

# **CLC 12 – Capstone Project Report**

## **Covitionary: Development of a Covid-19 Vocabulary using Text-Mining**

**Harjap Gosal**

We have been facing a challenge of our lifetimes in the form of the Covid-19 pandemic since the start of 2020. The scientists and researchers of the world are working hard to control the deadly virus and the disease. It has led to the creation of a large amount of scientific and non-scientific literature (1-4) covering the different aspects of the pandemic, such as surveillance, prevention, diagnosis, transmission, treatment drugs, and vaccines, post-treatment dynamics, to name a few. The automatic means of text-mining and standardized resources could help researchers in dealing with the overwhelming literature.

The objective of the project was to develop a resource for a standardized vocabulary of terms in the Covid-19 domain (Covitionary) by mining the literature and automatic generation of Covid terms from the text. For this, the Natural Language Processing (NLP) techniques, especially text-mining, were used and implemented in Python programming language to acquire the various terms from the text (target of 10000 terms) and organize these in a vocabulary. The Covitionary resource has currently 12184 most frequent terms used in the COVID-19 domain literature. This resource has many potential applications as it has been made publically available via GitHub, a public code hosting platform, at <https://github.com/HarjapGosal/Covitionary>.

## **EXPERIMENTATION**

### **Dataset**

#### **Input corpus data**

LitCovid is a curated literature hub for tracking up-to-date scientific information about the 2019 novel Coronavirus (1-3) that provides centralized access to more than 100000 (and

expanding) related articles in PubMed (124000 as of May 2021 ). The input data consisting of titles of these papers from LitCovid (1-3) has been used for garnering the most commonly used terms and building a large terminological resource (covitionary) in Covid-19 domain.

The source input data ( a snapshot shown in **Fig. 1**) could be accessed at

<https://github.com/HarjapGosal/Covitionary/blob/main/covitionary/data/input/covid-literature-input.csv>

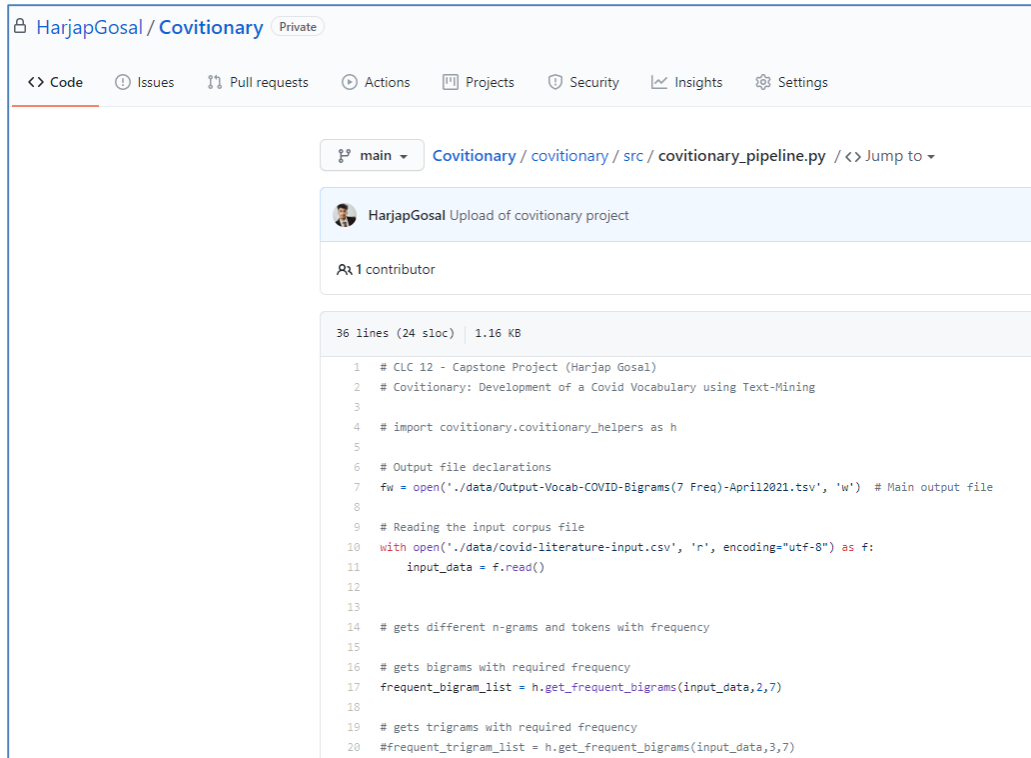
Paper	Title
•	<i>Early adoption of non-pharmaceutical interventions and COVID-19 mortality.</i>
•	<i>ABO blood groups, COVID-19 infection and mortality.</i>
•	<i>Patient satisfaction scores with telemedicine in the neurosurgical population.</i>
•	<i>An international perspective of out-of-hospital cardiac arrest and cardiopulmonary resuscitation during the COVID-19 pandemic.</i>
•	<i>Acute Fulminant Cerebellitis in Children With COVID-19: A Rare But a Treatable Complication.</i>
•	<i>Eight-month follow-up of olfactory and gustatory dysfunctions in recovered COVID-19 patients.</i>
•	<i>Resveratrol-zinc nanoparticles or pterostilbene-zinc: Potential COVID-19 mono and adjuvant therapy.</i>
•	<i>Objectively measured digital technology use during the COVID-19 pandemic: Impact on depression, anxiety, and suicidal ideation among young adults.</i>
•	<i>Phenoxazine nucleoside derivatives with a multiple activity against RNA and DNA viruses.</i>
•	<i>Mental disorder prevalence among populations impacted by coronavirus pandemics: A multilevel meta-analytic study of COVID-19, MERS &amp; SARS.</i>
•	<i>Screening of potent phytochemical inhibitors against SARS-CoV-2 protease and its two Asian mutants.</i>
•	<i>From the Cochrane Library: Tele dermatology for Diagnosing Skin Cancer in Adults.</i>
•	<i>Discovery of naturally occurring inhibitors against SARS-CoV-2 3CL(pro) from Ginkgo biloba leaves via large-scale screening.</i>
•	<i>A cell-based assay to discover inhibitors of SARS-CoV-2 RNA dependent RNA polymerase.</i>
.....	
.....	

**Figure 1:** A snapshot of COVID 19 corpus of 12800 paper titles (Source: 1-3)

## IMPLEMENTATION

### Software

The term extractor tool is implemented in Python language. The algorithm is based on the “text-mining” technique. The “[covitionary\\_pipeline.py](#)” (a snapshot shown in **Fig. 2**) is the main program which implements the system. The implemented code for applying this technique performs certain functions implemented in terms of several routines available in [covitionary\\_helpers.py](#) (a snapshot shown in **Fig. 3**).



HarjapGosal / Covitionary Private

<> Code Issues Pull requests Actions Projects Security Insights Settings

main Covitionary / covitionary / src / covitionary\_pipeline.py / <> Jump to

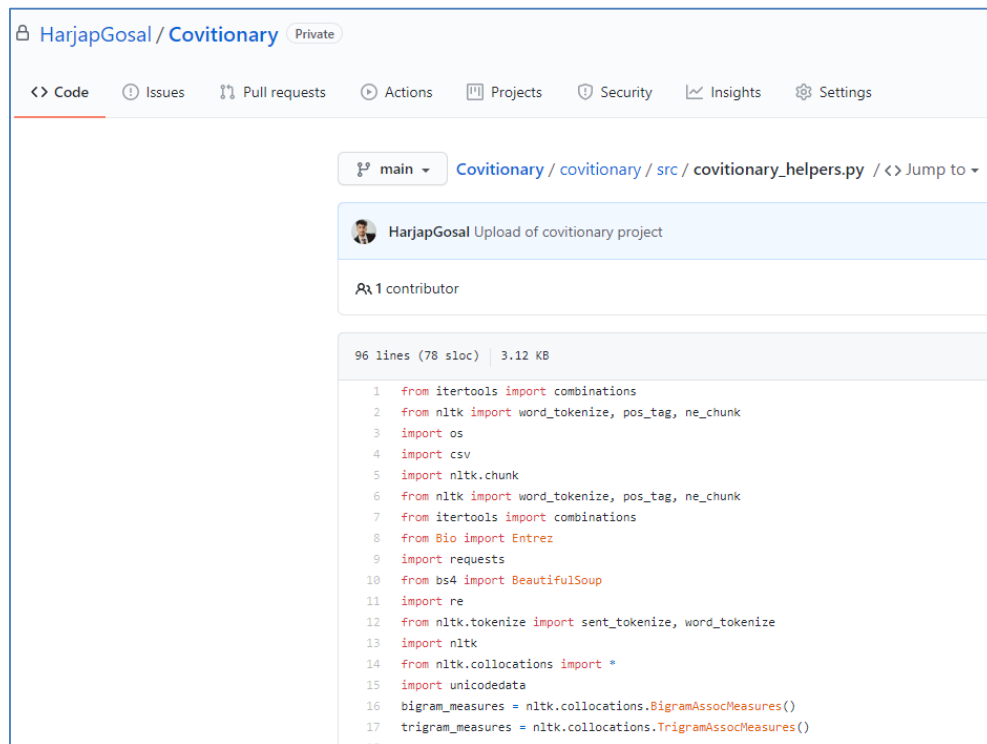
HarjapGosal Upload of covitionary project

1 contributor

36 lines (24 sloc) 1.16 KB

```
1 # CLC 12 - Capstone Project (Harjap Gosal)
2 # Covitionary: Development of a Covid Vocabulary using Text-Mining
3
4 # Import covitionary.covitionary_helpers as h
5
6 # Output file declarations
7 fw = open('./data/Output-Vocab-COVID-Bigrams(7 Freq)-April2021.tsv', 'w') # Main output file
8
9 # Reading the input corpus file
10 with open('./data/covid-literature-input.csv', 'r', encoding="utf-8") as f:
11     input_data = f.read()
12
13
14 # gets different n-grams and tokens with frequency
15
16 # gets bigrams with required frequency
17 frequent_bigram_list = h.get_frequent_bigrams(input_data,2,7)
18
19 # gets trigrams with required frequency
20 frequent_trigram_list = h.get_frequent_bigrams(input_data,3,7)
```

**Figure 2:** A snapshot Covitionary pipeline (covitionary\_pipeline.py)



HarjapGosal / Covitionary Private

<> Code Issues Pull requests Actions Projects Security Insights Settings

main Covitionary / covitionary / src / covitionary\_helpers.py / <> Jump to

HarjapGosal Upload of covitionary project

1 contributor

96 lines (78 sloc) 3.12 KB

```
1 from itertools import combinations
2 from nltk import word_tokenize, pos_tag, ne_chunk
3 import os
4 import csv
5 import nltk.chunk
6 from nltk import word_tokenize, pos_tag, ne_chunk
7 from itertools import combinations
8 from Bio import Entrez
9 import requests
10 from bs4 import BeautifulSoup
11 import re
12 from nltk.tokenize import sent_tokenize, word_tokenize
13 import nltk
14 from nltk.collocations import *
15 import unicodedata
16 bigram_measures = nltk.collocations.BigramAssocMeasures()
17 trigram_measures = nltk.collocations.TrigramAssocMeasures()
18
```

**Figure 3:** A snapshot Covitionary pipeline (covitionary\_helpers.py)

## RESULTS AND EVALUATION

### Output

The outputs are generated for individual word frequency, most frequent bigrams i.e 2-word phrase (based on different frequencies), most frequent trigrams i.e 3-word phrase (based on different frequencies), and most frequent quadgrams i.e 4-word phrase (based on different frequencies). Assumption is that more than 4-word phrases as terms will be rare and not most frequent. The following is the list of output files (available at - <https://github.com/HarjapGosal/Covitionary/tree/main/covitionary/data/output>) from which the Covitionary is created:

- Output-Vocab-COVID-April2021.tsv
- Output-Vocab-COVID-Bigrams(7 Freq)-April2021.tsv
- Output-Vocab-COVID-Bigrams-April2021.tsv
- Output-Vocab-COVID-Frequency-April2021.tsv
- Output-Vocab-COVID-Quadgrams-April2021.tsv
- Output-Vocab-COVID-Trigrams-April2021.tsv

The **Tables 1-5** show the snapshot of various outputs to reflect the terms automatically generated by the system.

**Table 1:** The top most 10 most frequent individual words with frequency.

Word	Frequency
covid19	84018
pandemic	23119
patients	16306
sarscov2	15637
coronavirus	12260
disease	10065
health	8247
study	7559
infection	7104
care	6416

**Table 2:** The top 10 most frequent bigrams (frequency>500).

- covid-19 pandemic
- coronavirus disease
- covid-19 patients
- sars-cov-2 infection
- mental health
- covid-19 infection
- systematic review
- covid-19 outbreak
- novel coronavirus
- severe covid-19

**Table 3:** The top 10 most frequent trigram (frequency>200).

- coronavirus disease 2019
- critically ill patients
- personal protective equipment
- intensive care unit
- health care workers
- New York city
- retrospective cohort study
- novel coronavirus disease
- multisystem inflammatory syndrome
- acute kidney injury

**Table 4:** The top 10 most frequent quad-grams (frequency>100).

- severe acute respiratory syndrome
- acute respiratory syndrome coronavirus
- coronavirus disease 2019 pandemic
- acute respiratory distress syndrome
- critically ill covid-19 patients
- coronavirus disease 2019 patients
- severe coronavirus disease 2019
- acute respiratory syndrome coronavirus-2
- novel coronavirus disease 2019
- corona virus disease 2019

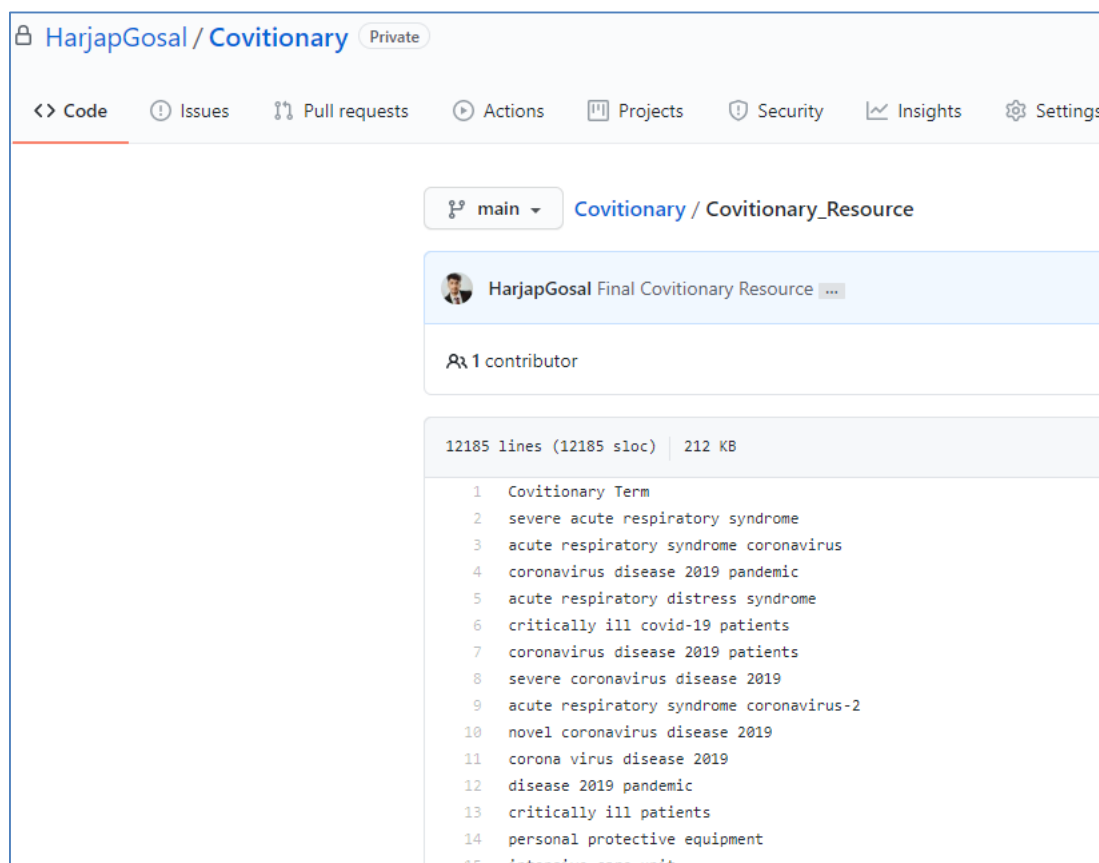
**Table 5:** The top 40 most frequent terms (more than 100 occurrences).

- severe acute respiratory syndrome
- acute respiratory syndrome coronavirus
- coronavirus disease 2019 pandemic
- acute respiratory distress syndrome
- critically ill covid-19 patients
- coronavirus disease 2019 patients
- severe coronavirus disease 2019
- acute respiratory syndrome coronavirus-2
- novel coronavirus disease 2019
- corona virus disease 2019
- disease 2019 pandemic
- critically ill patients
- personal protective equipment
- intensive care unit
- health care workers
- New York city
- retrospective cohort study
- novel coronavirus disease
- multisystem inflammatory syndrome
- acute kidney injury
- extracorporeal membrane oxygenation
- inflammatory bowel disease
- coronavirus disease 2019
- covid-19 pandemic
- coronavirus disease
- covid-19 patients
- sars-cov-2 infection
- mental health
- covid-19 infection
- systematic review
- covid-19 outbreak
- novel coronavirus
- severe covid-19
- case report
- public health
- covid-19 pneumonia
- respiratory syndrome
- health care
- cohort study
- United States

## COVITIONARY RESOURCE

From the output files and after analysis of the results, the following dictionary of Covid-19 domain terms, **Covitionary**, is generated that contains **12184** frequent terms (**Fig. 4** shows a snapshot of Covitionary resource available at GitHub):

[https://github.com/HarjapGosal/Covitionary/blob/main/Covitionary\\_Resource](https://github.com/HarjapGosal/Covitionary/blob/main/Covitionary_Resource)



**Figure 4:** A snapshot of Covitionary resource available at GitHub.

## POTENTIAL APPLICATIONS

Covitionary could be used for spell checking, making ontologies (that are standard hierarchical vocabularies in a particular domain). It could also be used for further classification of terms in pre-defined categories in the Covid19 domain such as Treatment, Diagnosis, Transmission, Prevention, Research and Study, Travel, Forecasting, Mechanism/ Device/Tool, etc. One such example of classification as aproo-of-concept is given in the **Table 6**.

**Table 6:** Proof-of-concept of classification of terms in pre-defined categories in the Covid19 domain (showing 10 terms for 5 categories from the gathered terms)

<b>Prevention</b>	<b>Treatment</b>	<b>Potential Covid Drugs</b>	<b>Covid Related Travel Terms</b>	<b>Transmission Terms</b>
infection prevention	corticosteroid	azithromycin	travel restrictions	aerosol transmission
antibody detection	intravenous fluids	favipiravir	air travel	airborne transmission
biosecurity concern	oral fluids	umifenovir	covid-19 travel	asymptomatic transmission
close proximity	orogastric fluids	colchicine	international travel	community transmission
early detection	medication agent	lopinavir	travel medicine	contact transmission
serological testing	experimental agent	chloroquine	travel history	covid-19 transmission
antibody response	antimalarial agent	ritonavir	returning travelers	cross-species transmission
control measures	antifungal agent	ilaris	travel bans	disease transmission
preventive measures	antiviral medication	hydroxychloroquine	travel health	droplet transmission

## CONCLUSIONS

With every passing day, an overwhelming amount of literature is being generated by the researchers working in the Covid-19 domain. The terminological resources are the basis of many useful applications in any domain, and Covitionary (a dictionary of most frequent terms in Covid-19 domain) is one such useful terminological resource. The use of an automatic text-mining pipeline here has resulted in 12184 terms in Covid-19 related texts that have a frequency of 7 or more in the text. This resource has potential applications in many natural language processing tasks in the Covid-19 domain as some of these are outlined earlier.

## REFERENCE

1. Chen Q, Allot A, Lu Z. Keep up with the latest coronavirus research. *Nature*. 2020 Mar;579(7798):193-.
2. Chen Q, Allot A, Lu Z. LitCovid: an open database of COVID-19 literature. *Nucleic Acids Research*. 2021 Jan 8;49(D1):D1534-40.
3. Wei CH, Allot A, Leaman R, Lu Z. PubTator central: automated concept annotation for biomedical full text articles. *Nucleic acids research*. 2019 Jul 2;47(W1):W587-93.
4. Wang LL, Lo K, Chandrasekhar Y, Reas R, Yang J, Eide D, Funk K, Kinney R, Liu Z, Merrill W, Mooney P. Cord-19: The covid-19 open research dataset. *ArXiv*. 2020 Jul 9.