query Number
0
10
20
30
40
50

- As you mentioned, it would be beneficial for me to go through some typical IR techniques, looking at ones from TREC to find the best way of querying CoronaCentral. Here was my process:

  Take the “qrel” dataframe from CORD-19 which stores all the docnos for documents which have been judged in TREC, and gives them the label indicating relevance, partial relevance or no relevance. I filtered this down, to only include documents which are in CoronaCentral.

  I tried some very simple ways of representing the documents as benchmarks such as TF, TF_IDF, BM25, PL2 etc. **The winner was TF_IDF, with BM25 a close second.**

  Next I had a look at training some features. At the moment, I picked 3 features which were whether the document has a DOI, whether the Altmetric is over 100, and whether the publication date was 2020 or beyond. I then trained this pipeline using the data split I mentioned above, using three different techniques. The first was coordinate ascent, the second was a random forest regressor and the third was LambdaMart. These were trained over a TF_IDF representation of the documents. **The clear winner was coordinate ascent.**

  I also had a look at some second stage neural reranking, during which time I encountered some problems. I used TF_IDF as the document representation, in line with my previous findings. I started by trying a vanilla bert, and also a bert which had been trained on SLEDGE (SLEDGE was successful in TREC-COVID). I do not have any results, because the querying took so long and required so many batches, that my Colab session crashed twice. This is obviously not something which we want to be happening, as the results should be provided to the user almost instantaneously. I therefore tried a MonoT5 reranker, which degraded performance and also took 43 minutes to give me the metrics...definitely not appropriate for our purpose.

  I think my main takeaway from all of these experiments is that it is easy to over-complicate the process. Using rerankers did not increase performance and took a ridiculously long time to give me answers. Instead, a simple retrieval where we can leverage some of the features of CoronaCentral such as altmetric and document topic really seems to be the best way forward.

  I looked into some neural indexing as well, which definitely seems promising, especially since the results pyterrier got on CORD-19 work well. It does, however, take a massive amount of time to expand the documents using something like doc2query or DeepCT. I used doc2query on CoronaCentral which ran for 6 hours before crashing. I’m not sure if there is some way we can do this but for a subset of the documents? For example, only the most recent or the highest altmetrics? That way we are only indexing the documents which are most likely to be relevant anyway.

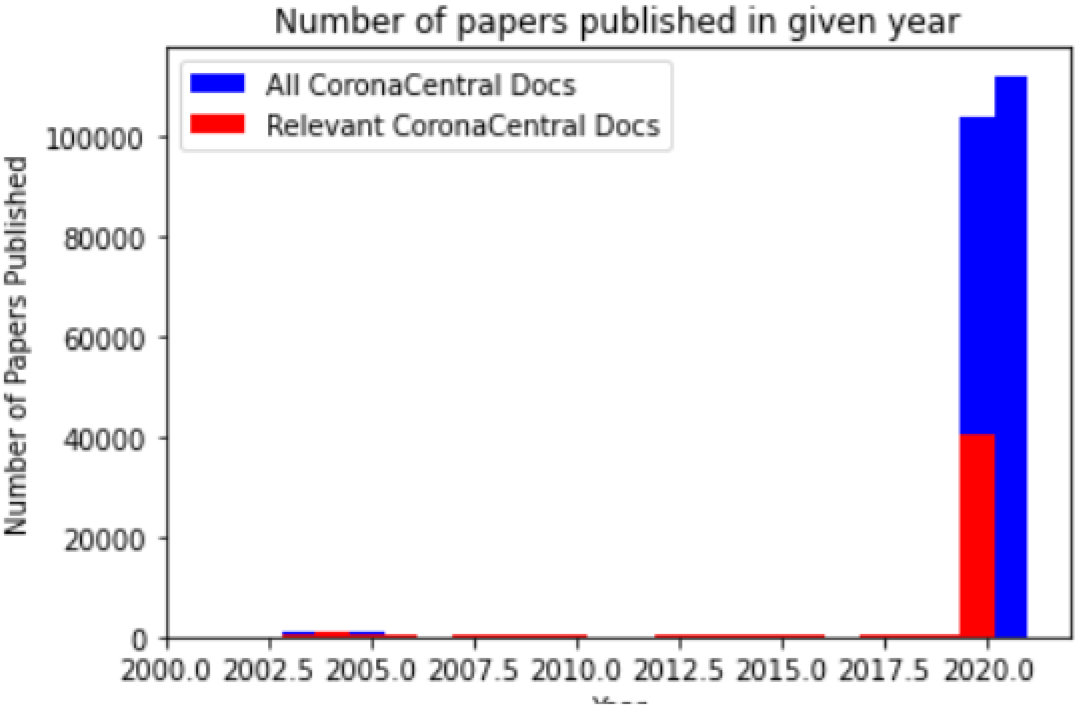
Week 6 - Pre-Meeting Status Report

What was my plan for the week?

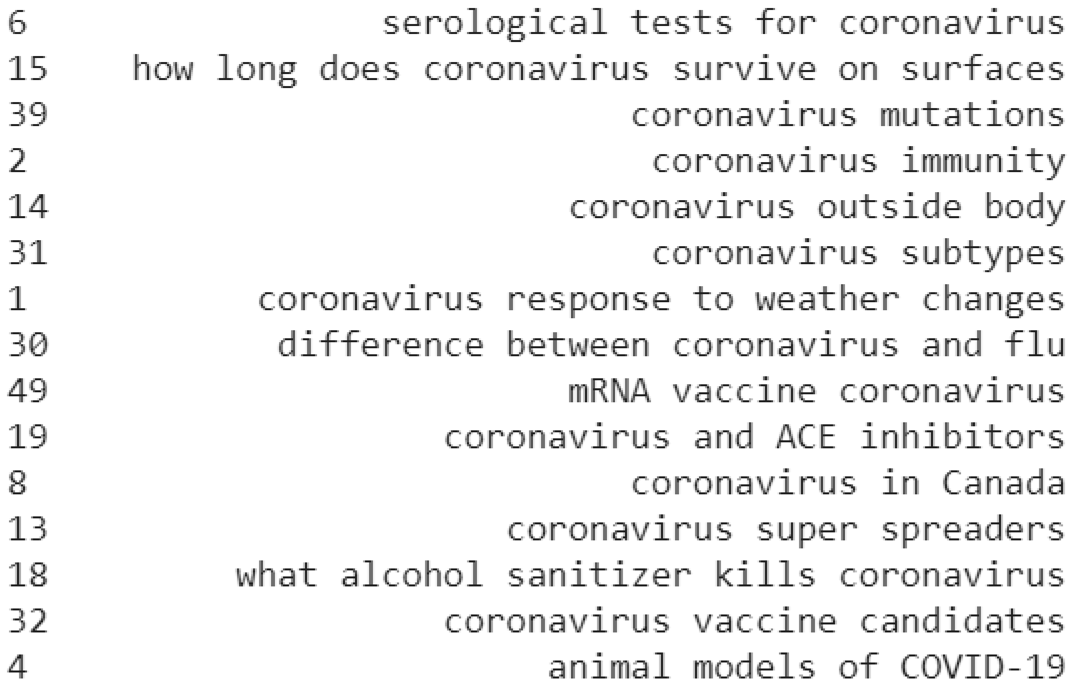
- Find out whether there has been any critique of CORD-19
- Look deeper into date filtering
- Set aside train, validate and test data
- Look into how we can remove irrelevant information from abstracts

What have I done throughout the week?

- Looked at date filtering in CoronaCentral more specifically. Last week I showed you the histogram of relevant documents from cord-19 and when they were released. This week I have done the same for CoronaCentral and the same trend appears:



- I looked at all the queries available, and looked at the documents which are classed as relevant for them, calculating the average year these documents were published. So, here is a list of the queries which are related to the set of documents with the earliest average publication year.

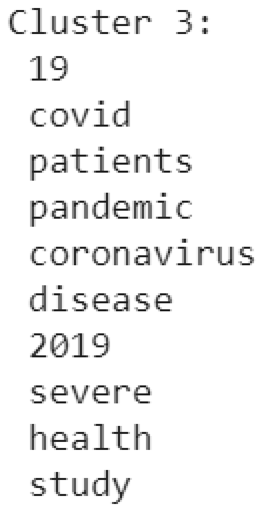


```
6                   serological tests for coronavirus
15    how long does coronavirus survive on surfaces
39                              coronavirus mutations
2                                coronavirus immunity
14                           coronavirus outside body
31                               coronavirus subtypes
1             coronavirus response to weather changes
30             difference between coronavirus and flu
49                          mRNA vaccine coronavirus
19                     coronavirus and ACE inhibitors
8                               coronavirus in Canada
13                         coronavirus super spreaders
18         what alcohol sanitizer kills coronavirus
32                      coronavirus vaccine candidates
4                           animal models of COVID-19
```

What pops out for me is that the queries about coronavirus can be from a long time ago but the queries about covid-19 tend to be much later. Obviously this is because covid 19 only arrived last year, whereas coronaviruses have been around since 1965. One thought I had was to remove the words coronavirus and covid-19 from the queries, essentially taking out the words which are probably biasing the date a little.

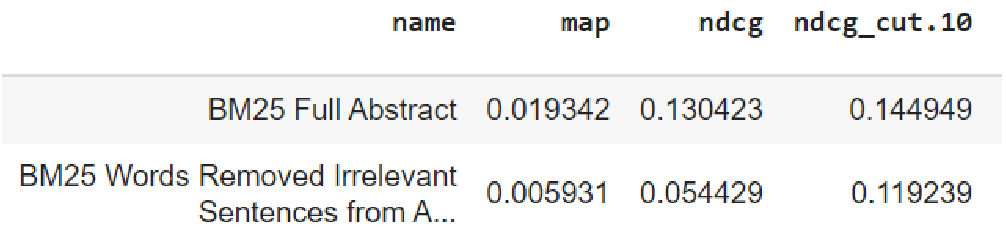
- I split the CoronaCentral dataset into train, validate and test and have set these aside for future use.

- I used the train dataset to train my model for classifying sentences, trying to identify irrelevant sentences. This was trained on around 145000 sentences and came up with the following cluster, which I think could represent the generic sentences quite well. It has around 141 tokens in it, but here is the top 10:



```
Cluster 3:
 19
 covid
 patients
 pandemic
 coronavirus
 disease
 2019
 severe
 health
 study
```

  I then classified all the sentences in the abstracts in the test data and removed them if they were from cluster 3. This resulted in the following, which shows the technique did not improve retrieval.



name	map	ndcg	ndcg_cut.10
BM25 Full Abstract	0.019342	0.130423	0.144949
BM25 Words Removed Irrelevant Sentences from A...	0.005931	0.054429	0.119239

  I decided to have a deeper look at each of the queries and see what was happening. It works quite well for the first query - coronavirus origin. But then afterwards, when you actually read the titles of the papers being retrieved it is not working very well at all. Therefore, I think this idea of removing boilerplates could potentially be removing important information for other queries and we have just become too focussed on answering one query.

- Unfortunately, when we remove the tokens “coronavirus” and “covid-19” from the abstracts, this does not produce better results. It very marginally makes things worse.
- One thought I have had - are we getting poor results because the documents which were labelled relevant in CORD-19 aren’t actually in CoronaCentral? That way when I evaluate the retrieval on CoronaCentral using the CORD-19 benchmarks it is performing

sub-par. So, I took a look at the number of documents that are classed as relevant, and checked how many of those are in coronacentral. There are around 70000 relevant documents and CoronaCentral only has around 18000 of them. So, this makes it clearer why the metrics are so poor for CoronaCentral as opposed to CORD-19. The annotations are not balanced well so it might be necessary to come up with a new way of evaluating CoronaCentral queries.

- I looked into reviews of CORD-19 - some interesting stuff out there actually. General takeaway is that there is a lot of irrelevance in CORD-19, but it does a pretty good job of encompassing all the coronavirus literature that is available. There is also a mention that a lot of the discussion is done online so one paper even recommends Altmetrics!

Week 5 - Pre-Meeting Status Report

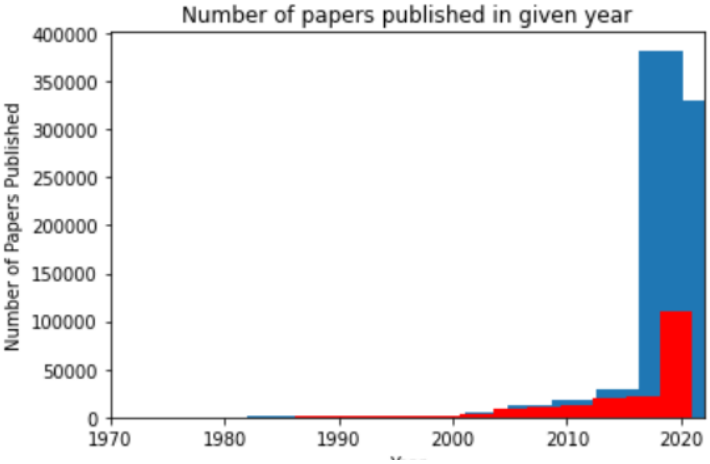
What was my plan for the week?

- Continue research before coming up with more concrete plan
- Experiment with removing irrelevant background information from abstracts
- Experiment with date filtering
- Look at mapping CoronaCentral documents to their associated CORD-19 documents

What have I done throughout the week?

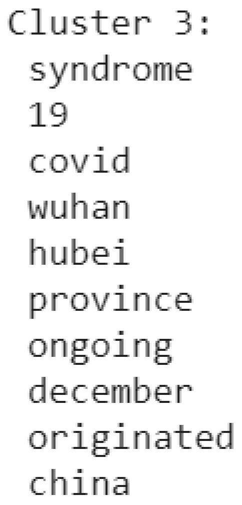
- I have formed a narrative that I think works quite well for my research and come up with a rough structure of how I think my literature review could work. I've started writing a very rough draft following this structure (around 800 words).

- I had a look at date filtering and here is what I found:



  The blue represents all the papers in CORD-19, the red represents only the papers which were classed as relevant in TREC-COVID. Obviously, we can see the more recent the paper, the more likely it is to be relevant. What I also thought was strange, was the fact that the majority of papers before 2000 were actually classed as relevant. This was shown by the fact that when I ran experiments on a subset of CORD-19, only documents post 2020, the performance was actually very slightly degraded.

- I also had a look at removing irrelevant information from the abstracts, the "boilerplate" background information on COVID-19. From what I could see, a lot of the documents retrieved for the first query "covid origin", it was picking up very strongly on the "origin" token which was causing confusion. I looked at DeepCT, which took emphasis away from the generic information in the abstracts and marginally improved results. I also tried expanding the queries with DeepCT which I know is a strange idea and didn't really work, but might be worth exploring further. I tried some k-means clustering on the abstract sentences. I thought we could train a clustering model, which identified the following cluster which seems generic:



```
Cluster 3:
 syndrome
 19
 covid
 wuhan
 hubei
 province
 ongoing
 december
 originated
 china
```

I thought we could classify each sentence in the abstracts and remove them if they are put in this cluster.

- Finally, I used the amended CoronaCentral data that you sent me to map back to CORD-19 documents. Works well to be able to evaluate.

- I do not have any specific questions at the moment.

Week 4 - Pre-Meeting Status Report

What was my plan for the week?

- Continue reading research papers, specifically looking into tools such as doc2query.
- Continue working through the pyterrier tutorials
- Take a deeper look at the CoronaCentral dataset, specifically areas such as altmetrics.

What have I done throughout the week?

- I have completed the pyterrier tutorials which have been very useful in my research. The tutorial on doc2query and DeepCT was especially useful and prompted me to explore further information about them.
- I have taken a look at altmetrics and read a paper written by the startup company themselves. I definitely think this is an avenue which can be explored further.
- I have researched more into the reranking stage, looking at pyterrier solutions for TREC challenges which utilised BERT-based models for this. I had a look at ColBERT and SciBert and how they worked on the CORD-19 data. Since we don't have much training data, it might be better to use a sequence-to-sequence model such as T5?
- I have been working on pulling all this research together into a narrative so that it is easier to write my literature review when the time comes.
- I have been playing about with the CoronaCentral dataset, trying to query it with some pyterrier techniques - both to consolidate my knowledge of pyterrier and also get me comfortable manipulating the dataset.

A heads up on some questions I have:

- Do you have any thoughts on how we can evaluate the effectiveness of the system when searching on the CoronaCentral dataset? TREC has already been manually annotated by judges, so is this something I might have to consider doing?
  Corduid to map back to cord-19
- How would it be best to go about getting Altmetric scores for documents retrieved in TREC-COVID solutions? Does the school have an account already registered for them?

Week 3 - Pre-Meeting Status Report

What was my plan for the week?

- Set up version control and any other admin along with it.
- Generally research the CoronaCentral website, TREC-COVID challenge and any other resources which will be useful in planning for the project, in preparation for a literature review.

What have I done throughout the week?

- I have set up version control.
- I have become familiar with the CoronaCentral website and read the corresponding paper.
- I have become familiar with the TREC-COVID challenge and read the corresponding paper.
- I have read several papers related to the TREC-COVID challenge in order to see how others combatted the challenge and what methods they used.
- I have worked my way through some of the PyTerrier tutorials on Colab in order to become acquainted with a tool which might be useful for the project. It teaches you how to use two-stage information retrieval on the CORD-19 dataset using BM25 then Colbert which seems very similar to some of the strategies utilised by high-performing TREC-COVID participants.

A heads up on some questions I have:

- Would you recommend writing up a literature review while doing this background research, or waiting until the second semester when writing the dissertation?
- Do you think it would be a good idea to get in touch with Craig McDonald to find some background information about Terrier and what his experiences have been applying it to CORD-19?