Case Studies in Data Science (COSC2669)

WIL Project: Final Report

Group 17

Akar Gupta, Akash Sunil Nirantar, Anurag Dinesh Karmarkar, Chinmay Pradeep Karangutkar, Cleon Ozzie Rodrigues, Rahul Sanjay Halappanavar

RMIT University

Akar(S3808546@student.rmit.edu.au)

Akash (S3813209@student.rmit.edu.au)

Anurag(S3829070@student.rmit.edu.au)

Chinmay(S3827733@student.rmit.edu.au)

Cleon(S3826800@student.rmit.edu.au)

Rahul(S3815553@student.rmit.edu.au)

17th October 2020

Table of Contents

Member	% Contribution
Akar (S3808546@student.rmit.edu.au)	**16.67**
Akash (S3813209@student.rmit.edu.au)	**16.67**
Anurag (S3829070@student.rmit.edu.au)	**16.67**
Chinmay (S3827733@student.rmit.edu.au)	**16.67**
Cleon (S3826800@student.rmit.edu.au)	**16.67**
Rahul (S3815553@student.rmit.edu.au)	**16.67**

Abstract

Coronavirus is a group of RNA viruses that can cause illness, which can vary from common cold and cough to sometimes more severe disease. SARS-CoV-2 (n-coronavirus) or popularly known as COVID-19 is the new virus of the coronavirus family, which first discovered in 2019, which has not been identified in humans before. It was declared as Pandemic by WHO due to high rate spreads throughout the world. Currently (on the date 10 August 2020), this leads to a total of 750K+ Deaths across the globe. This Project is an effort to evaluate search algorithms and systems for helping scientists, clinicians, policy makers, and others manage the existing and rapidly growing corpus of scientific literature related to COVID-19, and to discover methods that will assist with managing scientific information in future global biomedical crises using the data provided by The Semantic Scholar team at the Allen Institute of AI Built[1].

Introduction

In the city of Wuhan there were many pneumonia cases. The cause of many pneumonia cases was unknown. This was bought to the notice of the World Health Organization in December,2019. This virus was never identified in humans before. This virus initially known as new virus was then identified and renamed to 2019 novel coronavirus. This was then further renamed to Coronavirus Disease 2019 (COVID-19) by WHO in February of 2020. The virus is referred to as SARS-CoV-2 and the associated disease is COVID-19. Coronavirus is a zoonotic, which means that it is transmitted between humans and animals.

Coronavirus has mild to moderate symptoms which are almost like seasonal flu. Some other symptoms are fever, cough, shortness of breath, fatigue, breathing difficulties, respiratory symptoms, and sore throat. Sometimes people may even show severe symptoms such as pneumonia, sepsis, etc. Older people and people with severe illness such as diabetes, cancer, etc tend to be affected more by this virus.

Transmission of this virus takes place through large respiratory droplets and direct or indirect contact with infected secretions. It takes about10-14 days for they symptoms to show. Once detected, there is a quarantine period of 14 days to get totally cured[4].

This virus started spreading worldwide and the number of cases were increasing day by day. Later in February, there were outbreaks in Middle East and Asian countries. Soon a couple of deaths were recorded in Europe as well. Countries started to impose restrictions which further went to turn into locking down the entire city. Cases started to rise drastically worldwide. All non-essential things were banned. People could exit their houses only for essential things such as groceries, medical services etc.

Currently, United States of America is the state with the greatest number of cases with a count of 8.09 million followed by India and so on. Overall, there are almost 40 million cases and 1.1 million deaths. Out of which almost 30 million people have recovered[9].

Problem statement

The period of this virus has been very tumultuous to the research community. The researchers find it very difficult to conduct their research during this pandemic. The organization Research Australia invited all the researchers on their database to participate in the survey[11]. An overwhelming response was received. It was found that 79·6% of participants stated that their research was affected due to the COVID19 pandemic. A further 9·7% of participants indicated that their research is likely to be affected in the future. This is an indication that the research community is facing a lot of difficulties. The current pandemic scenario has had devastating consequences on the research for the COVID 19 vaccine.

The laboratories are currently closed or partially open. The infrastructure is partially compromised. The researchers must go through a lot of unnecessary articles to find the right one. It is very difficult to examine the whole of the internet to find relevant research.

COVID19 has had a huge and long-lasting impact on the research. Research related to various disciplines has been curtailed by the pandemic. Most of the clinical trials have been put on hold indefinitely. These do not include the ones being conducted to develop the COVID19 vaccine. On-going clinical trials have been restricted to home administration to restrict the infections[12]. The scientific community has been divided as the resources towards the research have been divided towards the pandemic response and hence the research community is falling short of the resources. Our solution can help the researchers working from home. We have created a search engine to manage the existing and rapidly growing corpus of scientific literature related to COVID-19, and to discover methods that will assist with managing scientific information in future global biomedical crises using the data provided by The Semantic Scholar team at the Allen Institute of AI Built.

Data Preparation

Information Recovery (IR) is the process of extracting information system data that are important to the need for information from a database of such data[10]. To retrieve the required information, we need to train the system on the existing textual data and gear up for searching the relevant information against a user query.
Firstly, we scrapped the data by removing the insignificant columns and by deleting the missing values. At last, we had data for 70k articles, and we used them to explore the data.
During the exploration, we visualized the columns year, authors, title, and abstract. Due to this process, we understood the golden nuggets from the data, which helped us in selecting the best feature to train the TFIDF model.
We created the bag of words (BOW) and the dictionary using the data from the column **title**. We used the BOW to train our TFIDF model, which gives a specific weight to a word depending on its frequency.
Then lastly, we calculated the cosine similarities using the 'gensim' library and saved them into a sparse matrix.
Finally, we used this matrix to calculate the similarities between the user query and the document, and then we displayed the result using the search engine website.
We developed the search engine website using the Flask framework and integrated it with the python files.

Data collection and data dictionary

We have used the data provided by the Semantic Scholar team at the Allen Institute of AI built[1]. The Allen Institute for Artificial Intelligence created the data in collaborations with (AI2), the National Institute of Standards and Technology (NIST), the National Library of Medicine (NLM), Oregon Health & Science University (OHSU), and the University of Texas Health Science Center at Houston (UTHealth).

The dataset consists of every research article regarding diseases related to influenza published until now. The research articles are in the form of JSON and XML format. It also contains a metadata file comprising the data about their authors, titles, and article file locations.
The metadata comprises of the following columns: -

1. cord_uid: unique ID for an article
2. sha: unique code for an article
3. source_x: the source of the article
4. title: title of the article
5. DOI: date of issue
6. license: type of license
7. abstract: brief abstract about the article
8. publish time: published date and time
9. authors: authors of the article
10. journal: journal name
11. pdf_json: article pdf JSON file location
12. pmc_json: article PMC JSON file location
13. URL: URL of the article's website

Data preprocessing and scrapping

As we needed to create a search engine or an information retrieval system, we needed to work heavily on the metadata file. We started the data preprocessing by deleting the unwanted columns in our analysis, followed by scrapping some of the data. The metadata consisted of around 200k articles. Hence, it was impossible to process all this on a local PC. Therefore, we decided to scrap the data by 60% to build our MVP (Minimal Viable Product).

Steps to preprocess: -

1. We only kept the following columns in the metadata and discarded the rest - **cord_uid, title, abstract, publish_time, authors, journal, pdf_json_files,** and **URL**
2. We scanned the data for any errors and unexpected values. Fortunately, we didn't find any.
3. We deleted all those tuples who had missing values.

After performing all the above steps, we had around 70k rows and 8 columns.
The most important columns in the data were title, author, abstract, and URL.

Data exploration and feature selection

We needed to find crucial features/columns from the dataset, which will help us in building an information retrieval system. Users may search the articles by publish_year, by authors, or by the content of the research articles. So we decided to explore these columns and select the best one to build our model.

Publish year

We found out that the maximum number of articles was published in 2020, followed by 2019 and 2018. So, researchers released many papers recently.

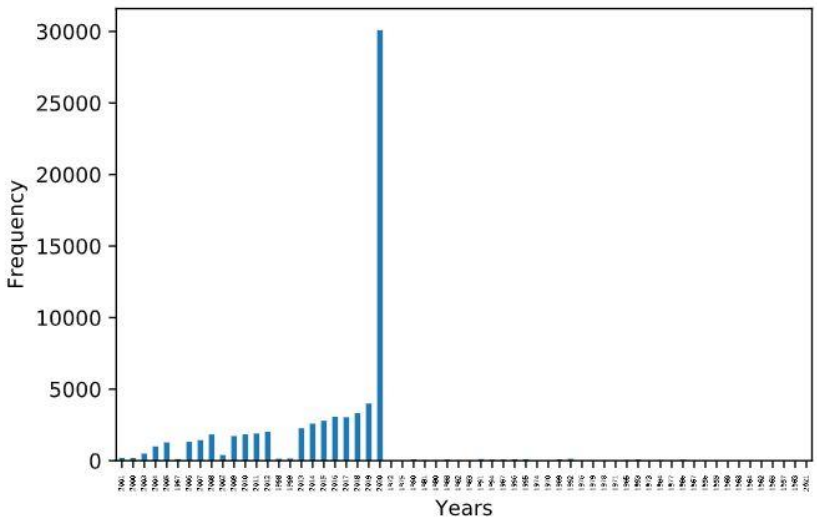

Figure 1 Years vs frequency of publishing articles

Article title

To visualize the column **title,** we decided to mark the occurrence of every word. After visualizing the occurrence graph, we concluded that the most occurring terms are the **virus**, **respiratory, coronavirus, infection, and patients**. It was one of the best features to build our model.

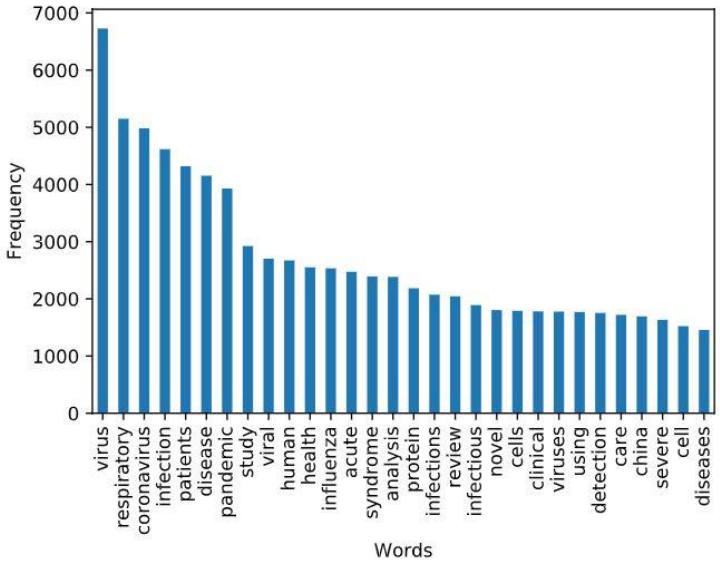

Figure 2 Words from titles and their frequency

Abstract

The abstract contains a summary of the corresponding article. After visualizing it, we concluded that the terms and their frequency found out are similar to the title. But the frequencies were higher due to a large number of words present in the summary of articles.

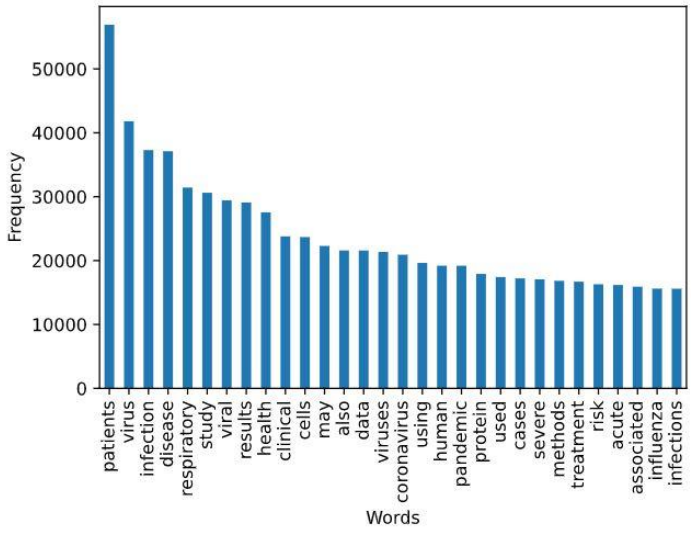

Figure 3 Words from abstract and their frequency

Inference and decisions

Title contains the required data to search about the articles and the train model could handle simple user queries. Abstract on the other hand had numerous words and a bit more complicated to process. After considering the complexity and the effectiveness of the system, we decided to select the **'title'** to train our model and use it for the minimal viable product - information retrieval system.

Data Modelling

After performing attribute selection, we will focus on Tokenization. So, we tokenized every title from metadata and saved it in serializable file. Tokenization basically refers to splitting up a larger body of text into smaller lines, words or even creating words for a non-English language. It has various types of tokenization such as Line Tokenization, Non-English Tokenization[5] and Word tokenization.

Later, we fetched the tokenized list of titles from .dat file (serializable file) and we split the document list and save it into a corpus. Thus, we got the list which was required for model building. Now, we have built a TF-IDF model.

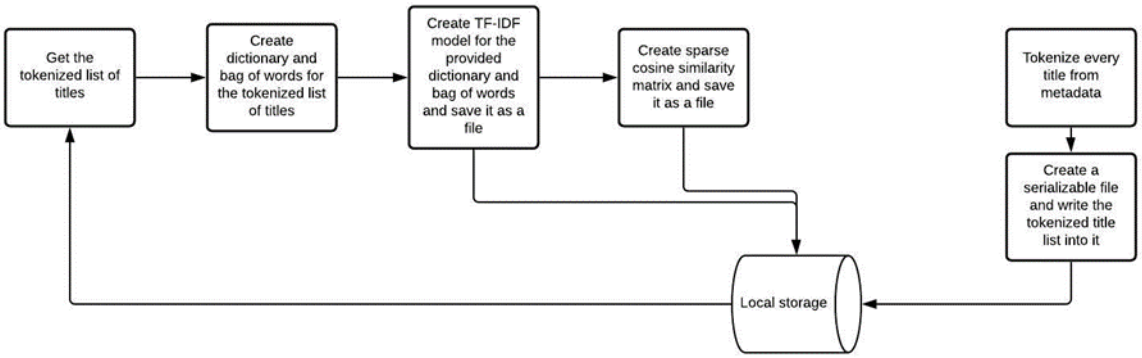

Figure 4 Creating Bag of Words, Dictionary and TF-IDF Model

TF-IDF

The term TF stands for "term frequency" while the term IDF stands for the "inverse document frequency". Its basically a statistical measure that evaluates how relevant a word is to a document in a collection of documents steps in tf-idf theory. Tf-IDF is basically used in automated text analysis and is very useful for scoring words in machine learning algorithms for Natural Language Processing (NLP). Tokenization is pre-requisite for TF-IDF model. Once you have tokenized the sentences, the next step is to find the TF-IDF value for each word in the sentence. It is important to mention that the IDF value for a word remains the same throughout all the documents as it depends upon the total number of documents. On the other hand, TF values of a word differ from document to document. Now we have IDF values of all the words, along with TF values of every word across the sentences. The next step is to simply multiply IDF values with TF values. Thus, we calculated TF-IDF weights and saved it into a file[8].

Gensim

Gensim is Natural Language Processing's model which basically does topic modelling for humans. It is based on topic modelling, which is technique to extract topics from huge texts. Tf-Idf is computed by multiplying a local component like term frequency (TF) with a global component, that is, inverse document frequency (IDF) and optionally normalizing the result to unit length. There are multiple variations of formulas for TF and IDF existing. Gensim uses the information retrieval system that can be used to implement these variations for TF-IDF. Genism has many uses, most importantly in automated text analysis, and is very useful for scoring words in machine learning algorithms for Natural Language Processing (NLP). Gensim is designed to handle large text collections using data streaming

and incremental online algorithms, which differentiates it from most other machine learning software packages that target only in-memory processing[7].

Cosine similarity matrix

Cosine similarity is a metric used to determine how similar the documents are irrespective of their size. Mathematically, it measures the cosine of the angle between two vectors projected in a multi-dimensional space. Our objective was to quantitatively estimate the similarity between the two words which appeared at least 2 times in the corpus. Thus, we stored all things into a newly created dictionary and push into local storage[6].

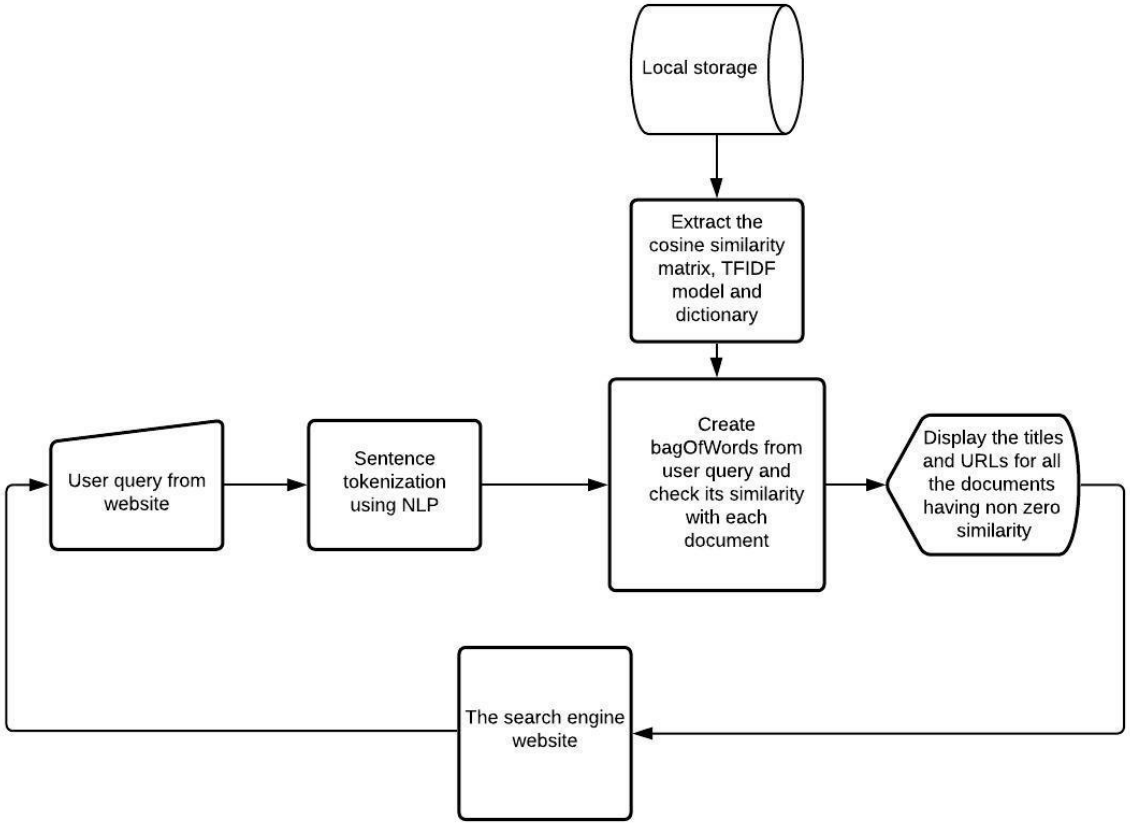

Figure 5 Processing the User Query

Processing user query

We have built a search Engine, which has a proper search box where user can enter his query, Thus, his query will be fetched by tokenizer and using NLP it will be tokenized. After tokenization, bag of words will be created based on user query. Now the query will be used by similarity model and similar content related to the processed user query will be returned. Therefore, all the data which is mostly containing the URL for the documents with non-zero similarity will be returned to the user dashboard.

UI Designing

We designed a user interactive search dashboard using Flask. Flask is basically a lightweight web development framework. It is basically used to scale upto complex application. The basic working of Flask is based on HTML and CS. We created the dashboard using Flask's visual objects. We have a search bar where the user can enter his query and after pressing search button the dashboard fetches the results and displays to user.

UI Snippets

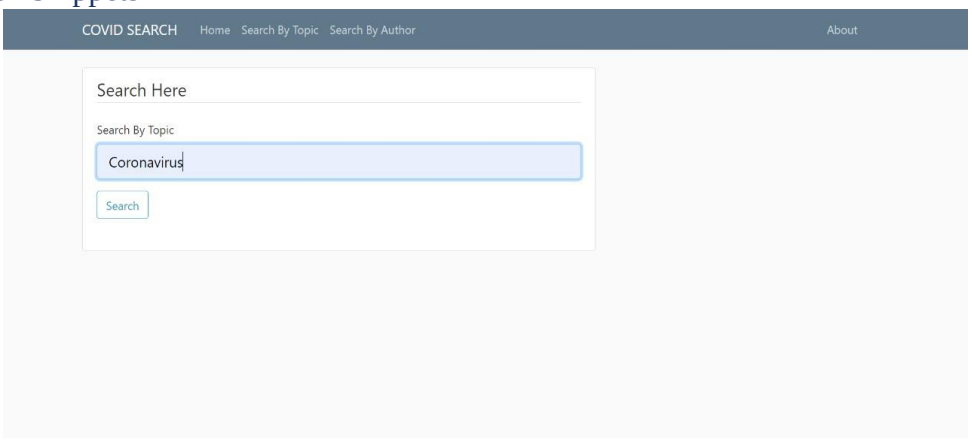

Figure 6 Search Page

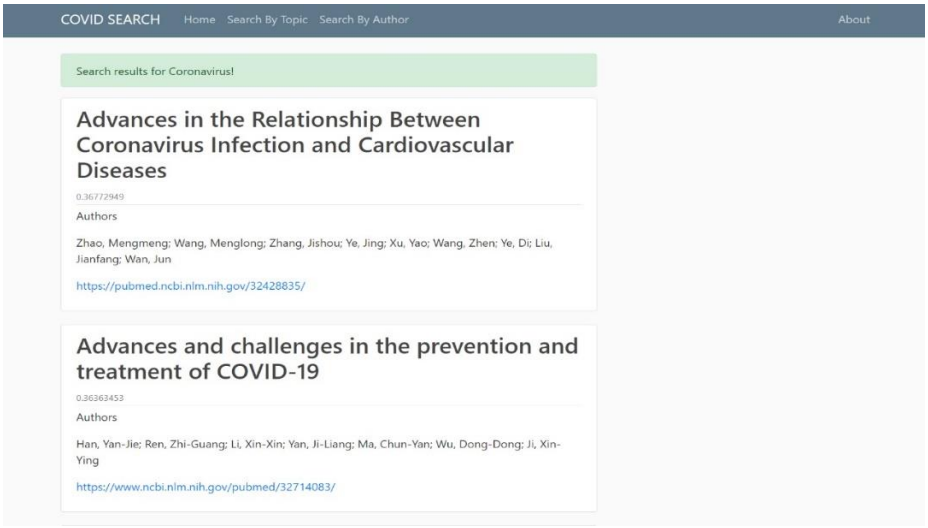

Figure 7 Results Dashboard

Project Management Approach:

The project used an Agile Methodology which consisted of 5 sprints. Each sprint was of 2 weeks and weekly we had 2 stand-ups for task updates and to discuss any blockers.

Sprint Information:

Sprint 1 (1st August - 16th August)

Motivation of this Sprint was to decide what our MVP will be and all technologies that we must work on. This sprint was dedicated to get familiar with Git , NLP and python libraries ,Flask for UI.

Sprint 2 (17th August - 30th August)

Motivation of this sprint was to write milestone 1 report and the decide the role responsibility of each team member. Secondly, we did pre-process tasks in this sprint and created the final data for analysis. For more information on this please refer section(data preprocessing)

Sprint 3 (31st August - 13th September)

This sprint was dedicated to coding and creating the business logic for our MVP. For more information on this please refer section(data modelling)

Sprint 4 (14th September - 27th September)

This Sprint was dedicated for cater to all the existing issues in our business logic and make it more accurate. In this sprint we made our UI interface too.

Sprint 5 (28th September - 19th October)

This sprint mainly had two tasks making presentation and report submission. With all the other final assignment and exams, we increased the span of this sprint from 2 weeks.

	August 2020		September 2020		October 2020
Sprint No	1	2	3	4	5
Sprints Motivation	MVP identification	Milestone 1 Report	Analysis and Logic building Code	UI development	Final Presentation & Final Report Submission

Team Capabilities

We all are highly motivated post graduate pursuing Master of Data Science. Topic of our project selection was chosen in a way that we all can gain knowledge about new Data mining techniques and any new software which we can work on. We dedicated a whole sprint in identifying & learning about any new language necessary for progressing further in the project. We divided our work in two major sections. One dedicated to process metadata and other to process the user query and creating UI. Team was divided in groups of two 1st group took care or metadata processing and 2nd group took care of User query and UI.

Workflow

As discussed earlier we did our project as an agile methodology and two-week sprints. We created a channel on the Microsoft teams for discussion and file sharing. Each week we had 2 stand ups. Each Tuesday we decided on the tasks to be done for the week. Tasks were given to each member. In the Saturday's stand up we discussed any personal blockers, or any help required. If required member used to demonstrate their work.

We created a Kanban board with 4 work buckets to-do, in progress, completed and backlog. Tuesday's tasks were created in to-do bucket and assigned to members. Person who started working on his task used to move it to in progress and after completing moved the task to completed.

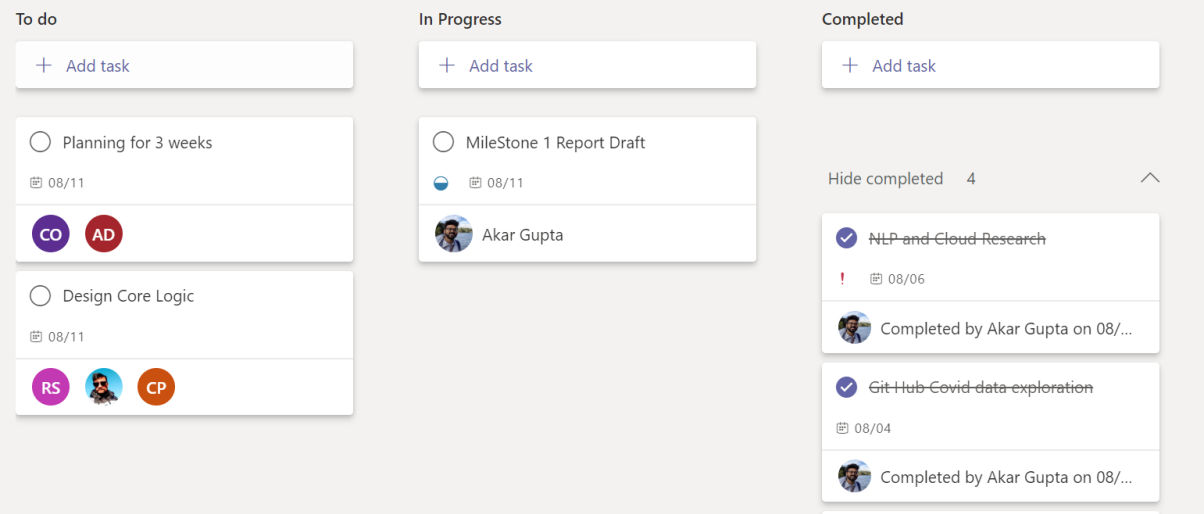

Figure 8 Describing the Motivations of each sprint

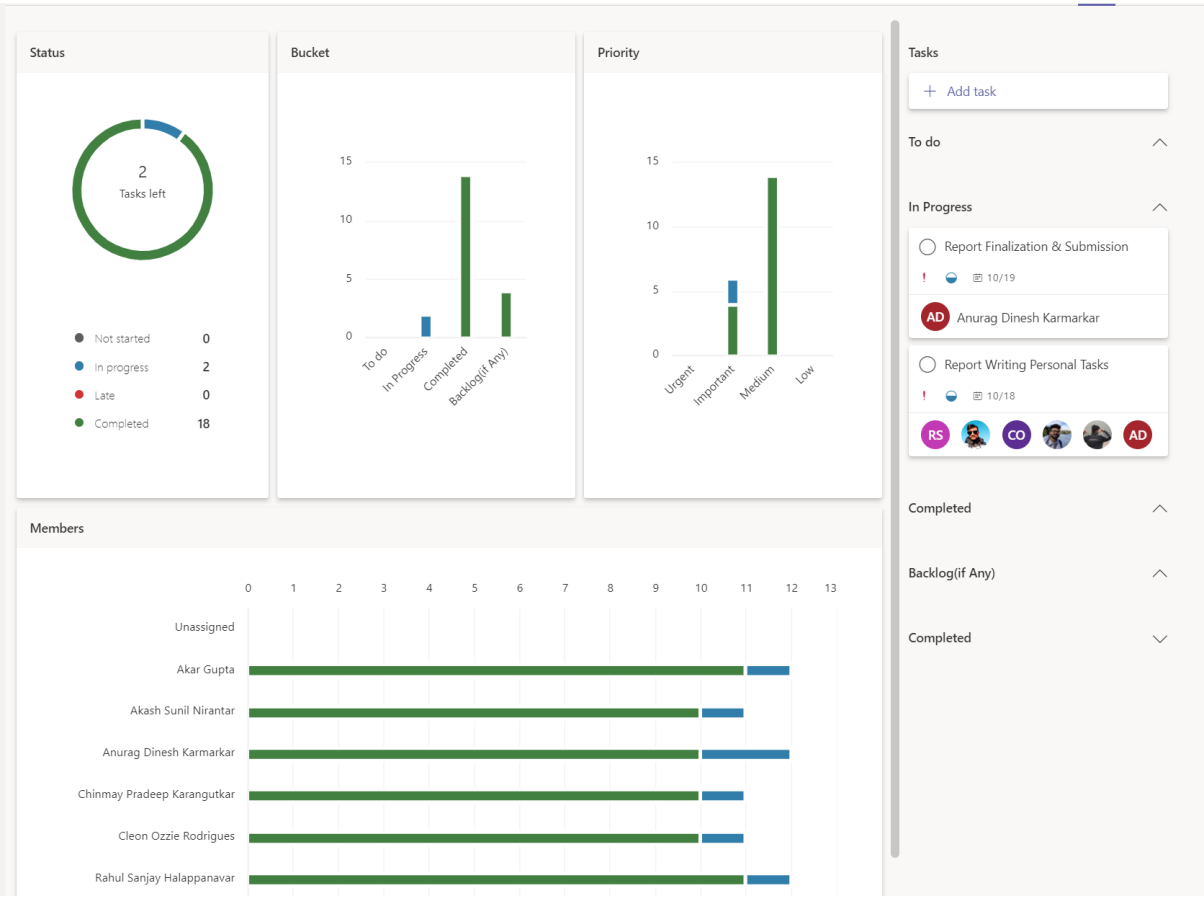

Figure 9 Describing the No of tasks done for project

Retrospect

As per agile methodology after each sprint we need to have a retrospect meeting. But due to time crunch we moved it at the end of last sprint. Here a highlight of what we loved, learned, issues and what can be learned.

Loved	Loathed	Learned	To Learn
Lots to learn Good Similarity index results. Able to process such big metadata file Team Work	Processing similarities for titles and abstract took a large time. Data was too big around 8GB and hence we had to scrape it.	Learned using Git Working with multiple people on a main Branch using Git. Studying about NLP, Gensim, Similarities. Working with Flask for creating UI	Publishing our site online Rectifying the user query for spelling and grammatical mistakes

Conclusion/Wrap Up/Next Steps

This report describes the effort we put in to create a COVID19 search system that could extract the relevant COVID19 related research papers from a huge database of scientific research papers from a given query. Our solution is mainly based on natural language processing (NLP) and machine learning. Currently we believe that our designed system can process queries given to it and output the relevant papers by comparing the queries with thousands of titles of research papers. This might limit the systems knowledge to only the titles which may not contain the story of the whole paper sometimes. Hence, in future we would like to make the following changes:

1. To focus the search engine algorithm more on the content or body of the paper to give more relevant results and much accurate too rather than just titles.
2. Also, in terms of scalability this project needs to be shifted to cloud services such as AWS or Azure, so that the researchers have the access to such tools worldwide.
3. Further improvements in terms of accessibility could be by shifting the use case to a dedicated chatbot for COVID19 related queries.
4. Integration with Government Application like COVIDSAFE by using Android and IOS APIs.

Coronavirus being such an important matter in 2020, we think that this solution can prove to be useful to fight against the global pandemic by helping researchers and scientists to answer their queries by referring each other's work at much faster pace using this search engine. This could lead to a potential cure even earlier than expected.

References

1. Kaggle.com. 2020. COVID-19 Open Research Dataset Challenge (CORD-19). [online] Available at: <https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge> [Accessed 17 August 2020].
2. Kaggle.com. 2020. CORD Research Engine - Search And Similarity. [online] Available at: <https://www.kaggle.com/dgunning/cord-research-engine-search-and-similarity> [Accessed 15 August 2020].
3. Discovid.ai. 2020. [online] Available at: <https://discovid.ai/> [Accessed 12 August 2020].
4. Physiopedia. 2020. Coronavirus Disease (COVID-19). [online] Available at: <https://www.physio-pedia.com/Coronavirus_Disease_(COVID-19)> [Accessed 17 October 2020].
5. Tutorialspoint.com. 2020. Python - Tokenization - Tutorialspoint. [online] Available at: <https://www.tutorialspoint.com/python_text_processing/python_tokenization.htm#:~:text=In%20Python%20tokenization%20basically%20refers,in%20programs%20as%20shown%20below> [Accessed 11 August 2020].
6. Prabhakaran, S., 2020. Cosine Similarity - Understanding The Math And How It Works? (With Python). [online] ML+. Available at: <https://www.machinelearningplus.com/nlp/cosine-similarity/> [Accessed 13 August 2020].
7. Prabhakaran, S., 2020. Gensim Tutorial - A Complete Beginners Guide - ML+. [online] ML+. Available at: <https://www.machinelearningplus.com/nlp/gensim-tutorial/> [Accessed 12 August 2020].
8. Malik, U., 2020. Python For NLP: Creating TF-IDF Model From Scratch. [online] Stack Abuse. Available at: <https://stackabuse.com/python-for-nlp-creating-tf-idf-model-from-scratch/> [Accessed 11 August 2020].
9. Kantis, C., 2020. UPDATED: Timeline Of The Coronavirus | Think Global Health. [online] Council on Foreign Relations. Available at: <https://www.thinkglobalhealth.org/article/updated-timeline-coronavirus> [Accessed 10 August 2020].
10. En.wikipedia.org. 2020. Information Retrieval. [online] Available at: <https://en.wikipedia.org/wiki/Information_retrieval> [Accessed 10 August 2020].
11. Peeters, A., Mullins, G., Becker, D., Orellana, L. and Livingston, P., 2020. COVID-19's impact on Australia's health research workforce. The Lancet, 396(10249), p.461.
12. Weiner, D., Balasubramaniam, V., Shah, S. and Javier, J., 2020. COVID-19 impact on research, lessons learned from COVID-19 research, implications for pediatric research. Pediatric Research, 88(2), pp.148-150.