Assignment 2 of Big Data Analysis and Project
COMP SCI 7209

COVID-19 Open Research Dataset Challenge (CORD-19)

Man Him Li, a1817190

The University of Adelaide, manhim.li@student.adelaide.edu.au

Abstract:

The primary objective of this project was going to use natural language processing techniques to build a question and answering tool for answering COVID-19 queries. The dataset for this task was publicly available on Kaggle and it contains more than 500,000 scholarly articles in the medical field. In this project, exploratory data analysis was the first section to explore statistical insights of the given dataset. The result from the data analysis benefited this project from filtering unnecessary scholarly articles for finding relevant answers in the question tool. The criteria of articles selection were based on the occurrence of some common COVID-19 related keywords in their titles and abstracts and only articles published in 2021 were considered. This project was trying to build a question and answering tool to deliver 10 relevant COVID-19 related answers to the given questions input by individuals who are queries about the pandemic. The Word2Vec algorithm was utilised for the word embedding process after data selection and preprocessing. The initial word embedding model was then tested with a simple vocabulary to find synonyms from the given body contexts in the selected scholar articles prior to the final delivery of the question and answering tool. The final model effectively generated 10 most relevant answers from the body contexts of the selected scholarly articles based on the score of the cosine similarity to the given question input related to COVID-19.

1. INTRODUCTION

With regards to the occurrence of the COVID-19 pandemic, this project was going to set up a Question and Answering Tool for individuals who had any queries related to the COVID-19. This project was originally established in Kaggle as a COVID-19 open research dataset challenge to ask for participated programmers to utilise AI techniques to develop a Question and Answer tool to deliver high-quality scientific answers in response to the COVID-19 pandemic based on the text from scholarly articles. The given dataset contained more than 500,000 scholarly articles with 200,000 full body texts related to coronaviruses. However, the primary objective of this project was to generate a question and answering tool related to COVID-19, the only occurrence of 'COVID-19' in the titles and abstracts of scholarly articles were selected as legitimate dataset for natural language processing in this project. All unnecessary scholarly articles were removed and some statistical insights of the selected data were generated in the section of exploratory data analysis with the matplotlib functions in Python. However, this project would eventually take 30,000 relevant scholarly articles for model training due to the limitations of my computer performance. The next section was data preprocessing including null values removal, tokenisation, lowercase conversion and punctuation removal to make the selected data ready for word embedding implementation. This project utilised the Word2vec algorithm to convert input words into a set of mathematical vectors representation. Word2Vec algorithm had the capability to assign similar words with similar meanings into a similar mathematical vectors representation. The selected 30,000 articles were then passed to the Word2Vec model for word embedding processing. A cosine similarity function imported from the Word2Vec library was used to compare the sentence

similarity between the selected body contexts and the input question given by the users. The ten sentences with the highest score of cosine similarity were selected and displayed as a result in the Question and Answering tool.

2. DATA COLLECTION

This project was going to utilise the dataset given by Kaggle. First of all, the metadata spreadsheet summarised the articles in terms of their titles, authors, abstracts, publish dates and paper_id,.etc. The sentences from the abstracts of different articles were extracted directly from the spreadsheet while the body contexts of the selected articles were extracted from the local file based on the given JSON file names retrieved from the paper_id's column of the spreadsheet.

3. METHODOLOGY

This project was divided into several parts, including exploratory data analysis, data pre-processing, Word2Vec model word embedding processing, testing model efficiency by using one single vocabulary and finally delivering the top 10 answers based on the cosine similarity scores with the questions input by users. Each part will be further explained in the next session of this report.

3.1 Exploratory Data Analysis

Prior to the application of model training, exploratory data analysis was the essential first step to help researchers better understand the data, such as detecting missing data and recognising data nature and distribution. Graphs and tables were used in this session to visualise the statistical information of the dataset. The statistical information benefited this project from mining appropriate datasets which deliver better results for the tool of question and answers related to COVID-19.

3.2 Data Preprocessing

Prior to the implementation of the Word2Vec technique to learn word embedding, selected data of body contexts was needed to be preprocessed. This project is going to utilise the preprocessing tools from an open-source library called Gensim. The preprocessing procedure included null values removal, tokenization, lowercasing and punctuations removal.

- Removing null values
  Null values in the data frame were removed in this section based on the information given from the previous section.

- Tokenisation
  It was the first step to preprocess a piece of text data prior to model training. Text data was separated into one piece of a unit which was known as a token. In general, tokenisation could be either converting information into a word, a character or a group of specified characters. However, in this project, all words in a sentence were tokenised into pieces of a single word as a token.

- Lowercasing
  Lowering cases of all text data was the simplest data pre-processing for the application of natural language processing. It was important as it helped resolve the sparsity issue within the input text. For instance, the word 'Covid' and COVID' were seen as the same word after lowercasing.

- Removing punctuations
  It was also considered as noise removal as some of the unnecessary characters existed in the dataset. The data became clean and consistent without the existence of punctuations.

- Stemming/ Lemmatization/ Stopwords removal

  Some other common texts preprocessing such as stemming and lemmatization and removing stopwords were not used in this project. Stemming and lemmatization both converted words to their basic root form. For instance, the word 'going' and gone' would be transformed into 'go' as one word. However, this process was ineffective in that the transformed word had altered its original meaning due to its different grammatical structure. Moreover, stop word removal was another common text preprocessing to reduce the number of features prior to model training. For example, some words such as 'a' and 'the' were removed from the dataset and only important words were left for model training. However, this project is going to apply the Word2Vec model and the dependency information of one word was calculated based on the previous and next words. Therefore, remaining all the stop words in a sentence without lemmatization for the Word2Vec model could be a better option to provide a high-quality answer for the input question.

3.3 Word2Vec model word embedding process

A word embedding model was initialized in this section from Gensim which was an open-source library for natural language processing. A Word2Vec algorithm was selected to learn and train word embedding from the input tokenized text corpus. Word2Vec was an unsupervised learning algorithm to create a vector representation of a single word in the vector space. The general concept of the Word2Vec algorithm was to create a sliding window of each target word in a sentence and to learn the context between the neighbour words and the target word. The window size could be controlled by the users in the parameter setting. Word2Vec algorithm contained two different architectures which were The Continuous Bag of Words and Skip-grams. This project was going to utilise CBOW as its training time was faster than skip-gram and it performed better for predicting some repeated words as our project was mainly focusing on the COVID-19 spectrum (Riva, 2021). The working concept behind CBOW was firstly selecting the centre word as a target based on the window size. Then the neighbour words were used as features to predict the middle target word. Building an effective Word2Vec model needed customised parameters setting which include size, window, minimum count, workers, and sg.

- Size

  It was a parameter to control vector dimension of each input tokenized word. Default setting was 100 which were sufficient for this project.
- Window

  The numbers of neighbors words to have direct impact for each target word.
- Min_count

  This setting referred to control the minimum number of words to be presented in each iteration of model training. A text corpus less than the threshold would not be processed.
- Workers

  This parameter setting controls number of threads that the users desired to use to allow the number of cores in the computer CPU to run on this model training.
- sg

  This parameter selected either CBOW algorithm or skip gram algorithm for the Word2Vec model.

3.4 Cosine Similarity comparison

It was a common methodology to compare similarities between two different words or sentences in a vector space. The concept behind it was to calculate the cosine angle difference between two vectors in the same dimension. For sentence comparison, an average vector would be calculated from a list of individual vectors representing different words in one sentence. Two average vectors representing two different sentences were compared based on the cosine similarity function from the Word2Vec application. In our project, the question input by a user was treated as a sentence and converted to its

average vectors and compared to all sentences selected from the COVID-19 related scholar articles in terms of cosine similarity. The higher the similarity score obtained, the more similar two sentences are related to each other and the selected sentences were high likely to match the information needed for the given questions.

4 EXPERIMENTS AND FINDINGS

4.1 Exploratory data analysis and Data Selection

Exploratory data analysis and data selection were combined into one section to select relevant data based on the statistical insight. Since this project was going to create a Question and Answering tool for the COVID-19 query, the selected scholarly articles should be directly related to COVID-19. One way of selecting relevant articles was to check whether there was keyword “COVID-19” presented in the title and abstract of the scholarly articles. Only scholarly articles with titles and abstracts having the keyword “COVID-19” were selected. Figure 1 illustrated the occurrence number of “COVID-19” in the column of the title whilst Figure 2 illustrated the occurrence of the keyword “COVID-19” in both columns of title and abstract. There was a total of 55,624 articles containing the keyword “COVID-19” in the column of the title whilst there were 52,508 articles containing the keyword in both title and abstract.

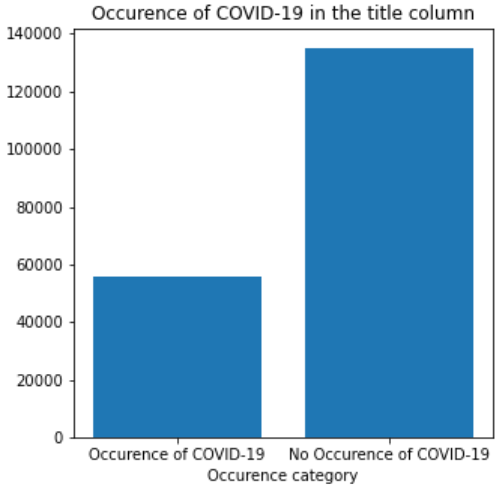

Figure1. Occurrence of COVID-19 in the title column

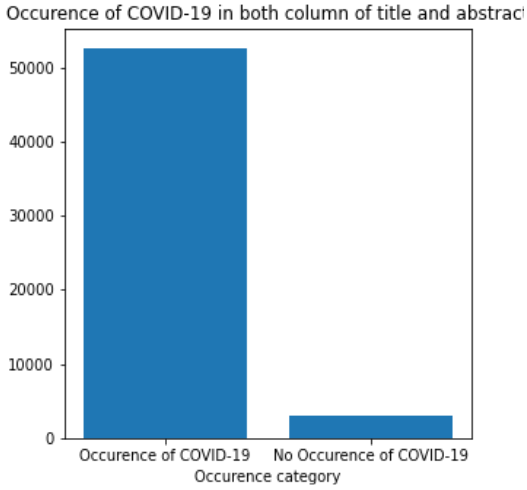

Figure2. Occurrence of COVID-19 in both abstract and title column

Therefore, 52,508 rows in the data frame remained. Afterwards, this project was going to utilize the most recent articles to support the final question and answering tool. Only the articles having published date on and after 01/01/2021 were selected. At this stage, only 27,203 rows of data in the dataset were left for further selection. Figure 3 showed the general distribution of the selected 27,703 articles published in different months in 2021 up to October.

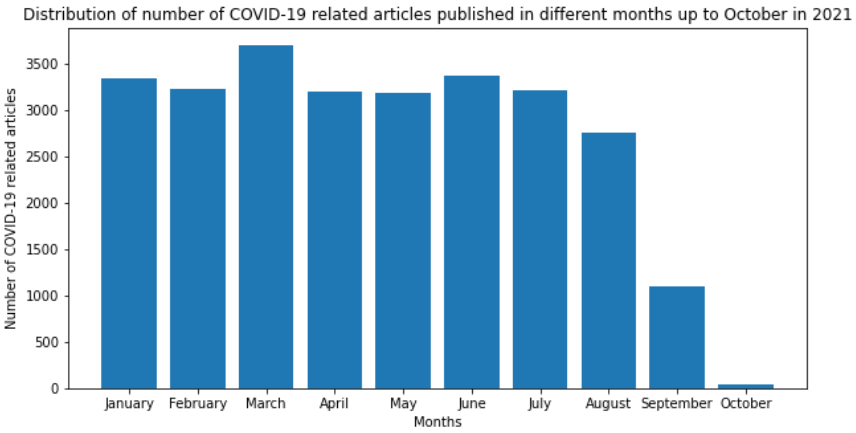

Figure3. Distribution of number of COVID-19 related articles published in different months in 2021

One of the objectives of this question and answering tool was to answer some COVID-19 related questions for individuals who were living in the countries with most COVID-19 cases. The top 10 countries of reported COVID-19 cases were the US, India, Brazil, UK, Russia, Turkey, France, Iran, Argentina and Spain (Worldometer, 2020). The next step was to check the occurrence of these countries names in the titles column of the selected dataset. Figure 4 illustrated the distribution of the selected articles having keywords related to the top 10 countries of reported COVID-19 cases. Therefore, it could be said that the selected COVID-19 related articles were most relevant to the countries with the highest COVID-19 cases.

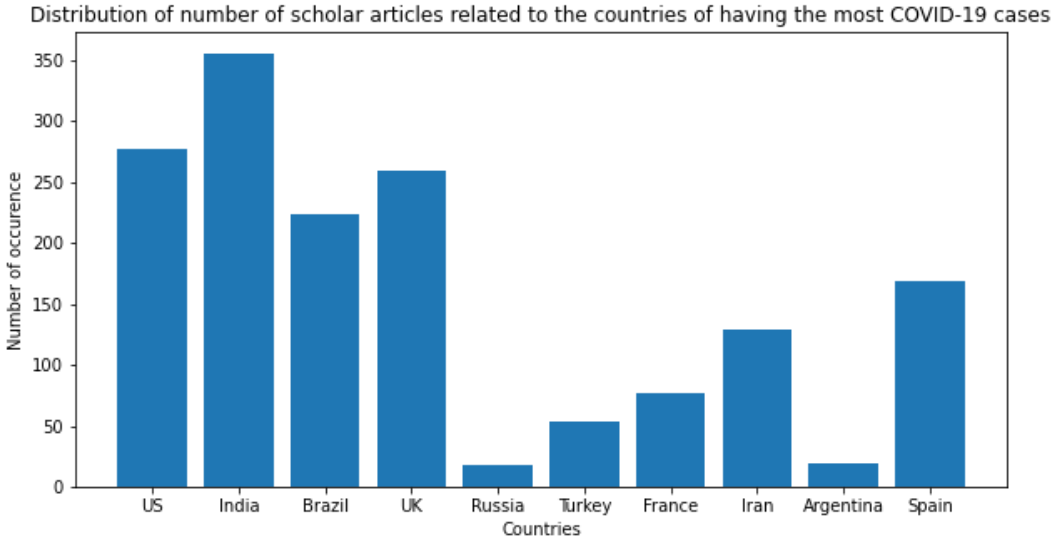

Figure4. Distribution of the selected articles having keywords related to the top 10 countries of reported COVID-19

The next step was to extract body context from the chosen full-text articles. By using the function of ultra-fast JSON decoder with Python, 398,002 body contexts were firstly selected from the chosen articles. The top leading countries in publishing science research were the US, China, Germany, the UK, Japan, France, Canada, Switzerland, South Korea and Australia (Nature index, 2020). To provide a quality answer, the chosen body contexts should be mostly published by the top leading science research published countries. A summary of the distribution of the selected body contexts published by

the top leading countries in publishing science research was illustrated in Figure 5.

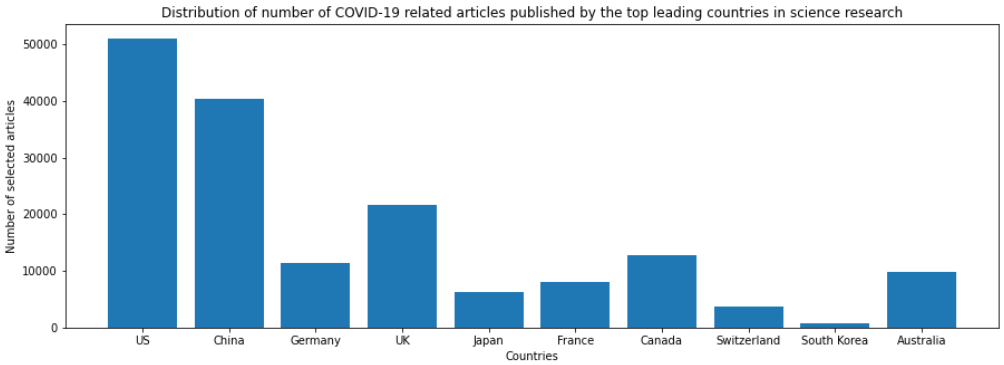

Figure5. Distribution of number of selected body contexts published by the top leading countries in publishing research

The next step was extracting body contexts from the abstract column in the data frame and concatenating it with the body context selected from the previous section as in one finalised data frame prior to the implementation of Word2Vec word embedding processing. Before sending the data to the Word2Vec model, a final check in the selected sentences of occurrence of COVID-19 frequently asked question (FAQ) keywords were investigated. The keywords existing in the COVID-19 related FAQ were coronavirus, symptoms, transmissions, incubation period, treatment, death rate, prevent, travel face mask, social distancing, children, asymptomatic, contagious and contact tracing (Chp.gov.hk, 2021). The distribution was shown in Figure 6 below.

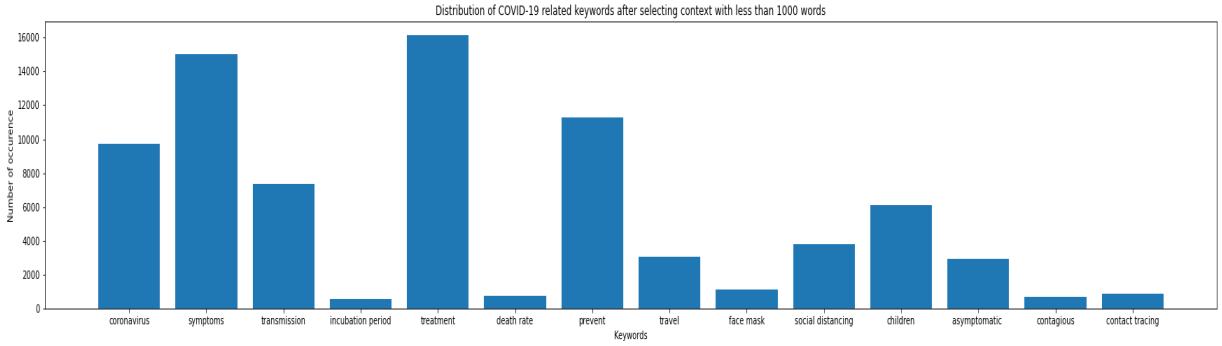

Figure6. Distribution of occurrence of some keywords selected from the top 10 most commonly asked questions related to COVID-19

To provide a quality answer without showing a prolonged paragraph of body context, only body context with less than 1000 words were selected as relevant sentences for the Question and Answering Tool in this project. Another illustration in the selected sentences with less than 1000 words of occurrence of COVID-19 keywords was also was shown in Figure 7 below to make sure the filtering procedure did not remove most of the body contexts with the relevant keywords. The result illustrated the distribution of keywords in the body contexts with less than 1000 words was almost identical to the original distribution in full texts.

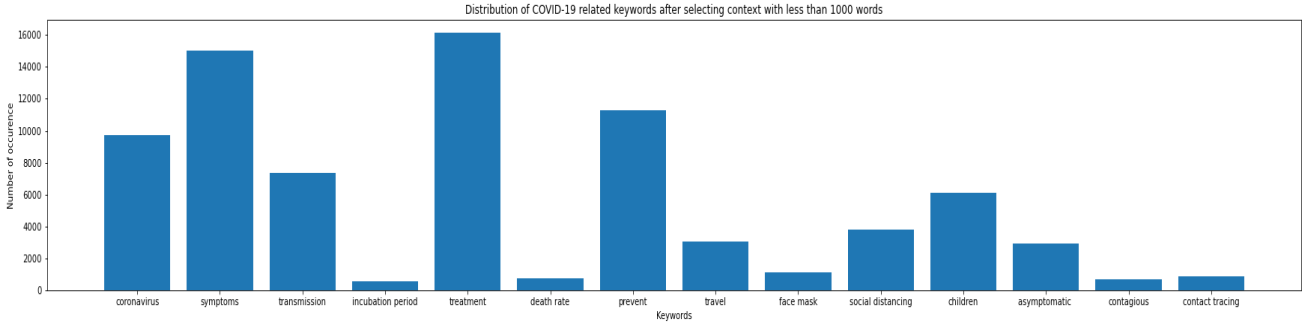

Figure7. Distribution of occurrence of some related keywords in the body contexts with less than 1000 words

4.2 Data pre-processing

All the keywords were included and the data was ready to be implemented to the Gensim library for a simple preprocessing process. Gensim was an open-source public library in Python for natural language processing. This project was going to utilise the Gensim library for simple preprocessing and implementing the Word2Vec algorithm. Gensim library offered a utility tool of simple preprocessing which converted the input documents into a list of lowercased, tokenized sentences without punctuations. This project was going to use the Word2Vec algorithm which created a window of a target word in a sentence and to learn the context between the neighbour words and the target word. Therefore, stopwords removal and lemmatization were not included due to the meaning dependency of each word from a sentence applied in a Word2Vec algorithm.

4.3 Word2Vec model implementation

The Word2Vec algorithm was implemented from the Gensim library. In this project, the parameter setting was shown in table 1 below. This project set the size of the sliding window as 5 that 5 words behind and after the target word would have a direct impact on its prediction outcome. 303,327 body contexts were finally used in the Word2Vec model to learn word embedding.

Table 1. Parameter setting

Window	5
Min_count	1
Workers	4
Sg	Default (CBOW)
Epochs	5
total example	303,327

4.4 Model testing with a single word cosine similarity

Prior to finalizing the question and answering tool, the model was tested by implementing a function of most_similiar from the library of Word2Vec algorithm to find similarity of different single vocabulary based on the cosine similarity algorithm. The result was shown in Table 2 below. In this case, the model successfully predicted the input word 'covid' and 'vaccine' in different vocabularies extracted from the body contexts which had similar meanings or different forms.

Table2. Cosine similarity of single vocabulary

Input: "covid"		Input: "vaccine"	
'disease'	0.5302287936210632	'vaccines'	0.8753696084022522
'cancer'	0.5025320649147034	'vaccination'	0.8389375805854797
'symptomatic'	0.4835883677005768	'vaccinations'	0.6372769474983215
'the'	0.4769422113895416	'vaccinated'	0.6372769474983215
'viral'	0.46182093024253845	'immunization'	0.6336705088615417
'influenza'	0.44902393221855164	'bnt'	0.5663754940032959
'infection'	0.4425373673439026	'fosun'	0.5370407104492188
'opth'	0.4408847987651825	'pfizer'	0.5253798961639404
'tochild'	0.43895629048347473	'biontech'	0.5144389867782593
'alonzi'	0.43624410033226013	'coronavac'	0.5115625858306885

4.5 Question and Answering Tool

After model testing, the model was ready to be implemented as a Question and Answering Tool. Word2Vec algorithm has a function of wv.n_similarity to compare the cosine similarity of two sets of tokenized words. The methodology of generating a question and answering tool in this project was to firstly get the input question from the user as a string from the input() Function in Python. The input string was then preprocessed through the function of gensim.utils.simple_preprocess. Then a set of strings was compared with a set of body contexts in different articles based on the similarity score. An average of individual vectors corresponding to the words that appeared in the sentences was calculated and compared the cosine similarity with the input question. The top 10 highest scores of cosine similarity body context were selected as the most relevant answers for the tool. An example was shown in Table 3 below with the title indicating the source of data. The overall result showed that most of the answers are relevant and logical for answering the query given by the users regarding the signs and symptoms of the COVID-19 in this example.

Table 3. Result from the Question and Answering Tool

Enter a question: What are signs and symptoms of the COVID-19?

Answer 1:
The occurrence of hyposmia among Indian COVID-19 patients is 26.1% and that of hypogeusia is 26.8%. The proportion of patients presenting with hyposmia as the first symptom is 7.67% and hypogeusia as first symptom is 3.13%. There was no statistically significant difference between presence or absence of hyposmia/hypogeusia and severity of stage of COVID-19 disease. More than 96% of the patients fully recovered their sense of smell and taste sensation by the end of 5 weeks. Increased public awareness measures regarding these symptoms and its prognosis are recommended, which can help in early diagnosis, isolation and prevention of spread of pandemic. It should be highlighted to the patients and clinicians that the hyposmia and hypogeusia are neither predictors nor protective of severity of COVID-19 disease.

Extracted from (ie.title of scholar article):
Course of Hyposmia and Hypogeusia and their Relationship with Severity of COVID-19 Disease among Indian Population

---

Answer 2:
Our students identified cough, fever, and fatigue as the most frequent symptoms and nausea, vomiting, and diarrhea as the less frequent, which agrees with what we found in recent studies concerning the prevalence of symptoms [37] [38] [39] [40] [41] . The amount of information available and its dissemination might justify this correct identification of symptoms. One of the relevant measures to reduce the transmission of the disease is the fast and correct identification of cases, where the recognition of the symptoms is crucial.

Extracted from (ie.title of scholar article):
The Impact of Health Literacy on Knowledge and Attitudes towards Preventive Strategies against COVID-19: A Cross-

Sectional Study

Answer 3:
Cutaneous symptoms can be the first in appearance, accompany respiratory symptoms, or be subsequent. Given the limited number (in this pandemic) of internal medicine, emergency, and intensive care clinicians, it is critical that all physicians can identify signs and symptoms associated with COVID-19. Considering that Latin America is the current epicenter of the pandemic, and Europe presented the third wave of new cases, this article's objectives were, first, to present images of cutaneous lesions observed in COVID-19 patients and, second, to perform a review of the available literature on COVID-19 and skin manifestations.

Extracted from (ie.title of scholar article):
Cutaneous manifestations of COVID-19 in Mexican patients: A case series and review of literature Case Report

---

Answer 4:
Any abnormal symptoms in the elderly should be taken seriously, as COVID-19 has a variety of symptoms. Heart, gastrointestinal, and respiratory problems are some of the symptoms of COVID-19, but the point is that the symptoms of the disease in the elderly are elliptical. Sometimes the disease occurs in them only with drowsiness. We always advise taking any change in the elderly seriously, because the slightest change can signify the onset of COVID-19. An elderly person in stable condition who suddenly has a change in condition should see a doctor immediately.

Extracted from (ie.title of scholar article):
COVID-19 Has Made the Elderly Lonelier

Answer 5:
Although symptoms and signs are often added to the Centers for Disease Control and Prevention's list of indicative of COVID-19, we chose these definitions of "typical" and "atypical" presentations because (i) fever, cough, and shortness of breath remain the "hallmark" symptoms of COVID-19 despite that other symptoms or signs are considered common; (ii) our categorization is based on symptoms and signs considered typical or atypical at the time at which this manuscript was prepared; and (iii) we followed the categorization of what is considered typical and atypical in the existing COVID-19 literature (2) (3) (4) (5) 7, 9, 11) .

Extracted from (ie.title of scholar article):
Medical Sciences cite as

Answer 6:
The most common symptoms of COVID-19 are: dry cough, fever, muscle ache, fatigue, and shortness of breath, as depicted in Figure 5 . Along with these, other less-observed symptoms are diarrhea, headache, and hemoptysis. The subject who possesses all these conditions is a person infected with the COVID-19 virus. As time passes, the virus eventually affects the lungs' functionality with the impact increasing up to 14 days. Among the symptoms, research has found that body temperature and dry cough are the vital diagnosis parameters of COVID-19. In Table 3 , we summarize the recent studies conducted in screening and tracking the symptoms of COVID-19, as well as the technologies that can be adopted to prevent people from becoming infected by this deadly virus.

Extracted from (ie.title of scholar article):
The Rise of Wearable Devices during the COVID-19 Pandemic: A Systematic Review

---

Answer 7:
COVID-19 has respiratory and systemic implications. The clinical and epidemiological characteristics are comparable with SARS [60] . Diagnosis is not only based on the symptoms but also on the history of exposure to the virus. Thus, effective tests are required to recognize patients regardless of the presence of symptoms.

Extracted from (ie.title of scholar article):
Journal Pre-proof Assessment and management of asymptomatic COVID-19 infection: A Systematic Review Systematic Review Assessment and management of asymptomatic COVID-19 infection: A Systematic Review

---

Answer 8:
To our knowledge, worsening of PD symptoms as the sole initial manifestation of SARS-CoV-2 infection, in the absence of other cardinal features of COVID-19 infection, has not been reported, expanding the disease clinical spectrum. In the era of COVID-19, we suggest that any unexplained worsening of PD symptoms, especially in advanced cases, should warrant for COVID-19 testing, even in the absence of its usual cardinal features.

Extracted from (ie.title of scholar article):
N/A

Answer 9:
• The shortage in medical equipment needed to perform RT-PCR tests, and thus, the inability to perform the 2 M. Alrahhal And K.P. Supreethi Studies have shown that most cases of COVID-19 disease develop to mild respiratory and constitutional symptoms such as fever, cough, dyspnea, myalgia and fatigue [6, 7] . The symptoms vary according to the degree of infection that may be a mild, moderate or severe infection. Studies have also shown that the examination of the chest X-ray reveals the changes that occur on the patient's lungs 4-5 days after the onset of symptoms [8] , whereas the computed tomography (CT) scan reveals these changes 2 days after the onset of symptoms [9, 10] or when symptoms begin to appear [11] . The effect of the COVID-19 on the patient's lungs in the case of an X-ray and CT scan is illustrated in Table 1.

Extracted from (ie.title of scholar article):
COVID-19 Diagnostic System Using Medical Image Classification and Retrieval: A Novel Method for Image Analysis

---

Answer 10:
COVID-19 may manifest as mild, moderate or severe disease with each grade of severity having its own features and post-viral implications. With the rising burden of the pandemic, it is vital to identify not only active disease but any post-recovery complications as well. This study was conducted with the aim of identifying the presence of post-viral symptomatology in patients recovered from mild COVID-19 disease. Presence or absence of 11 post-viral symptoms was recorded and we found that 8 of the 11 studied symptoms were notably more prevalent amongst the female sample population. Our results validate the presence of prolonged symptoms months after recovery from mild COVID-19 disease, particularly in association with the female gender. Hence, proving the post-COVID syndrome is a recognizable diagnosis in the bigger context of the post-viral fatigue syndrome.

Extracted from (ie.title of scholar article):
Follow-up of COVID-19 recovered patients with mild disease

5 Discussion and Conclusion

This project was divided into several sections including exploratory data analysis, data selection, data preprocessing, Word2Vec model training, model testing and finally building a question and answering tool. In this project, data selection was one of the most essential sections which had a direct impact on the quality of the output of the question and answering tools. Since this project was going to build an updated COVID-19 related tool, only scholar articles published in 2021 and had the keyword COVID-19 in their abstracts and titles were selected as the data trained in the Word2Vec model. Moreover, some of the other statistical insights such as the articles' published countries, the frequent occurrence of keywords in real-world COVID-19 frequently asked questions, the countries where had the most severe COVID-19 cases were also highlighted in the selected dataset to make sure the Question and Answering tool could be trained to provide a high quality answers to the individuals who had queries about this pandemic. Before delivering the final model, it was tested by inputting some COVID-19 related vocabularies to match the similarity of other words that appeared in the selected articles. And finally, the Word2Vec model effectively generated the top 10 most relevant answers extracted from the body context of articles based on the function of cosine similarity to answer the question regarding COVID-19.

6 REFERENCES

Chp.gov.hk. 2021. Centre for Health Protection, Department of Health - Frequently Asked Questions on Coronavirus Disease 2019 (COVID-19). [online] Available at: <https://www.chp.gov.hk/en/features/102624.html> [Accessed 28 October 2021].

Nature Index. 2020. The ten leading countries in natural-sciences research. [online] Available at: <https://www.nature.com/articles/d41586-020-01231-w> [Accessed 28 October 2021].

Riva, M., 2021. Word Embeddings: CBOW vs Skip-Gram. [online] baeldung.com. Available at: <https://www.baeldung.com/cs/word-embeddings-cbow-vs-skip-gram> [Accessed 30 October 2021].

Worldometer (2020). Countries where Coronavirus has spread - Worldometer. [online] www.worldometers.info. Available at: https://www.worldometers.info/coronavirus/countries-where-coronavirus-has-spread/.