Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

School of Computer Science and Statistics

Fighting the Infodemic: Finding Important COVID-19 Related Imagery

Sean Brieffies

Supervisor: Prof. Khurshid Ahmad

April 2021

A Final Year Project

Presented to the University of Dublin, Trinity College

in partial fulfilment of the requirements for the degree of

B.A. (Mod.) Computer Science

Declaration

I, the undersigned, declare that this work has not previously been submitted as an exercise for a degree at this, or any other University, and that unless otherwise stated, is my own work.

Sean Brieffies

April 30, 2021

Permission to Lend and/or Copy

I, the undersigned, agree that Trinity College Library may lend or copy this thesis upon request.

Sean Brieffies

April 30, 2021

Abstract

The COVID-19 pandemic has resulted in an explosion of new information. This includes academic literature as well as deliberately false information designed to undermine public health responses. The World Health Organisation has declared this explosion of new information as an “infodemic”. During such a rapidly changing situation, there is a clear need to identify the important information from the unimportant. Here, a system is proposed to find important imagery related to COVID-19 to assist scientists and academics in their work. Importance is measured using the underlying citation network present among academic publications, this is referred to as the “impact” of a publication. The impact of a publication is used to lend credibility to the images contained inside it. Publications which have been identified as being high-impact are retrieved from the PubMed Central Open Access Subset and images are then extracted from these publications, indexed and made searchable by the corresponding figure caption. A user can then search this index to find high-impact images which match their search query. Three medical researchers evaluated the system and noted the usefulness of such a system as well as providing useful feedback for future improvements. Lastly, a new metric called the “Image Relevance Index” is proposed. This metric combines a number of common bibliometrics to more directly rank biological images contained within publications, rather than purely using the impact of the parent publication. This topic could be expanded upon in future research.

Acknowledgments

I wish to thank a number of people who have helped me this year.

Firstly, I would like to thank my supervisor Professor Khurhsid Ahmad for taking me on and introducing me to this area. I sincerely appreciate your time throughout the past few months and for constantly steering me in the right direction and answering my many questions.

I would also like to thank Dr. Aamir Ahmed, Dr. Nollaig Burke and Dr. Liam Townsend for taking the time to speak with me and providing invaluable feedback to help me improve my project.

I would like to thank my friends for their support during the year.

Last but not least, I would like to thank my family for their support throughout my education and life in general, without you none of this would be possible.

Sean Brieffies

University of Dublin, Trinity College
April 2021

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Project Description & Motivation

On March 13th 2020, in collaboration with the White House Office of Science and Technology Policy, the National Library of Medicine, the Chan Zuckerberg Initiative, Microsoft Research, Georgetown University Centre for Security and Emerging Technology and Kaggle, the Allen Institute for AI released the COVID-19 Open Research Dataset (CORD-19) (Wang et al. 2020b). This dataset is a collection of publications and preprints on COVID-19 and other historical coronaviruses. At the time of release, this dataset contained approximately 29,500 scholarly articles. CORD-19 is updated weekly and as of March 29th 2021, there are now 468,406 articles in the dataset; this represents an average increase of 1,230 articles per day since the dataset was first released.

The World Health Organisation (WHO) has described these newfound excesses of information, coupled with deliberate attempts to spread false information which undermines public health advice, as an "infodemic" (World Health Organisation 2020b). They say that the infodemic:

- Causes confusion and risk-taking behaviours such as using unproven medical treatments.
- Leads to mistrust in health authorities and undermines the public health response.
- Costs Lives.

The spread of false information, often colloquially referred to as "fake news", is commonly seen as a political problem, most famously in relation to the 2016 U.S. Presidential Election. This current wave of fake news could have far worse consequences, costing lives rather than votes. It has been observed that a clear correlation exists between susceptibility to misinformation, compliance with public health advice and willingness to undergo vaccination (Roozenbeek et al. 2020). This presents a clear and urgent need to detect these acts of disinformation and to provide reliable, scientifically accurate information to scientists and the wider public.

To combat the infodemic, the WHO notes the need for “Free, reliable, trustworthy, factual, multilingual, targeted, accurate, clear and science-based information” (World Health Organisation 2020a). The goal of this project is to help in fighting this infodemic by creating a system which will help scientists to find important medical images related to COVID-19 e.g., microscopic imagery, X-rays, CT scans to assist their research and perhaps their own fight against the infodemic. To do this, these images will have to be collected, organised, and made readily available to those who search for them. A methodology will need to be developed to bring credibility to these images and to make their retrieval as fast and reliable as possible.

1.2 Ethical Considerations

There are several ethical issues which must be considered in designing this system. Firstly, any images stored and distributed by the system must have been made publicly available or failing this, the appropriate permissions received from the copyright holder to store and distribute the image. Secondly, none of the images can contain information which could potentially be used to identify the patient, their location or any of their personal details.

In collecting and ranking the credibility of images, the utmost care must be taken to ensure that certain genders and/or races are not favoured over others and that there is equal opportunity for the work of these demographics to be presented to the user. There is the possibility that pre-existing underlying imbalances in the academic world could skew the probability of encountering work from a particular demographic. For example, according to the UNESCO Institute for Statistics as of 2016, only 29.3% of researchers are women (UNESCO Institute for Statistics 2019). Furthermore, there is evidence which suggests that the pandemic is disproportionally affecting female academics. Frederickson (2020) quantified this effect by comparing the number of preprints submitted by male and female authors on arXiv and bioRxiv year-on-year. This showed a greater relative increase in the number of submissions by male academics compared to their female colleagues. These underlying factors must be acknowledged and accounted for where possible.

COVID-19 has affected all corners and people of the world and so any system which aims to help combat the infodemic stemming from the virus, or any other problem it has caused, should be designed in a way that can best represent and be useful to all people, regardless of gender, race, or nationality.

1.3 Report Outline

1.3.1 Background

This chapter discusses existing ways in which scientists find images, text searching methods, and how scientific publications can be ranked using the underlying citation network present within academia.

1.3.2 Implementation

The Implementation Chapter provides a detailed description of how images are collected and indexed from scientific publications and then made available for the user to search through a Graphical User Interface (GUI).

1.3.3 Evaluation

The Evaluation Chapter focuses on interviews with medical-focused academics who evaluated the system as well as a discussion on the proposed Image Relevance Index.

1.3.4 Conclusions & Future Work

Finally, building on the evaluation, there is a view to future work in this area as well as recapping the work undertaken and drawing conclusions from it.

Chapter 2

Background

2.1 Image Retrieval

As the old adage goes "a picture is worth a thousand words". Images are a very dense form of information and play an important role in conveying information in scientific publications. In their guide to authors, Springer Publishing note that figures are often the best way to quickly convey large amounts of complex information to the reader, and play an important role in attracting readers (Springer n.d.). They also note that readers will frequently only look at figures when deciding whether or not to read a publication.

The medical and academic communities have made a concerted effort to share potentially useful information to combat the ongoing pandemic. There is now an abundance of medical imaging datasets available for public use (European Institute for Biomedical Imaging Research 2021) (Region et al. 2020) (The Cancer Imaging Archive 2021) (Cohen et al. 2020). These datasets mainly consist of labelled CT and radiographic scans and are intended primarily for training machine learning models to aid in diagnostics, such as the COVID-Net deep learning model (Wang et al. 2020). Many of these models have also been made freely available to the public (Stanford University Center for Artificial Intelligence in Medicine & Imaging 2021), again highlighting the remarkable collaboration which currently exists within the medical and academic communities.

Despite the apparent abundance of images, there is not an abundance of ways to search for COVID-19 related images directly. The aforementioned imaging datasets are mainly tailored towards training machine learning models, with a minimal amount of additional information included which could enable effective searching. Most existing tools which aim to sift through the growing amounts of new literature do not focus on finding images, but rather focus on finding full-text publications (Bhatia et al. 2020) (Google 2020) (Trewartha et al. 2020) or extracting textual information from them (Allen Institute for

AI 2020). With the exception of the search tool by Bhatia et al. (2020), which allows results to be sorted by citation counts, these other tools do not allow for sorting using any citation-based metric. Even with this, raw citation counts are somewhat problematic as they are biased against newer publications which have not had enough time to acrue citations. Alternatives to this metric will be discussed in Section 2.2.

Over the years there have been several systems designed to allow users to search for images in biomedical publications. The Biotext Search Engine (Hearst et al. 2007), Yale Image Finder(YIF) (Xu et al. 2008) and Open-i (Demner-Fushman et al. 2012) are the most notable of these. Biotext and YIF both allowed for search over figure captions, with YIF also allowing for search over text embedded within an image, abstracts and publication titles. These two systems have been discontinued in recent years. The U.S. National Library of Medicine's Open-i is still in operation. Open-i supports image search over captions, abstracts, titles and authors. As well as extracting these textual features, it also extracts visual features from images. These visual features are used to classify the type of image e.g. 'X-ray', 'CT', 'graphic', the user can then filter results based on these classifications. This also allows for content-based image search whereby the user can enter an image as a query and similar looking images will be returned. While Open-i does supports various image ranking methods such as 'oldest', 'newest', 'diagnosis', 'outcome', there is no way to rank images by any importance/impact metric so the issue of finding important images is still unsolved.

2.2 Ranking Academic Publications

In a rapidly-changing situation such as an infodemic, identifying high-quality information is of the utmost importance. While this is difficult in normal times, the explosion of new coronavirus literature, perhaps exasperated by the "publish or perish" mantra which has become pervasive in academia (Sarewitz 2016), makes this even more difficult. This raises the question: How do we decide what information is important? Perhaps an obvious answer is to let academics decide. Doing this directly would be impractical given the sheer volume of publications, but this can be done indirectly by using the underlying citation network to assess the impact of a publication within academia. While a citation does not necessitate quality, a publication could have exclusively critical citations, this criticised publication is still important in a way, perhaps "bad important" but important nonetheless. We could rank publication importance based on how many citations they receive from other publications. This is somewhat problematic however as citation counts are biased against newer publications which have not had enough time to accrue citations. This also treats all citations equally, taking no account of citation quality. Treating citations equally is especially problematic in a field where authors can self-cite to boost their rankings.

This problem is similar to the one faced by Page, Brin, Motwani and Winograd at the dawn of the World Wide Web in 1999. They sought to “bring order to the web” through their “PageRank” algorithm (Page et al. 1999). This was perhaps even more pertinent because anyone can make a website, academia at least has some barrier to entry. PageRank uses the idea that not all citations (or links in that context) are equal. Rather, a citation from a high-importance website is worth more than a citation from a low-importance website. PageRank does this by simulating a “random surfer”. Putting this in an academic context, the surfer is given a random paper from the citation network. The surfer keeps clicking on outbound citations, only stopping when they lose interest. They then start this process again on another random paper. The probability that the surfer visits a paper is its PageRank value. This promotes papers which are more central in the citation network and can therefore be viewed as being more important.

PageRank has been to shown to be very effective at measuring the importance of academic papers (Ma et al. 2008) (Chen et al. 2007) (Kanellos et al. 2019). PageRank has also inspired many spin-off algorithms. Some of these put a twist on the “random surfer” model, “Focused PageRank” models a “focused surfer” where the probability of choosing an outbound citation is proportional to its citation count (Krapivin and Marchese 2008). Others methods such as CiteRank (Walker et al. 2007) and Timed PageRank (Yu et al. 2005) incorporate time-awareness. Time-awareness decays the importance of a paper with time. This helps alleviate the aforementioned time-bias inherent in citation networks and is a useful way to find newer important papers.

In their evaluation of thirty-two different paper ranking methods, Kanellos et al. (2019) distinguished two forms of citation-based impact metric. The first is influence, this is the long-term impact of a paper. The second is popularity, which is the short-term or “current” impact of a paper. The evaluation concluded that the PageRank and Retained Adjacency Matrix (RAM) (Ghosh et al. 2011) methods performed best on calculating influence and popularity respectively. This study also birthed the PaperRanking library [1]. This library has since been used produce the “BIP4COVID19” dataset (Vergoulis et al. 2021). This dataset contains influence and popularity metrics for almost 300,000 coronavirus-related publications from the CORD-19 and LitCovid (Chen et al. 2020) datasets.

[1] `https://github.com/diwis/PaperRanking`

2.3 Searching With Text

A simple starting point for text search is to use a “Bag of Words” (BOW) model. This ignores the word order inherent in sentences, hence the “bag of words” name. A BOW model splits the text corpus into ’tokens’, these could be singular words, groups of words or both, and then counts how frequently each token appears in a given document. This could be applied to each document in the corpus and we would now know how frequently every word appeared in every document. These counts can then be used to match queries to documents by creating a BOW model for the query and matching the token counts to the closest BOW models in the document corpus. This overly-simple approach would cause issues however as there is no consideration of how ’special’ a word is. Common words which appear in every document, and therefore do nothing to differentiate one document from another, are valued the same as less common words which could reveal noteworthy documents.

To overcome this problem, the Term Frequency-Inverse Document Frequency (TF-IDF) of a query can be used instead. Term frequency is the number of times a token appears in a document. Document frequency is the number of times that term appears in all documents. The inverse of this document frequency is used because if a term does not appear often in the document corpus (has low document frequency), it is more likely to be an informative or differentiating term which could signal that a document is highly relevant to a given query.

TF-IDF was further improved and expanded upon by Robertson et al. (1995), resulting in “Best Match 25 (BM25)”. In addition to rewarding term frequency and penalising document frequency like TF-IDF, BM25 accounts for document length and term frequency saturation, making it more suitable to rank documents based on their relevance to a search query. Accounting for document length is important as longer documents, by their very nature, are more likely to have higher term frequencies. BM25 solves this by assigning a weight to every document in the corpus based on their length relative to the average length of documents in the corpus. BM25 accounts for term frequency saturation by capping the influence of terms after an upper-limit is reached. Intuitively, this makes sense as if there are two documents mentioning “covid” 100, and 1,000 respectively, they are both highly relevant to the query, the latter is not 10 times more relevant. BM25 is currently used in Apache’s Lucene search library (The Apache Software Foundation n.d. a) which serves as the backbone of the Solr (The Apache Software Foundation n.d. b) and Elasticsearch (elastic n.d.) search platforms. There have been several modifications to base BM25 since it’s inception. Lv and Zhai (2011) showed that BM25 over-penalised very long document and so proposed a modified version of BM25 called “BM25L” to fix this. It should be noted that a search engine for images indexing on captions, abstracts or titles would never reach sufficient document length for this over-penalisation to occur.

Chapter 3

Implementation

3.1 System Overview

The system operates in a similar way to a standard search engine (see Figure 3.1). Instead of "crawling" the web, images are collected by extracting figures from inside a publication's PDF file. The location of the images are stored in a search index along with metadata associated with the image and parent publication. Items in the search index are sorted using a citation-based impact metric. The user can then enter a search query and will be returned a list of images ranked by a combination query relevance and impact.

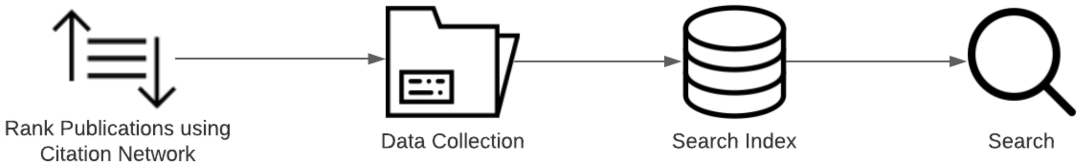
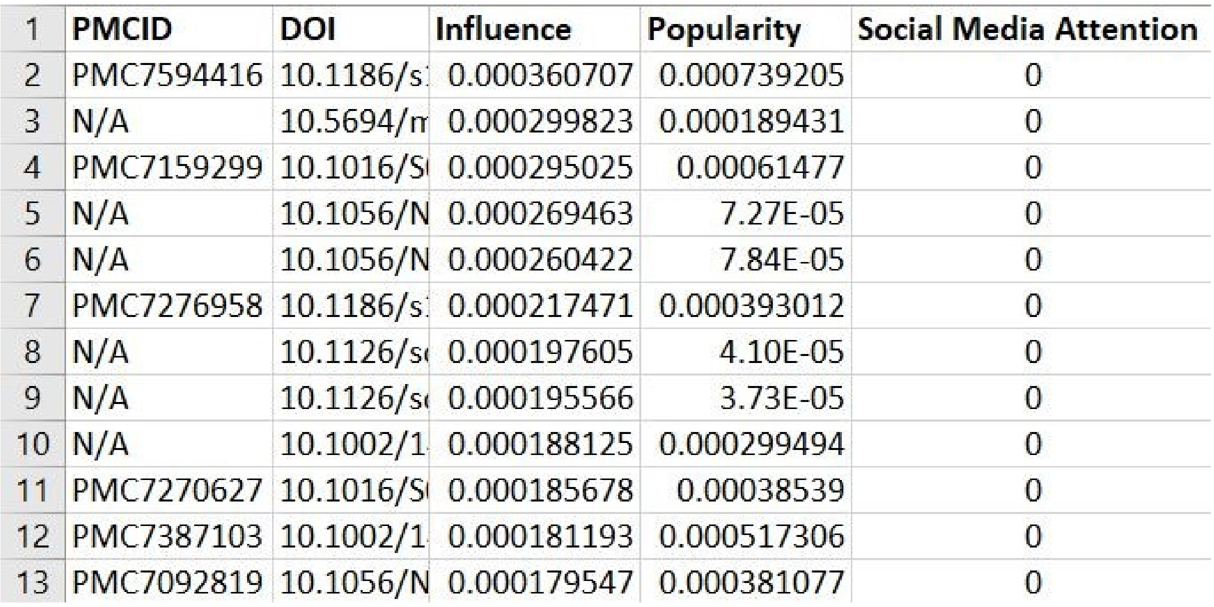


Figure 3.1: High-Level System Architecture

3.2 Ranking COVID-19 related Scientific Publications using the Citation Network

The BIP4COVID19 (Vergoulis et al. 2021) dataset was used to identify high-impact publications and thereby lend credibility to the images contained within them. The PageRank-calculated "influence" metric was used to rank these publications. While any impact metric can be used, given that the BIP4COVID19 dataset is updated regularly, it was better to use a longer-term metric which would be less liable to change during the course of the project, as would have been the case with popularity. Figure 3.2 shows an snippet of the dataset which has been abridged for readability.



1	PMCID	DOI	Influence	Popularity	Social Media Attention
2	PMC7594416	10.1186/s	0.000360707	0.000739205	0
3	N/A	10.5694/m	0.000299823	0.000189431	0
4	PMC7159299	10.1016/S	0.000295025	0.00061477	0
5	N/A	10.1056/N	0.000269463	7.27E-05	0
6	N/A	10.1056/N	0.000260422	7.84E-05	0
7	PMC7276958	10.1186/s	0.000217471	0.000393012	0
8	N/A	10.1126/s	0.000197605	4.10E-05	0
9	N/A	10.1126/s	0.000195566	3.73E-05	0
10	N/A	10.1002/1	0.000188125	0.000299494	0
11	PMC7270627	10.1016/S	0.000185678	0.00038539	0
12	PMC7387103	10.1002/1	0.000181193	0.000517306	0
13	PMC7092819	10.1056/N	0.000179547	0.000381077	0

Figure 3.2: BIP4COVID19 Snippet

3.3 Data Collection

3.3.1 Web Scraping

Originally, web scraping was to be used to retrieve PDF files for the publications in the BIP4COVID19 Dataset. This is well facilitated by the DOI names present in the dataset which can be converted to URLs by simply searching doi.org/+DOI NAME. However, this quickly became unfeasible due to a combination of factors:

1. A server may block the connection if too many requests are made, this would be increasingly likely as the system scaled.

2. Websites do not put download tags in the same location, so finding these for each individual site would require manual inspection of a site's HTML.

3. This method does not provide an easy way to ascertain the licensing information for an individual publication. This is important as this system aims to extract and redistribute images from the publications. This necessitated a change in approach and resulted in a new data collection pipeline.

3.3.2 PubMed Central FTP

PubMed Central (PMC) is a public repository which provides “free full-text archives of biomedical and life sciences journal literature at the U.S. National Institutes of Health’s National Library of Medicine”[1]. 73% of publications in the BIP4COVID19 dataset are available on the PMC Open Access Subset. This subset allows for more liberal redistribution of materials with attribution. In addition to being a reputable resource, the PMC repository solves the earlier problems in three ways:

1. PMC has a File Transfer Protocol(FTP) server[2] which allows developers to easily access medical publications.

2. The Open Access Subset allows for the redistribution of materials. Approximately 97% of PMC’s COVID-19 related papers are included in this subset as part of PubMed’s ‘COVID-19 Initiative’.

3. There are minimal restrictions on downloads. PMC only asks that developers do not make concurrent requests or perform bulk downloads (>100 files) during peak hours.

Due to the above benefits, publications from the PMC Open Access Subset were used to source images.

[1] https://www.ncbi.nlm.nih.gov/pmc/about/intro/

[2] https://www.ncbi.nlm.nih.gov/pmc/tools/ftp/

3.3.3 Data Collection Pipeline

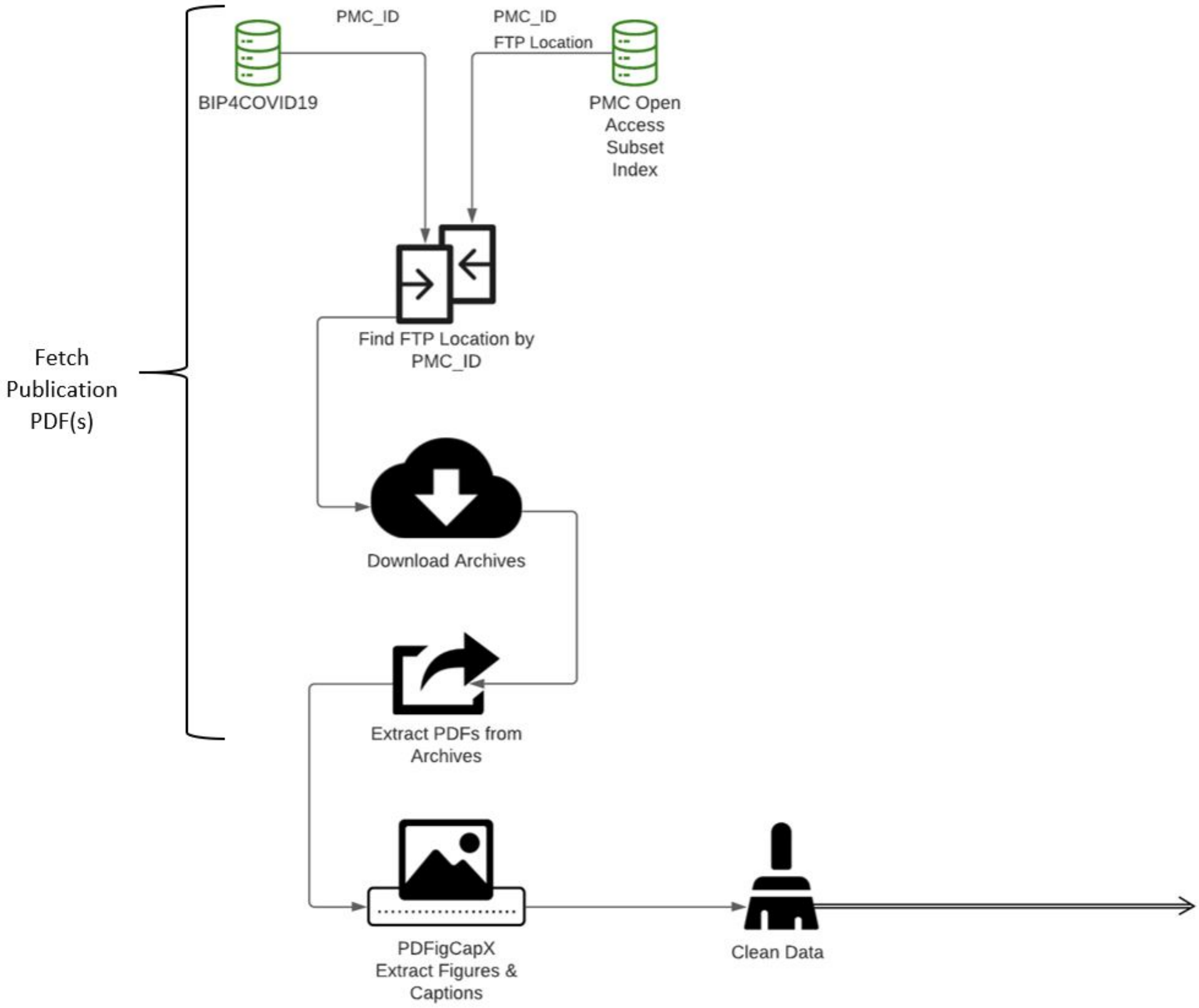

Fetch Publication PDF(s)
PMC_ID
PMC_ID
FTP Location
BIP4COVID19
PMC Open Access Subset Index
Find FTP Location by PMC_ID
Download Archives
Extract PDFs from Archives
PDFigCapX Extract Figures & Captions
Clean Data

Figure 3.3: Data Collection Pipeline Schematic

3.3.4 Tools & Technologies

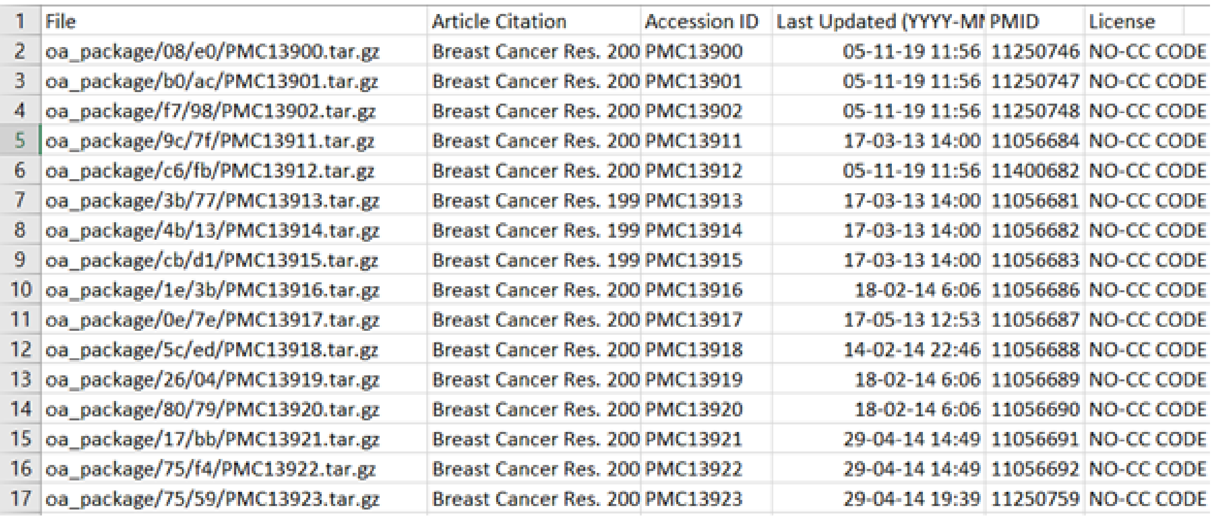
- Datasets
  - BIP4COVID19 (Vergoulis et al. 2021)
  - PMC Open Access Subset Index (National Center for Biotechnology Information 2021)
- Python 3.7
  - Pandas (McKinney et al. 2010) - Reading and Writing CSV Files.

  - Ftplib (Python Internet Protocols n.d.) - FTP Protocol Client Used to Interface with the PMC FTP Server.
  - Tarfile (Python Data Compression and Archiving n.d.) - Extracting Tar Files.
  - PDFigCapx (Li et al. 2019) - Extracts Figures & Captions from PDF Files.
  - Spacy (Honnibal and Montani 2017) - Tokenising Captions and User Queries.
  - Rank-BM25 (Brown 2020) - Determines the Relevancy of a Caption to a Given Query.
  - Tkinter (Lundh 1999) - Creates User Interface.

- Docker (Merkel 2014) - Virtualization Software Used to Run PDFigCapx in an Ubuntu Environment with the Necessary Dependencies.

3.3.5 Fetching Publication PDF Files

The PMC Open Access Subset Index File (Figure 3.4) contains the location of files associated with a given PMC publication on the PMC FTP server. This is searchable using the PMC Identifier (PMCID) of a publication. There are multiple versions of this index file, depending on whether the files will be used for commercial or non-commercial purposes. Given that that this system is non-commercial, we can access the entire Open Access Subset.



1	File	Article Citation	Accession ID	Last Updated (YYYY-MI	PMID	License
2	oa_package/08/e0/PMC13900.tar.gz	Breast Cancer Res. 200	PMC13900	05-11-19 11:56	11250746	NO-CC CODE
3	oa_package/b0/ac/PMC13901.tar.gz	Breast Cancer Res. 200	PMC13901	05-11-19 11:56	11250747	NO-CC CODE
4	oa_package/f7/98/PMC13902.tar.gz	Breast Cancer Res. 200	PMC13902	05-11-19 11:56	11250748	NO-CC CODE
5	oa_package/9c/7f/PMC13911.tar.gz	Breast Cancer Res. 200	PMC13911	17-03-13 14:00	11056684	NO-CC CODE
6	oa_package/c6/fb/PMC13912.tar.gz	Breast Cancer Res. 200	PMC13912	05-11-19 11:56	11400682	NO-CC CODE
7	oa_package/3b/77/PMC13913.tar.gz	Breast Cancer Res. 199	PMC13913	17-03-13 14:00	11056681	NO-CC CODE
8	oa_package/4b/13/PMC13914.tar.gz	Breast Cancer Res. 199	PMC13914	17-03-13 14:00	11056682	NO-CC CODE
9	oa_package/cb/d1/PMC13915.tar.gz	Breast Cancer Res. 199	PMC13915	17-03-13 14:00	11056683	NO-CC CODE
10	oa_package/1e/3b/PMC13916.tar.gz	Breast Cancer Res. 200	PMC13916	18-02-14 6:06	11056686	NO-CC CODE
11	oa_package/0e/7e/PMC13917.tar.gz	Breast Cancer Res. 200	PMC13917	17-05-13 12:53	11056687	NO-CC CODE
12	oa_package/5c/ed/PMC13918.tar.gz	Breast Cancer Res. 200	PMC13918	14-02-14 22:46	11056688	NO-CC CODE
13	oa_package/26/04/PMC13919.tar.gz	Breast Cancer Res. 200	PMC13919	18-02-14 6:06	11056689	NO-CC CODE
14	oa_package/80/79/PMC13920.tar.gz	Breast Cancer Res. 200	PMC13920	18-02-14 6:06	11056690	NO-CC CODE
15	oa_package/17/bb/PMC13921.tar.gz	Breast Cancer Res. 200	PMC13921	29-04-14 14:49	11056691	NO-CC CODE
16	oa_package/75/f4/PMC13922.tar.gz	Breast Cancer Res. 200	PMC13922	29-04-14 14:49	11056692	NO-CC CODE
17	oa_package/75/59/PMC13923.tar.gz	Breast Cancer Res. 200	PMC13923	29-04-14 19:39	11250759	NO-CC CODE

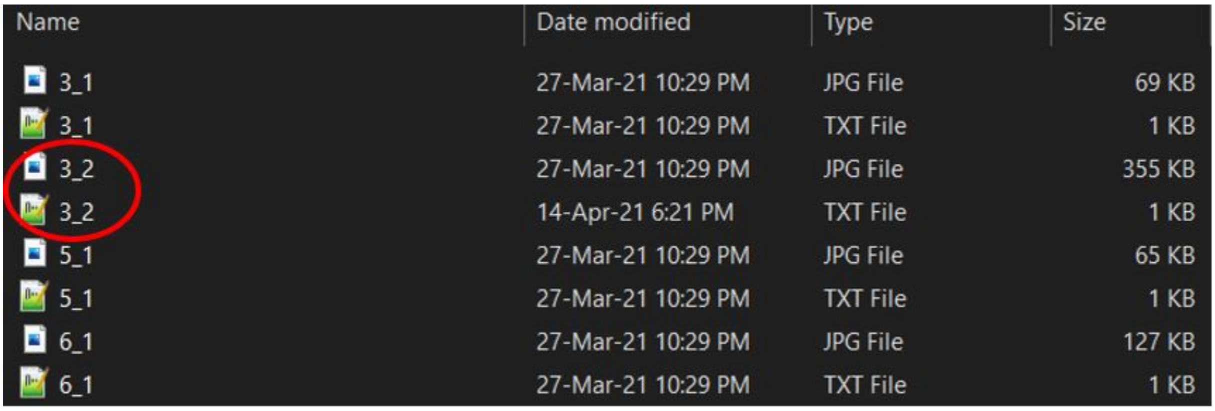
Figure 3.4: PMC Open Access Index Snippet

The FTP location for a BIP4COVID19 entry is found by matching the PMCIDs from BIP4COVID19 with the PMCIDs in the “Accession ID” column in the Open Access Subset Index file. We then use this FTP location to download an archive file relating to that entry from the PMC FTP server. Some archives contain additional PDF files, e.g. supplementary information sheets, in addition to the paper of interest. It is hard to discern between the paper of interest and other PDF files. The papers are usually bigger in size but this is a crude way to differentiate them and could easily be incorrect. Hence, these additional PDFs also pass through the pipeline and so PDFs from the same archive are differentiated with a postfix e.g. PMC13900_1, PMC13900_2, PMC13900_3 and so on. These postfixes are later ignored by the search indexer.

3.3.6 Extracting Figures & Captions

Figures and captions from the PDFs are then extracted using PDFigCapX. This tool segments the textual and graphical contents of a file, applies connected components analysis to identify figures, and then pairs figures with captions using spatial information encoded within the PDF. It was designed specifically for biomedical documents and has been shown to outperform the Allen Institute for AI's alternative "PDFFigures2" system (Clark and Divvala 2016) in the biomedical domain (Li et al. 2019). Both systems perform similarly on figure extraction, but PDFigCapX excels further in extracting captions and pairing them with the appropriate figure, which is vital in this context so that figures can be searched by their caption. Finally, a pass is performed on the pipeline output, and figures which do not have captions are deleted. This removes non-searchable figures.

3.3.7 Example Pipeline Output



Name	Date modified	Type	Size
3_1	27-Mar-21 10:29 PM	JPG File	69 KB
3_1	27-Mar-21 10:29 PM	TXT File	1 KB
3_2	27-Mar-21 10:29 PM	JPG File	355 KB
3_2	14-Apr-21 6:21 PM	TXT File	1 KB
5_1	27-Mar-21 10:29 PM	JPG File	65 KB
5_1	27-Mar-21 10:29 PM	TXT File	1 KB
6_1	27-Mar-21 10:29 PM	JPG File	127 KB
6_1	27-Mar-21 10:29 PM	TXT File	1 KB

Figure 3.5: Example of a Cleaned Output Folder for PMC7112410

3_2.jpg is the Second Figure on Page 3 of the publication and 3_2.txt is the Caption for that Figure. Figure 3.6 shows this figure & the associated caption).

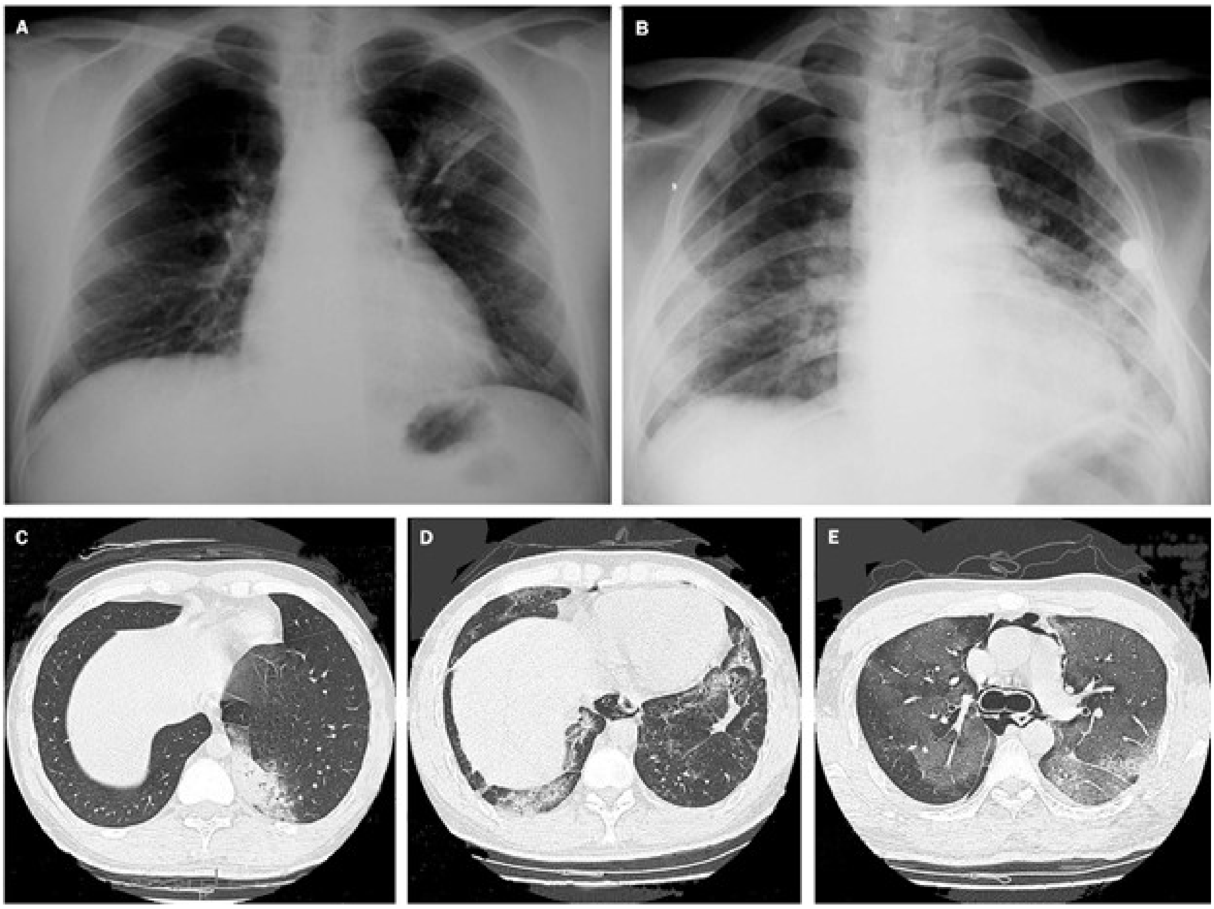
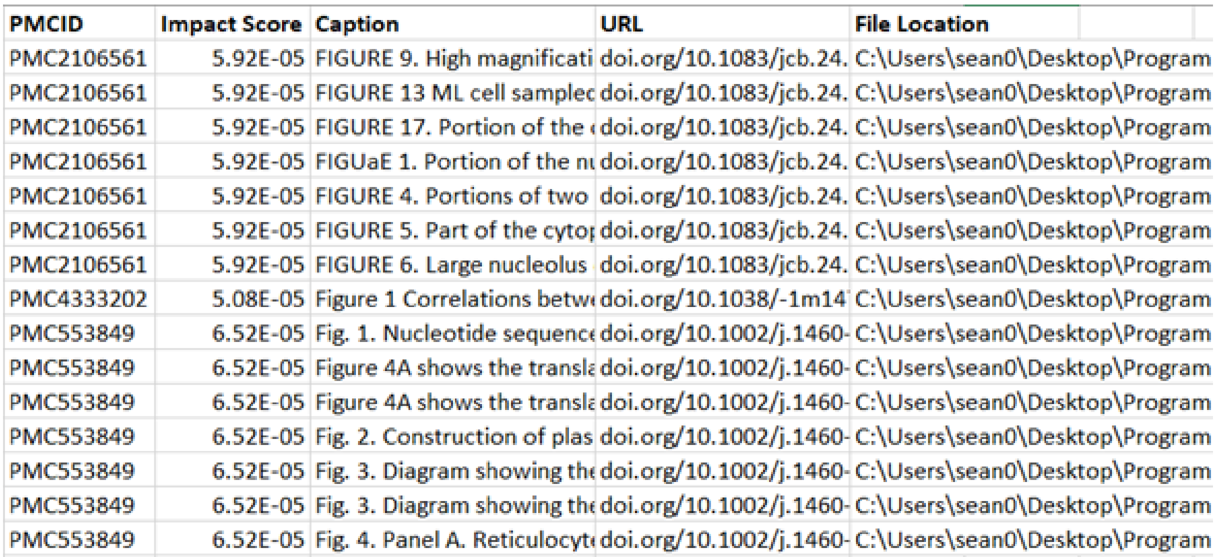


Figure 3.6: Figure & Caption Extraction Example

Caption: “Figure 2: Chest radiographs and high-resolution CT scans from two SARS patients. A Man aged 34 years admitted for high fever and cough. A: Consolidation seen in left upper and middle zones, which progressed maximally at day 7. B: At day 20, resolution of consolidation in the left upper and middle zones but new widespread air-space opacities noted; those in left lung base were confluent. Man aged 32 years, presented with fever, chills, rigors and myalgia, with clear chest radiograph at admission. C: High-resolution CT of thorax shows peripheral subpleural consolidation in medial basal segment of left lower lobe. D: Resolution of original left lower-lobe consolidation at day 18. E:Disease complicated by spontaneous pneumomediastinum.”

3.4 Search Indexing

A search index is then compiled by crawling through the pipeline output folders and, for every image, recording:
PMCID Impact Score Caption URL File Location
Figure 3.7 shows a snippet of the search index.



PMCID	Impact Score	Caption	URL	File Location
PMC2106561	5.92E-05	FIGURE 9. High magnificati	doi.org/10.1083/jcb.24.	C:\Users\sean0\Desktop\Program
PMC2106561	5.92E-05	FIGURE 13 ML cell samplec	doi.org/10.1083/jcb.24.	C:\Users\sean0\Desktop\Program
PMC2106561	5.92E-05	FIGURE 17. Portion of the	doi.org/10.1083/jcb.24.	C:\Users\sean0\Desktop\Program
PMC2106561	5.92E-05	FIGUaE 1. Portion of the nu	doi.org/10.1083/jcb.24.	C:\Users\sean0\Desktop\Program
PMC2106561	5.92E-05	FIGURE 4. Portions of two	doi.org/10.1083/jcb.24.	C:\Users\sean0\Desktop\Program
PMC2106561	5.92E-05	FIGURE 5. Part of the cytop	doi.org/10.1083/jcb.24.	C:\Users\sean0\Desktop\Program
PMC2106561	5.92E-05	FIGURE 6. Large nucleolus	doi.org/10.1083/jcb.24.	C:\Users\sean0\Desktop\Program
PMC4333202	5.08E-05	Figure 1 Correlations betwe	doi.org/10.1038/-1m14	C:\Users\sean0\Desktop\Program
PMC553849	6.52E-05	Fig. 1. Nucleotide sequence	doi.org/10.1002/j.1460-	C:\Users\sean0\Desktop\Program
PMC553849	6.52E-05	Figure 4A shows the transla	doi.org/10.1002/j.1460-	C:\Users\sean0\Desktop\Program
PMC553849	6.52E-05	Figure 4A shows the transla	doi.org/10.1002/j.1460-	C:\Users\sean0\Desktop\Program
PMC553849	6.52E-05	Fig. 2. Construction of plas	doi.org/10.1002/j.1460-	C:\Users\sean0\Desktop\Program
PMC553849	6.52E-05	Fig. 3. Diagram showing the	doi.org/10.1002/j.1460-	C:\Users\sean0\Desktop\Program
PMC553849	6.52E-05	Fig. 3. Diagram showing the	doi.org/10.1002/j.1460-	C:\Users\sean0\Desktop\Program
PMC553849	6.52E-05	Fig. 4. Panel A. Reticulocyte	doi.org/10.1002/j.1460-	C:\Users\sean0\Desktop\Program

Figure 3.7: Search Index Snippet

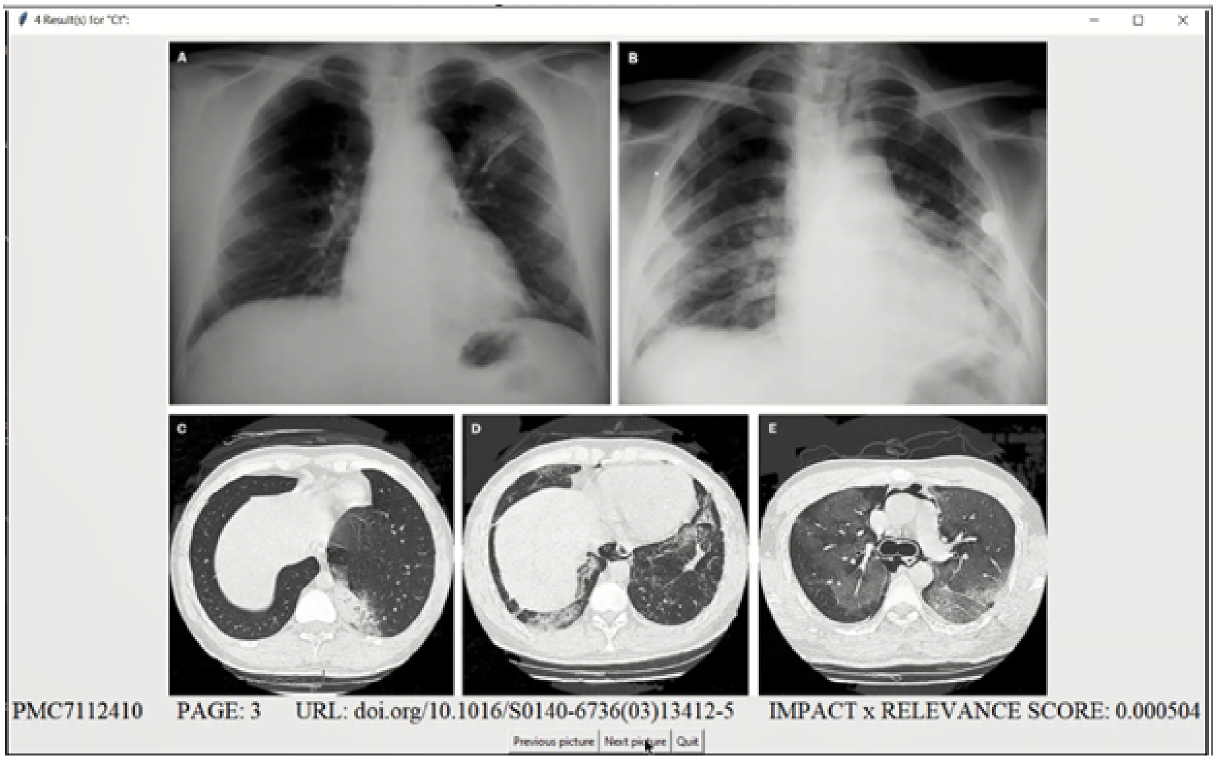
3.5 Searching

To perform a search, the user enters a query into a console. The caption corpus and query are then converted to lower-case and tokenised. BM25 (Robertson et al. 1995) is used to determine the relevancy score of each caption to the user's search query. It was noted in the Chapter 2 that BM25 has been shown to over-penalise very long documents (Lv and Zhai 2011). Given that indexing is done over figure captions, this over-penalisation was not an issue and vanilla BM25 could be used.

For captions with relevancy scores above zero, i.e. at least somewhat relevant, a 5-tuple is created consisting of {PMC_ID, Caption, URL, Ranking Score, Image File Location}. The "Ranking Score" is used to sort the search results and is the relevancy score multiplied by the chosen impact metric. This ensures that both relevancy and impact influence how the search results are sorted. An image from a paper with strong impact and low relevance should not come before an image from a paper with slightly weaker impact and a much higher relevancy, as the latter is a better match to the user's search query.

3.6 User Interface

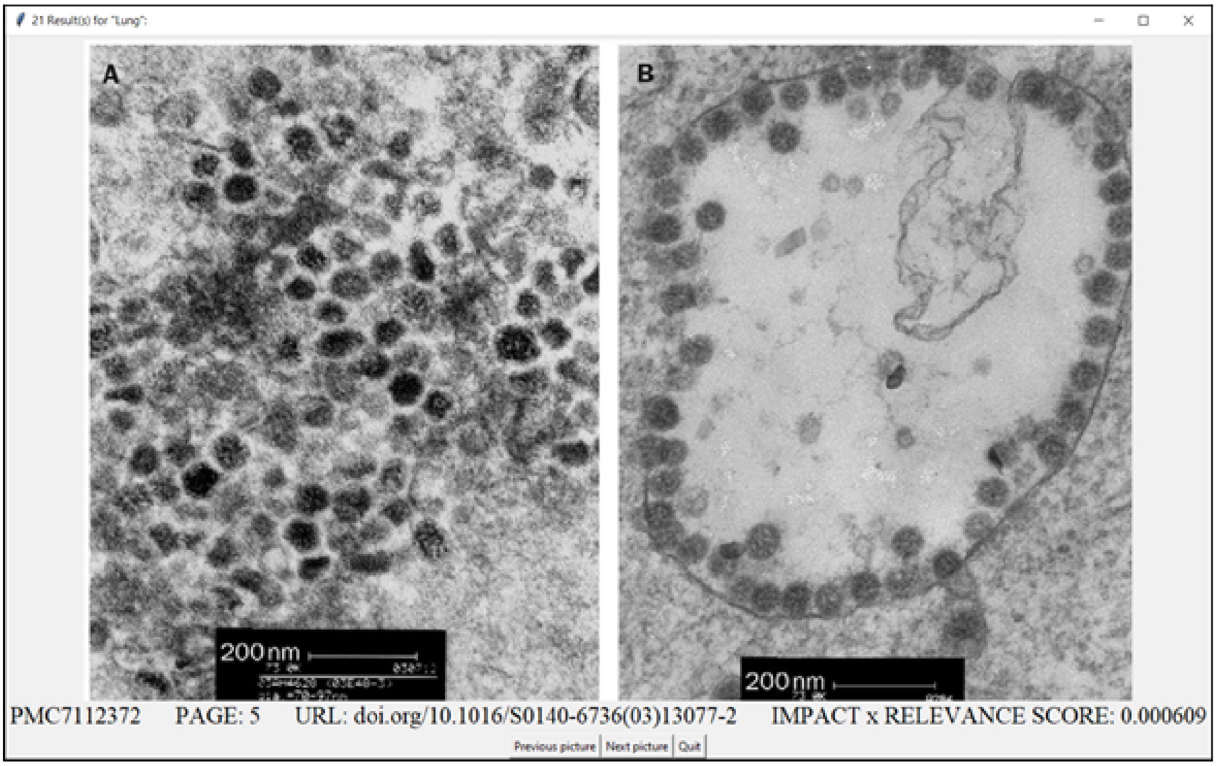
The results are then displayed to the user through a graphical user interface as seen in Figure 3.8 and Figure 3.9. Using the 5-tuple created during the searching step, and the page number inherent in the image file name, we can show the user the PMC_ID, Page Number, URL and Ranking Score for each image. The caption itself is currently printed to the console but ideally this would be shown within the UI itself, but the current UI functions as a proof-of-concept. The number of results is shown in the top left corner and the user can click back and forth through the image gallery.



4 Result(s) for "CT":

PMC7112410 PAGE: 3 URL: doi.org/10.1016/S0140-6736(03)13412-5 IMPACT x RELEVANCE SCORE: 0.000504

Figure 3.8: Search Example "CT"



21 Result(s) for "Lung":

PMC7112372 PAGE: 5 URL: doi.org/10.1016/S0140-6736(03)13077-2 IMPACT x RELEVANCE SCORE: 0.000609

Figure 3.9: Search Example “Lung”

Chapter 4

Evaluation

4.1 Data Collection

As mentioned in Section 3, the “influence” metric from the BIP4COVID19 was the impact metric used to rank publications. The 100 most impactful publications from the PMC Open Access Subset were used as a sample and passed through the Data Collection Pipeline. These 100 publications yielded 293 images in total, with 132 (45%) of these being biological images. These biological images were concentrated in 24 publications. Table 4.1 shows the average number of words and biological images in the entire publication corpus and also only for those publications which contained biological images.

	Words/Pub	**Bio Images/Pub**
All Publications	8532	1.32
Publications with Biological Images	5250	5.5

Table 4.1: Words & Biological Images Per Publication

While the “Words/Pub” appears to show a large difference, there are four outlier publications (see Figure 4.1) which when removed, bring the average number of words down to 5,994, so this difference is insignificant. Perhaps more interestingly, we can see that the biological images are clustered in a relatively small group of publications, with the majority of publications containing no biological images at all (see Figure 4.2). Publications which did have biological images contained an average of 5.5 images. While the old adage is “an image is worth 1,000 words”, the correlation coefficient between the number of words and biological images is insignificant at -0.1011.

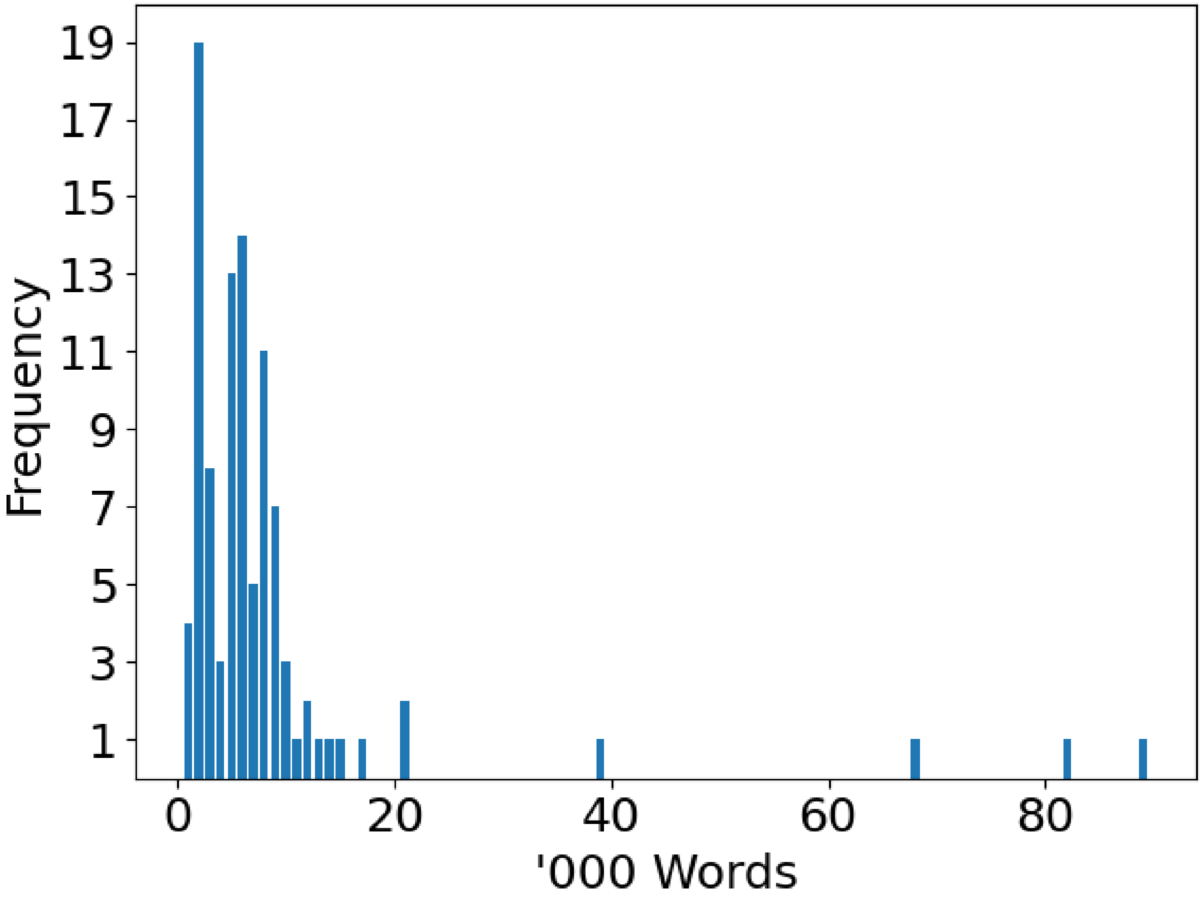


Figure 4.1: Thousand Words per Publication

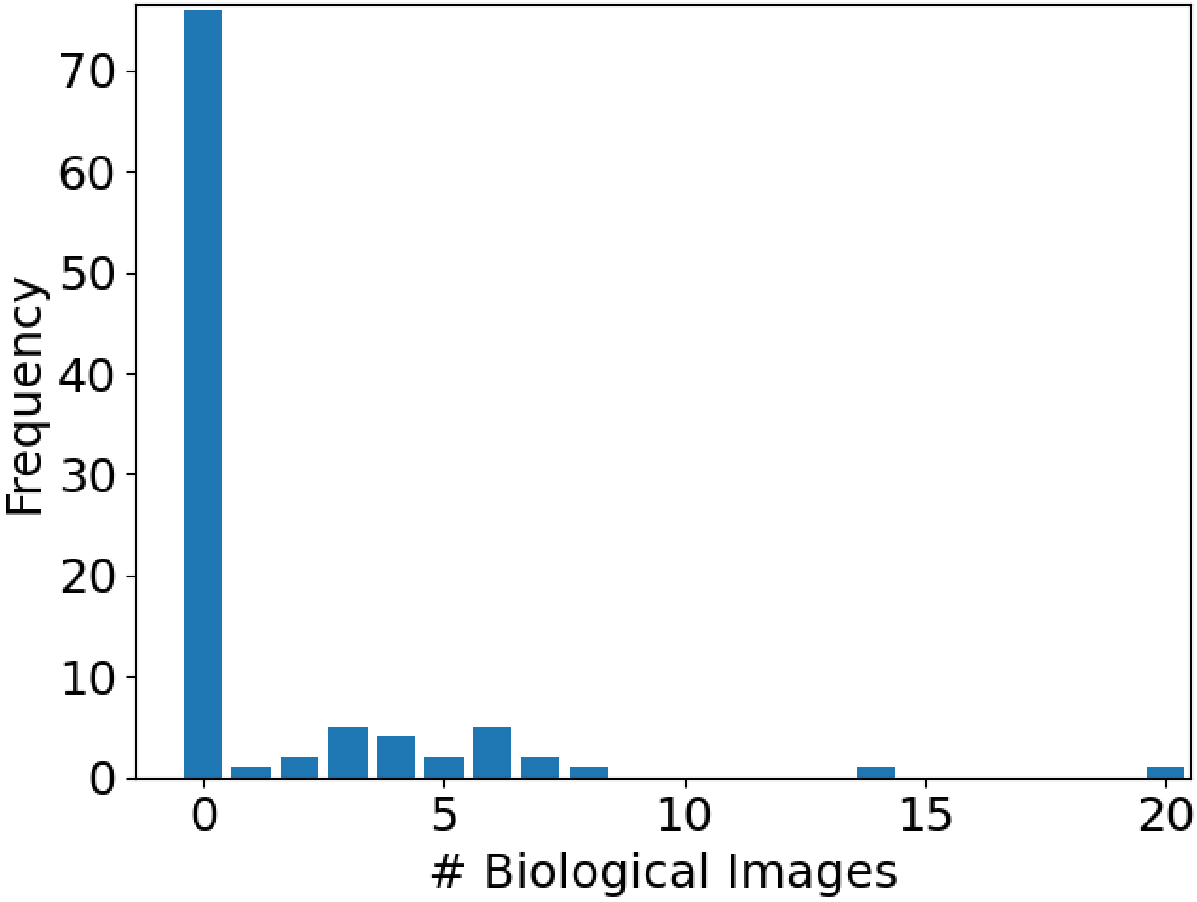


Figure 4.2: Biological Images per Publication

4.2 Image Relevance Index

The images being indexed are currently ranked by an impact factor which reflects the place of the parent publication in the citation network. Here, a new metric called the "Image Relevance Index" is proposed to more directly rank biological images using a combination of common bibliometrics. These metrics are:

- Number of Biological Images in the publication (#Bio).
- h-index (Hirsch 2005) - This metric seeks to quantify the output and impact of an author to their field. It is the number of papers h that an author has published which have received at least h citations.
- Journal Impact Factor (IF) (Garfield 1972) - IF indicates the importance of a journal. It is the average number of citations received by articles in a journal over the previous two years.
- Scimago Journal Rank (SJR) (SCImago 2007) - SRJ is similar to IF in that it measures the importance of a journal by dividing the number of citations received by articles in that journal by the number of the articles in the journal over a given period. The key difference is that SJR applies different weights to each citation, calculated by applying a PageRank-style algorithm to the citation network. It is also measured over a three year period rather than the two year period considered by IF.
- Number of Citations Received by the publication (#citations).

The Image Relevance Index is defined as:

$$\log(\#\mathit{Bio}) \,/\, [1 + \log(\mathit{h\text{-}index}) + \log(IF) + \log(SJR) + \log(\#\mathit{citations})]$$

By combining multiple bibliometrics, a more complete picture can be created versus only looking at the impact of the parent publication. Two of the three expert evaluators in Section 4.3 named journal reputation as a leading factor in judging the credibility of an image. This is incorporated in the Image Relevance Image by using two different journal ranking metrics. Combining these with the author's h-index will boost images which come from historically good sources. Using these longer-term metrics will also help ensure that images from new publications are not overly restricted by their lack of citations if they come from traditionally "good stock".

4.3 Expert System Evaluation

To help evaluate the proposed system, three experts from three different medical fields were enlisted:

1. Dr. Aamir Ahmed: Head of the Prostate Cancer Research Centre at King's College London.

2. Dr. Nollaig Burke: Ussher Assistant Professor in Inflammageing at the Translation Medicine Institute and The Irish Longitudinal Study on Ageing (TILDA) in Trinity College Dublin.

3. Dr. Liam Townsend: Specialist Registrar in Infectious Disease, Trinity College Dublin & St James's Hospital.

These evaluators were asked questions about their use of images (see Appendix A for questionnaire). They were then given a walk-through of the system and asked to rate the usefulness of such a system and provide feedback for future improvements. These evaluations took place over Zoom. See Appendix B, Appendix C and Appendix D for the full answers.

4.3.1 Judging the Credibility of Images

The evaluators were asked how they currently judged the credibility of images that they find in academic publications. Like the proposed system, they use factors surrounding the image to judge credibility. One evaluator checks if the publication that an image appears in is peer-reviewed. Another judges credibility based off the journal which the image appears in and the reputation of that journal. Something akin to a "popularity" ranking was highly desirable to one of the respondents. They noted the need to find images from important new publications which have not yet had the time to accrue a lot of citations. This is something which current medical image retrieval solutions do not offer.

4.3.2 Current Usefulness of the System

The response to the proposed system was generally positive. Each evaluator was asked to rate the usefulness of the system from 1-10. They gave scores of 4,7 and 8 respectively, leading to an average rating of 6.33. Two of the evaluators responded that searching for medical images was highly important for their work and/or research. They currently search for images by reading through papers on services such as PubMed and noted that this can be very time consuming. They appreciated the convenience of the proposed system as they would not have to manually scour journals to find images. One evaluator also liked the fact that the solution supports multiple languages. Many journals are single-language and so finding non-english publications requires someone to look through non-english journals or use a repository such as PubMed. Given that the proposed system uses PubMed exclusively, this is a not an issue.

4.3.3 Recommended Improvements

The evaluators were asked to suggest improvements to the system. These mainly focused on the user-side experience, which is understandable given that the demonstration was a user-side view of the system. Those recommendations will be discussed here, while back-end changes suggested by the author will be discussed in Chapter 5.

The first recommended improvement was to widen the search scope. Currently, searches are performed by searching a corpus of figure captions. It was suggested to widen this to include the abstract and possibly other textual elements such as the publication title. While it was noted that caption search is useful, it does rely on authors labelling their figures with enough detail to facilitate good searching. From manual inspection of the current corpus, the labelling is generally of a very high standard, but less descript captions are still present. It is also important to note that the high-quality labelling of these figures could be somewhat expected given that the 100 highest impact publications were used to generate the figure and caption corpuses. One of the evaluators noted that lesser quality publications generally have lesser quality captions. Widening the search scope by including titles and abstracts is well factilitated by the current data collection pipeline. Every archive file from which publication PDF files are extracted also includes an XML file which contains the full text of a publication. From this we can parse the title and abstract to widen the search scope. This additional information could also be added to the front-end, as supplementing figure captions with titles and abstracts has been shown to significantly increase reader comprehension of figures (Yu et al. 2009).

There were varying opinions on the inclusion of data figures/non-medical images in the search results. Currently, data figures are being returned in the search results. One evaluator found this useful as interesting data figures could point them towards noteworthy publications, while another felt that data figures did not mean much outside of the full-text publications, where a wider explanation of the data would be present. This latter evaluator suggested clustering data figures with the corresponding medical image as this would reveal more about the total context of the image. The example use-case they gave was grouping an ROC curve showing the performance of a machine learning model with an image that was used to test the performance of that model. As data figures and medical images are very different in appearance, an image classification system could effectively label the images and allow users to filter between data figures and medical imagery, this would strike a balance between the two diverging evaluator opinions.

One smaller suggestion was on the form of the system itself. It was suggested that in addition to being a fully-fledged image retrieval system operating online like a traditional search engine, the program could also act as a browser plugin. This plugin could detect when the user was viewing a specific publication on PubMed and offer to show the user the figures from that publication. This would provide the user with a quicker and more convenient way to view and save images of publications. In this scenario the user has already chosen a publication to inspect. This eliminates the need for a system to rank the images, although this could still be included to provide useful metrics. This suggestion was made late into the course of this project so a thorough consideration of this has not been made, although it is certainly technically feasible.

Chapter 5

Conclusions & Future Work

5.1 Conclusions

This report has outlined a system which allows users to search for images from scientific publications. The images in this case are related to COVID-19 and other historical coronaviruses but this could be adapted to any rapidly changing situation where important information must be filtered from an over-abundance of new information. This system is enabled by a data collection pipeline which fetches scientific publications and extracts their figures and captions. The images are indexed and made searchable by their captions. As discussed in Chapter 4, there is room to widen this search scope and a clear means to do so. Images are ranked by an impact factor which identifies the importance of the parent publication by its place in the underlying citation network. A new metric called the Image Relevance Index was proposed to more directly rank biological images using a combination of commonly available bibliometrics. Three experts were recruited to help evaluate the system and the feedback was generally positive, especially from an evaluator who specialised in infectious disease like COVID-19 which is encouraging. The changes suggested by these evaluators were discussed in-depth in Chapter 4. The search tool currently runs locally through a GUI as a proof of concept with the ultimate goal being to have make it available online for public use.

5.2 Future Work

There were a number of improvements suggested by the evaluators in Section 4. These focused on the user-side experience. There are a number of back-end improvements which could also be made to further enhance the current system. These mainly pertain to the Data Collection Pipeline.

Perhaps the biggest improvement to the Data Collection Pipeline would be performance. While this is not a problem currently, it is important to consider these improvements to allow the system to scale well. The main bottleneck here is the figure and caption extraction using PDFigCapx. Docker is used to run PDFigCapx in a virtualised Ubuntu environment along with the necessary dependencies. The PDF files are read into the virtualised environment from the Windows 10 host machine. This generates a warning from Docker that the file sharing "may perform poorly". Originally, the plan was to run PDFigCapx on the host OS but even after installing the listed dependencies, the program would not run correctly. After much time debugging, Docker was used as a fallback. Ideally, everything could be run on the host OS or could be ported and ran virtually in Docker to eliminate the file sharing slowdown. Secondly, and more significantly, the extraction is currently single-threaded. From experience, extracting figures and captions takes between 30-90 seconds depending on PDF size and the number of figures. While this was allowable for 100 publications and was no doubt in part caused by the file sharing issue, this would not be practical for 1,000 or 100,000 publications. The independent nature of the inputs makes them well suited to multi-threading. While not needed in this instance, this would be the biggest change needed for a more scalable system.

Fetching the publications PDF files could also take advantage of multi-threading to improve pipeline performance. Downloading the publication archive files is limited by PubMed's policy of "no concurrent downloads" but the downloaded archive files themselves could be divided into batches and extracted by multiple threads concurrently. Again, this was a non-issue in this instance and of lesser concert than the previous point but could be worthwhile if the system had to scale dramatically for real-world use.

It is also important to ensure that publications are not entering into the pipeline more than once when the pipeline is re-run to collect new images. This could be done quite easily by searching for the PMC Identifier of a potential publication in the most-recent search index to ensure that it has not yet passed through the data collection pipeline.

The Image Relevance Index arose during the course of this project as a potential way of better ranking biological images in scientific publications. Further work could be done on this by applying it to a large sample of publications which contain biological images and comparing it with other bibliometrics such as the ranking methods used in the BIP4COVID19 dataset. This could also be further explored by again consulting with people in the fields of academia and medicine. Hopefully the eventual subsidence of the COVID-19 pandemic will make this more feasible.

5.3 Final Remarks

On a personal level, I believe the project and the entire process behind it has been very beneficial and rewarding. It is the first solo software and research project that I have undertaken. There have been very few times in my formal education to date where there have effectively been no constraints to a problem and I have complete freedom of approach. The longer-term nature of the project also requires a higher level of self-motivation and discipline to stay on target. While the sheer number of possibilities and unknowns were daunting at first, gradually designing and later building a solution and then seeing the positive reaction that it got from potential users in academia and medicine was very rewarding.

I was also exposed to many new tools and technologies. My only previous experience with Python was for machine learning which was very domain-specific. Building this entire system in Python has given me a much more rounded skill set. I also gained valuable experience with Docker, which seems to be the industry standard for virtualisation. The project also revealed the amazing collaboration that exists within the scientific and academic communities through things like the PMC Open Access Subset and publicly available state of the art tools like PDFigCapx. Naturally, I also became very familiar with the (very topical!) COVID-19 literature, and academia more generally, which may be useful for my future.

Bibliography

Allen Institute for AI (2020). Spike: Search over covid-19. URL: https://spike.covid-19.apps.allenai.org/; accessed 16-April-2021.

Bhatia, P., Liu, L., Arumae, K., Pourdamghani, N., Deshpande, S., Snively, B., Mona, M., Wise, C., Price, G., Ramaswamy, S., Ma, X., Nallapati, R., Huang, Z., Xiang, B., and Kass-Hout, T. (2020). Aws cord-19 search: A neural search engine for covid-19 literature.

Brown, D. (2020). Rank-BM25: A Collection of BM25 Algorithms in Python. URL: https://doi.org/10.5281/zenodo.4520057; accessed 17-April-2021.

Chen, P., Xie, H., Maslov, S., and Redner, S. (2007). Finding scientific gems with google's pagerank algorithm. *Journal of Informetrics*, 1(1):8–15.

Chen, Q., Allot, A., and Lu, Z. (2020). Keep up with the latest coronavirus research. *Nature*, 579(7798):193–193.

Clark, C. and Divvala, S. (2016). Pdffigures 2.0: Mining figures from research papers.

Cohen, J. P., Morrison, P., Dao, L., Roth, K., Duong, T. Q., and Ghassemi, M. (2020). Covid-19 image data collection: Prospective predictions are the future. *arXiv 2006.11988*.

Demner-Fushman, D., Antani, S., Simpson, M., and Thoma, G. (2012). Design and development of a multimodal biomedical information retrieval system. *Journal of Computing Science and Engineering*, 6:168–177.

elastic (n.d.). Elasticsearch. URL: https://www.elastic.co/elasticsearch/; accessed 25-April-2021.

European Institute for Biomedical Imaging Research (2021). Covid-19 imaging datasets. URL: https://www.eibir.org/covid-19-imaging-datasets/; accessed 15-April-2021.

Frederickson, M. (2020). Covid-19's gendered impact on academic productivity. URL: `https://github.com/drfreder/pandemic-pub-bias`; accessed 20-April-2021.

Garfield, E. (1972). Citation analysis as a tool in journal evaluation. *Science*, 178(4060):471–479.

Ghosh, R., Kuo, T.-T., Hsu, C.-N., Lin, S.-D., and Lerman, K. (2011). Time-aware ranking in dynamic citation networks. In *2011 IEEE 11th International Conference on Data Mining Workshops*, pages 373–380.

Google (2020). Covid-19 research explorer. URL: `https://covid19-research-explorer.appspot.com/`; accessed 16-April-2021.

Hearst, M., Divoli, A., Guturu, H., Ksikes, A., Nakov, P., Wooldridge, M., and Ye, J. (2007). Biotext search engine: Beyond abstract search. *Bioinformatics (Oxford, England)*, 23:2196–2197.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, 102(46):16569–16572.

Honnibal, M. and Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Kanellos, I., Vergoulis, T., Sacharidis, D., Dalamagas, T., and Vassiliou, Y. (2019). Impact-based ranking of scientific publications: A survey and experimental evaluation. *IEEE Transactions on Knowledge and Data Engineering*.

Krapivin, M. and Marchese, M. (2008). Focused page rank in scientific papers ranking. In Buchanan, G., Masoodian, M., and Cunningham, S. J., editors, *Digital Libraries: Universal and Ubiquitous Access to Information*, pages 144–153, Berlin, Heidelberg. Springer Berlin Heidelberg.

Li, P., Jiang, X., and Shatkay, H. (2019). Figure and caption extraction from biomedical documents. *Bioinformatics (Oxford, England)*, 35.

Lundh, F. (1999). An introduction to tkinter. *URL: www.pythonware.com/library/tkinter/introduction/index.html.*

Lv, Y. and Zhai, C. (2011). When documents are very long, bm25 fails! SIGIR '11, page 1103–1104, New York, NY, USA. Association for Computing Machinery.

Ma, N., Guan, J., and Zhao, Y. (2008). Bringing pagerank to the citation analysis. *Information Processing Management*, 44(2):800–810. Evaluating Exploratory Search Systems Digital Libraries in the Context of Users' Broader Activities.

McKinney, W. et al. (2010). Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, volume 445, pages 51–56. Austin, TX.

Merkel, D. (2014). Docker: lightweight linux containers for consistent development and deployment. *Linux journal*, 2014(239):2.

National Center for Biotechnology Information (n.d.). FTP Service. URL: `https://www.ncbi.nlm.nih.gov/pmc/tools/ftp/`; accessed 15-April-2021.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab. Previous number = SIDL-WP-1999-0120.

Python Data Compression and Archiving (n.d.). tarfile — Read and write tar archive files. URL: `https://docs.python.org/3/library/tarfile.html`; accessed 17-April-2021.

Python Internet Protocols (n.d.). ftplib - FTP protocol client. URL: `https://docs.python.org/3/library/ftplib.html`; accessed 17-April-2021.

Region, B. M. I. D. o. t. V., Pertusa, A., and de la Iglesia Vaya, M. (2020). Bimcv-covid19+. URL: `osf.io/nh7g8`; accessed 15-April-2021.

Robertson, S., Walker, S., Jones, S., Hancock-Beaulieu, M. M., and Gatford, M. (1995). Okapi at trec-3. In *Overview of the Third Text REtrieval Conference (TREC-3)*, pages 109–126. Gaithersburg, MD: NIST.

Roozenbeek, J., Schneider, C., Dryhurst, S., Kerr, J., Freeman, A., Recchia, G., van der Bles, A. M., and van der Linden, S. (2020). Susceptibility to misinformation about covid-19 around the world. *Royal Society Open Science*, 7.

Sarewitz, D. (2016). The pressure to publish pushes down quality. *Nature*, 533:147–147.

SCImago, G. (2007). Sjr—scimago journal & country rank. *Consejo Superior de Investigaciones Cientificas (CSIC), University of Granada, Extremadura, Carlos III (Madrid) and Alcalá de Henares.*

Springer (n.d.). Figures and tables. URL: https://www.springer.com/gp/authors-editors/authorandreviewertutorials/writing-a-journal-manuscript/figures-and-tables/10285530; accessed 23-April-2021.

Stanford University Center for Artificial Intelligence in Medicine & Imaging (2021). Covid-19 + imaging ai resources. URL: https://aimi.stanford.edu/resources/covid19#models; accessed 15-April-2021.

The Apache Software Foundation (n.d.a). Apache lucene. URL: https://lucene.apache.org/; accessed 25-April-2021.

The Apache Software Foundation (n.d.b). Apache solr. URL: https://solr.apache.org/; accessed 25-April-2021.

The Cancer Imaging Archive (2021). Covid-19 datasets on tcia. URL: https://wiki.cancerimagingarchive.net/display/Public/COVID-19; accessed 15-April-2021.

Trewartha, A., Dagdelen, J., Huo, H., Cruse, K., Wang, Z., He, T., Subramanian, A., Fei, Y., Justus, B., Persson, K., and Ceder, G. (2020). Covidscholar: An automated covid-19 research aggregation and analysis platform.

UNESCO Institute for Statistics (2019). Women in science 2019. URL: http://uis.unesco.org/sites/default/files/documents/fs55-women-in-science-2019-en.pdf; accessed 19-April-2021.

Vergoulis, T., Kanellos, I., Chatzopoulos, S., Karidi, D. P., and Dalamagas, T. (2021). BIP4COVID19: Impact metrics and indicators for coronavirus related publications.

Walker, D., Xie, H., Yan, K.-K., and Maslov, S. (2007). Ranking scientific publications using a model of network traffic. *Journal of Statistical Mechanics: Theory and Experiment*, 2007(06):P06010–P06010.

Wang, L., Lin, Z. Q., and Wong, A. (2020a). Covid-net: a tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images. *Scientific Reports*, 10(1):19549.

Wang, L. L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Eide, D., Funk, K., Kinney, R. M., Liu, Z., Merrill, W., Mooney, P., Murdick, D., Rishi, D., Sheehan, J., Shen, Z., Stilson, B. B. S., Wade, A. D., Wang, K., Wilhelm, C., Xie, B., Raymond, D. A., Weld, D. S., Etzioni, O., and Kohlmeier, S. (2020b). Cord-19: The covid-19 open research dataset. *ArXiv*.

World Health Organisation (2020a). Cross-Regional Statement on “Infodemic” in the Context of COVID-19. URL: `https://onu.delegfrance.org/IMG/pdf/cross-regional_statement_on_infodemic_final_with_all_endorsements.pdf`; accessed 18-April-2021.

World Health Organisation (2020b). Infodemic. URL: `https://www.who.int/health-topics/infodemic`; accessed 18-April-2021.

Xu, S., McCusker, J., and Krauthammer, M. (2008). Yale image finder (yif): a new search engine for retrieving biomedical images. *Bioinformatics (Oxford, England)*, 24(17):1968–1970.

Yu, H., Agarwal, S., Johnston, M., and Cohen, A. (2009). Are figure legends sufficient? evaluating the contribution of associated text to biomedical figure comprehension. *Journal of biomedical discovery and collaboration*, 4:1.

Yu, P., Li, X., and Liu, B. (2005). Adding the temporal dimension to search - a case study in publication search. In *The 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI’05)*, pages 543–549.

Appendix

A Evaluation Questionnaire

1. Do you use diagnostic images in your work/research?

2. On a scale of 1-10, how important is looking up diagnostic images for your work/research?

3. How do you currently search for diagnostic images?

4. How do you currently judge the credibility of an image?

———Demonstrate Solution———

5. On a scale of 1-10, how would you rate the usefulness of this system?

6. What changes/improvements to the system would you recommend?

B Dr. Aamir Ahmed Questionnaire Responses

1. Do you use diagnostic images in your work/research?

Yes

2. On a scale of 1-10, how important is looking up diagnostic images for your work/research?

8

3. How do you currently search for diagnostic images?

Medical repositories e.g. PubMed, hospital or archival sources.

4. How do you currently judge the credibility of an image?

Based on whether they appear in peer-reviewed publications.

———Demonstrate Solution———

5. On a scale of 1-10, how would you rate the usefulness of this system?

4

Good that it reduces the need to scour journals and papers to find images.

Caption search is useful

6. What changes/improvements to the system would you recommend?

No figures in the search results would be better.

Instead, connect medical images to the corresponding data figures.

C Dr. Nollaig Bourke Questionnaire Responses

1. Do you use diagnostic images in your work/research?
No

2. On a scale of 1-10, how important is looking up diagnostic images for your work/research?
N/A

3. How do you currently search for diagnostic images?
N/A

4. How do you currently judge the credibility of an image?
N/A

———Demonstrate Solution———

5. On a scale of 1-10, how would you rate the usefulness of this system?
7

Could be a useful tool to wade through popular recent publications to find important ones.

6. What changes/improvements to the system would you recommend?
Searching by captions makes you very dependent on the author to label and describe figures in an adequate way.

Use the title, abstract or full-text elements to expand the search scope.

D Dr. Liam Townsend Questionnaire Responses

1. Do you use diagnostic images in your work/research?
Yes

2. On a scale of 1-10, how important is looking up diagnostic images for your work/research?
8

3. How do you currently search for diagnostic images?
Searching keywords in PubMed or Google Scholar.
Google Images sometimes, but this won't return as 'scientific' results.

4. How do you currently judge the credibility of an image?
What journal is the image in?
What is the reputation of that journals?
What is the author attempting to show?
Is this being represented in the images?

———Demonstrate Solution———

5. On a scale of 1-10, how would you rate the usefulness of this system?
8

6. What changes/improvements to the system would you recommend?
No glaring omissions, actually using the system would probably reveal potential areas of improvement.
Could include abstracts for additional search scope.
Bad figure captions in bad journals.
Journals are usually single language, often limited to English, this system is language independent which is good.
A plugin could be useful, e.g. the plugin detects what PubMed article I am currently reading and offers the images for viewing, rather than having to go through the paper first.