Question Answering Tool

CHEUK HANG NG (A1821087)

Outline

- Introduction
- Methodology
- Experiments and results
- Discussion and conclusions

Introduction

- Question Answering Tool
  - Aim: show the top 10 most relevant sentences from the articles in the dataset with their sources for a given query
  - Original dataset from the COVID-19 Open Research Dataset Challenge (CORD-19) Kaggle competition – Using a subset of the original dataset for this project
- Dataset
  - Around 6,000 articles in json format
- Natural language processing problem
  - Word embeddings

Methodology

- Data Preprocessing
  - Dataframe manipulation – Merging dataframe that contains all extracted articles with metadata.csv
  - Iterate over rows to unpack abstract, main content and country of the first author
    - Handle missing values
    - Drop useless columns
    - Reset index
- Exploratory Data Analysis
  - Plot number of publications against country of origin
  - Plot number of publications over time after 2003
  - Generate word cloud to observe frequent words

Methodology

- Text Preprocessing
- Text cleaning

```
def text_process1(text):
   #lowercase all characters
    text = text.lower()
    #remove punctuation
    text  =  re.sub(r'[%s]'  %  re.escape(punc),  '  ',
text)
    #remove unicode text
    text = re.sub(r'[^\x00-\x7F]+', ' ', text)
    #remove the numbers
    text = re.sub(r'[0-9]', '', text)
    #remove double space
    text = re.sub(r'\s{2,}', ' ', text)
    return text
```

- Text normalization
  - Remove Stop-words from the English Stop-words set from NLTK
    - Additional Stop-words: “et”, “al”, “may”, “also”
  - Lemmatization
    - Reduce a word to its base form by analyzing the morphological information of the word
  - Tokenization
    - Decompose a sentence into a list of words

Methodology

- Word Embeddings
  - Word2Vec – Continuous Bag-of-Words
    - Capture both semantic and syntactic information of words
    - Predict the centre word from surrounding context
    - Maximizes the probability of a word being in a particular context with the form

      $$P(wi|wi-c, wi-c+1, \dots, wi-1, wi+1, \dots, wi+c-1, wi+c)$$

      where wi is a word at position i with c being the length of window
    - `Word2Vec(sentence_list, vector_size = 100, window = 8, min_count = 1, sg = 0, workers = 4)`
- Similarity Search
  - Prompts for user input
  - Vectorize input query and compute similarity between the query and every sentence using `cosine_similarity()` from Sklearn
  - Return 20 most similar sentences to avoid duplicated sentences in abstract and main content from the same article
  - Output only 10 distinct articles
- Evaluation
  - Human judgement

Experiments and results

- Data Preprocessing
  - 5,228 instances

	paper_id	title	doi	abstract	publish_time	authors	journal	url	main_content	country
0	000e754142ba65ef77c6fdffcbcbe824e141ea7b	Laboratory-based surveillance of hospital-acqu...	10.1016/j.ajic.2017.01.009	Of 7,772 laboratory-confirmed cases of respira...	2017-05-01	Choi, Hye-Suk; Kim, Mi-Na; Sung, Heungsup; Lee...	Am J Infect Control	https://doi.org/10.1016/j.ajic.2017.01.009; ht...	The human respiratory viruses include adenovir...	South Korea
1	00218ecac4058261da156a6848e05e72f77b4dfc	COVID-19, type 1 diabetes, and technology: why...	10.1016/s2213-8587(20)30155-8		2020-05-05	Danne, Thomas; Limbert, Catarina	Lancet Diabetes Endocrinol	https://www.sciencedirect.com/science/article/...	In the current coronavirus disease 2019 (COVID...	NaN
2	002552e66a3c872d09c3559eab95d82704dcef57	"This is an opportunity for leadership to lead...	10.21203/rs.3.rs-310774/v1	In March of 2020, academic research centers in...	2021-03-16	Leonard, Chelsea; Connelly, Brigid; Albright, ...	Res Sq	https://doi.org/10.21203/rs.3.rs-310774/v1; ht...	The COVID-19 pandemic has caused far reaching ...	NaN
3	00339c93e11141d71c66e8562f5cb4020e6def2c	HIV- und AIDS-Patienten auf der Intensivstation	10.1007/s00390-003-0356-5	The number of HIV-infected patients in Germany...	2003	Mandraka, Falitsa; Salzberger, Bernd; Glück, T...	Intensivmed Notfallmed	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7...	n Summary The number of HIVinfected patients i...	Germany
4	003bb67952423a80571fd7f10b4ca9705384c4f7	A Cross Sectional Study of Midwifery Students'...	10.1016/j.nepr.2021.102988	The impact of COVID-19 on midwifery students i...	2021-02-09	Kuliukas, Lesley; Hauck, Yvonne; Sweet, Linda;...	Nurse Educ Pract	https://doi.org/10.1016/j.nepr.2021.102988; ht...	were collected through an online survey and se...	NaN
...	...	...	...	...	...	...	...	...	...	...
5223	ffc0b608f917bb85d07b8edbe54a26f29bb37604	EdNet: A Large-Scale Hierarchical Dataset in E...	10.1007/978-3-030-52240-7_13	Advances in Artificial Intelligence in Educati...	2020-06-10	Choi, Youngduck; Lee, Youngnam; Shin, Dongmin;...	Artificial Intelligence in Education	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7...	. A possible scenario of a student using Santa...	Republic of Korea
5224	ffc89f2bd7f513dae147b4bd0043164db16e8dc0	A SARS-CoV–Specific Protein Enhances Virulence...	10.1007/978-0-387-33012-9_87		2006	Pewe, Lecia; Zhou, Haixia; Netland, Jason; Tan...	The Nidoviruses	https://www.ncbi.nlm.nih.gov/pubmed/17037583/	SARS-CoV is tentatively classified as a group ...	NaN
5225	ffda8ef411ddce1e90e758b689fb25f4b7904322	The impact of modelling choices on modelling o...	10.1007/s00477-020-01965-z	The choices that researchers make while conduc...	2021-01-03	Briz-Redón, Álvaro	Stoch Environ Res Risk Assess	https://doi.org/10.1007/s00477-020-01965-z; ht...	The pandemic caused by the coronavirus disease...	NaN
5226	ffe718db1820f27bf274e3fc519ab78e450de288	Replication enhancer elements within the open ...	10.1093/nar/gkr237	We provide experimental evidence of a replicat...	2011-05-27	Tuplin, A.; Evans, D. J.; Buckley, A.; Jones, ...	Nucleic Acids Res	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3...	Tick-borne encephalitis virus (TBEV) is a huma...	NaN
5227	fff8b9e88db122ffcbaf1daf6b697e44eaaffd93	Septic shock caused by Mycobacterium tuberculo...	10.1007/s001340050825		1999	Angoulvant, D.; Mohammedi, I.; Duperret, S.; B...	Intensive Care Med	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7...	Sir: Septic shock due to Mycobacterium tubercu...	France

Figure 2: Sample of rows in processed dataframe

Experiments and results

- Exploratory Data Analysis

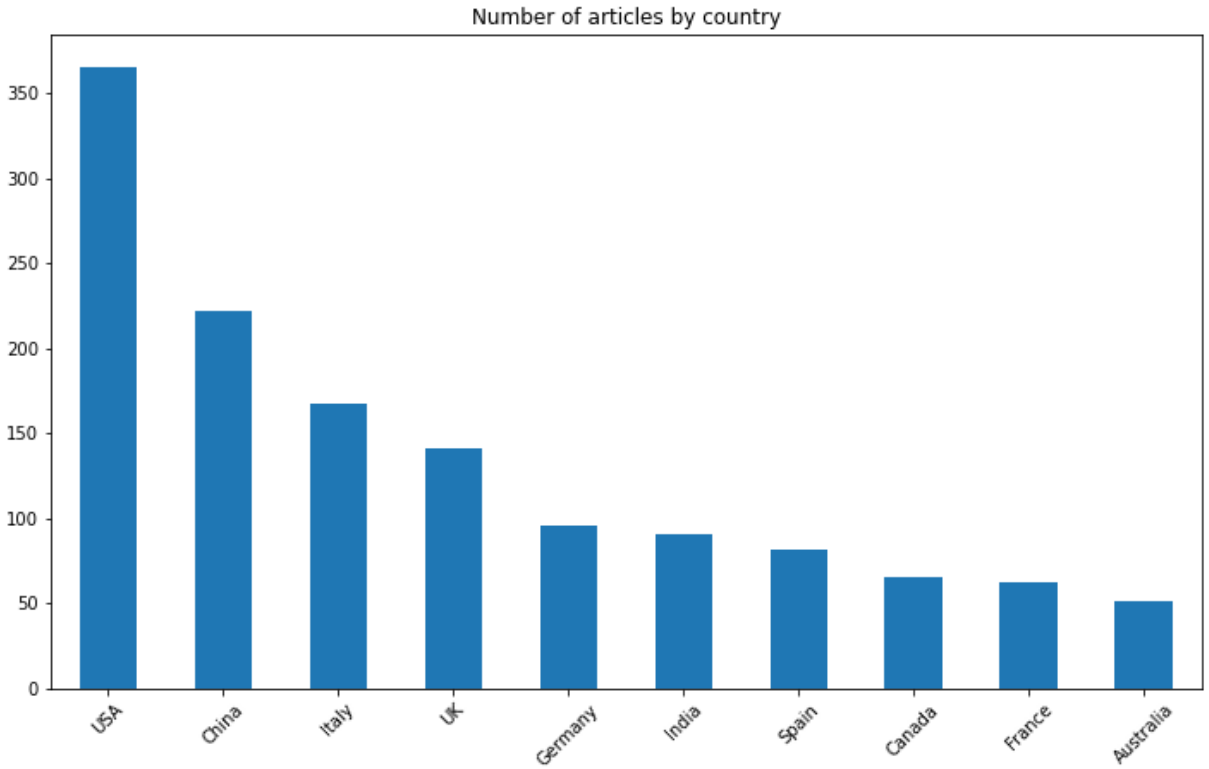

Figure 3: Number of publications with respect to country of origin

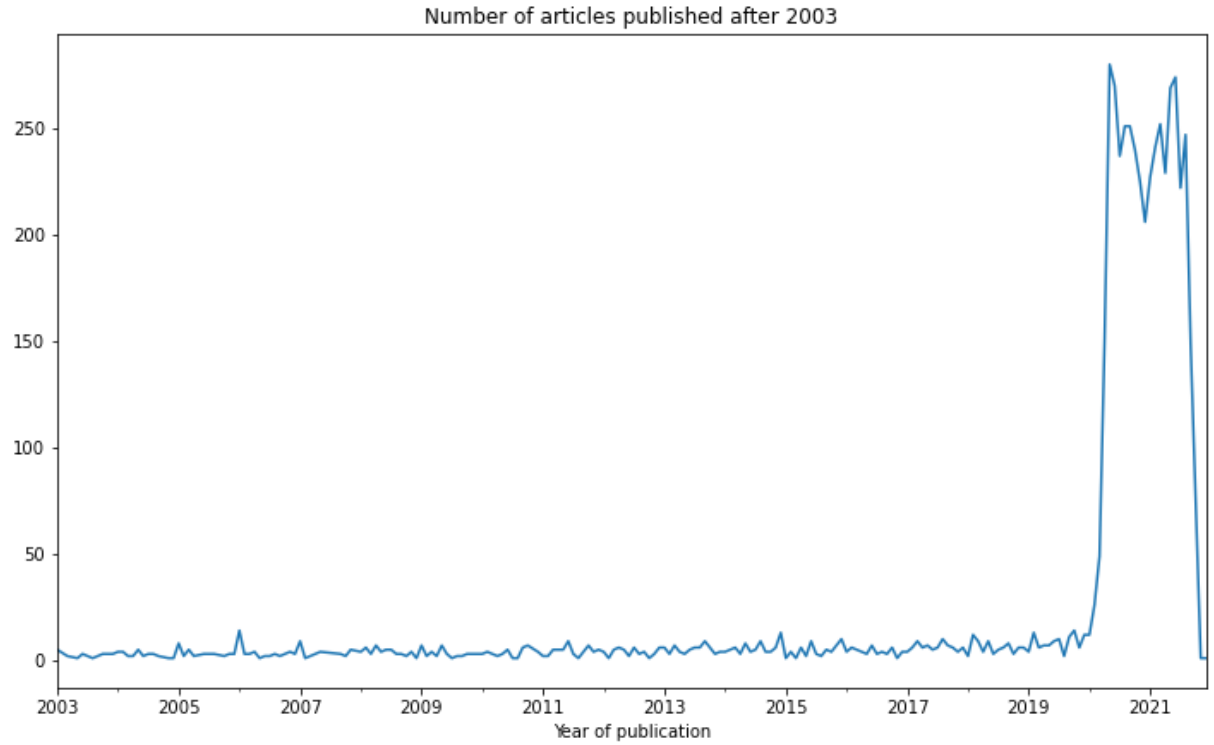

Figure 4: Number of publications over time after 2003

Experiments and results

- Exploratory Data Analysis

Figure 5: Word cloud generated from all sentences in the dataset

- Word Embeddings
  - Word2Vec – Continuous Bag-of-Words
  - Training data: 1,043,560 tokenized sentences
  - Example of tokenized sentences
    - [['laboratory', 'confirmed', 'case', 'respiratory', 'viral', 'infection', 'among', 'hospitalized', 'patient'], ['categorized', 'hospital', 'acquired', 'infection'], ['overall', 'incidence', 'hospital', 'acquired', 'respiratory', 'viral', 'infection'], ...]

Experiments and results

- Similarity Search and Search Results
  - “How long does it take to cure COVID” was used as an input and search was performed

Figure 6: 20 most similar tokenized sentences returned

```
['known', 'long', 'take', 'covid', 'patient', 'recover', 'infection']
['sick', 'leave', 'long', 'covid']
['new', 'zealand', 'recovery', 'covid', 'long', 'difficult']
['spite', 'proper', 'counselling', 'prognosis', 'long', 'term', 'follow', 'pemphigus', 'taking', 'prednisolone', 'unable', 'com
e', 'follow', 'covid', 'pandemic']
['present', 'uncertain', 'long', 'covid', 'pandemic', 'continue', 'thus', 'know', 'long', 'endoscopic', 'unit', 'return', 'func
tioning', 'normal', 'capacity']
['covid', 'patient', 'taken', 'longer', 'resolve', 'hyperglycaemic', 'crisis']
['covid', 'patient', 'taken', 'longer', 'resolve', 'hyperglycaemic', 'crisis']
['discussion', 'death', 'dying', 'put', 'long', 'confront', 'u', 'daily', 'longer', 'avoided']
['poliomyelitis', 'long', 'ago', 'feared', 'viral', 'disease', 'yet', 'eradicated', 'although', 'long', 'pursued', 'objective',
'seems', 'presently', 'within', 'reach']
['clinical', 'perspective', 'covid', 'long', 'incubation', 'period']
['initial', 'study', 'patient', 'long', 'term', 'complication', 'covid', 'aka', 'long', 'covid', 'mapping', 'risk', 'factor',
'complication', 'long', 'covid', 'patient', 'particularly', 'interesting', 'area', 'follow', 'research']
['sick', 'leave', 'protracted', 'sick', 'leave', 'long', 'covid', 'quite', 'common']
['herd', 'immunity', 'covid', 'never', 'occur', 'long', 'last', 'short', 'offer', 'level', 'protection', 'vulnerable', 'communi
ty']
['individual', 'suffering', 'post', 'covid', 'symptom', 'long', 'covid']
['difficulty', 'disease', 'detection', 'containment', 'long', 'course', 'covid']
['difficulty', 'disease', 'detection', 'containment', 'long', 'course', 'covid']
['repeated', 'asymptomatic', 'infection', 'help', 'maintain', 'long', 'standing', 'immunity']
['cancellation', 'sporting', 'event', 'lockdown', 'contain', 'spread', 'covid', 'short', 'long', 'term', 'consequence', 'especi
ally', 'athlete', 'coach']
['however', 'sizable', 'proportion', 'covid', 'survivor', 'fully', 'recover', 'suffer', 'postviral', 'syndrome', 'known', 'lon
g', 'covid', 'long', 'haul', 'covid', 'despite', 'released', 'hospital', 'tested', 'negative', 'sars', 'cov']
['sick', 'leave', 'least', 'week', 'thus', 'defined', 'long', 'covid']
```

Showing top 10 relevant articles

Assessment and Management of Diabetic Patients During the COVID-19 Pandemic3532

... in COVID-19 pathogenesis. [36] [37] [38] Therefore, disease intensity in COVID-19 patients is due both to the virus and the response of the host. 39 It is not known how long it takes for a COVID-19 patient to recover from infection. The infectivity period is estimated by detecting the virus or its ssRNA in respiratory tract samples. Though, presence of ssRNA of virus does not approve the contagious virus occurrence. Reports suggest that transmission of the virus may occur in the early stage of COVID-19 infection, as high virus loads are found in respiratory samples shortly after the onset ...

Original document at: https://www.ncbi.nlm.nih.gov/pubmed/34262317/; https://doi.org/10.2147/dmso.s285614

Patterns and predictors of sick leave after Covid-19 and long Covid in a national Swedish cohort1245

... to Covid-19. RESULTS: A total of 11,955 people started sick leave for Covid-19 within the inclusion period. The median sick leave was 35 days, 13.3% were on sick leave for long Covid, and 9.0% remained on sick leave for the whole follow-up period. There were 2960 people who received inpatient care due to Covid-19, which was the strongest predictor of longer sick leave. Sick leave the year prior to Covid-19 and older age also predicted longer sick leave. No clear pattern of socioeconomic factors was noted. CONCLUSIONS: A substantial number of people are on sick leave due to Covid-19. Sick leave ...

Original document at: https://www.ncbi.nlm.nih.gov/pubmed/34059034/; https://doi.org/10.1186/s12889-021-11013-2

On New Zealand's weak, strong and muddled management of a COVID-19 epidemic3327

... mental health problems, using e-health, virtual health teams and community health workers who are trained for this purpose and micro-credentialed.New Zealand's recovery from COVID-19 will be long and difficult. Best practice, both invented and borrowed, is essential. An explicit pandemic plan, which recognises the country's contextual strengths and weaknesses, and a national public health agency are essential ...

Original document at: https://doi.org/10.1111/imj.14928; https://www.ncbi.nlm.nih.gov/pubmed/32881265/

Kaposi Varicelliform Eruption in a Patient with Pemphigus Vulgaris: A Case Report and Review of the Literature435

... was diagnosed as pemphigus vulgaris (Figure 1). e patient was advised to take an oral steroid 60 mg daily along with other supportive treatments. In spite of proper counselling of the prognosis and long-term follow-up of pemphigus, he was taking prednisolone on and off and was unable to come for the follow-up because of the COVID-19 pandemic.e patient presented again 5 weeks before the admission with multiple erosions that extended to the anterior chest and was advised to continue prednisolone 50 mg once daily, azathioprine (100 mg once daily), doxycycline (100 mg once daily), and nicotinamide ...

Original document at: https://www.ncbi.nlm.nih.gov/pubmed/33489386/; https://doi.org/10.1155/2020/6695342

Endoscopy in inflammatory bowel diseases during the COVID-19 pandemic and post-pandemic period4445

... that, where possible, a separate endoscopy room should be used for these vulnerable patients. A flowchart for these patients is shown in figure 5 .At present, it is uncertain how long the COVID-19 pandemic will continue, and thus we do not know how long it will be before endoscopic units return to functioning at normal capacity. In the era of treating to target, with the goal of mucosal healing, it may be necessary to re-evaluate and prioritise endoscopic needs post-pandemic. Patients with IBD have a chronic disease with periods of relapse and remission 45 requiring endoscopy to monitor not ...

Original document at: https://www.ncbi.nlm.nih.gov/pubmed/32305075/; https://www.sciencedirect.com/science/article/pii/S2468125320301199; https://api.elsevier.com/content/article/pii/S2468125320301199; https://doi.org/10.1016/s2468-1253(20)30119-9

Discussion and conclusions

- With restricted storage and computing power, only around 5,000 articles were used to build the model.
- Hard to evaluate word embeddings model
- To improve the tool:
  - Data selection before data preprocessing
  - Data preprocessing strategies
  - Preprocess the text in a different way
    - Introduce stemming, spelling correction, part of speech tagging and negations handling to text preprocessing step
  - Continuous parameters tuning
    - Vector size, window length