CMPE 493 Extractive Text Summarization for COVID-19

Definition of the Project

In this assignment, I implement a text summarization model similar to the LexRank algorithm. LexRank is an unsupervised model which computes the relative importance of the sentences based on the cosine similarity and the PageRank algorithm. As the data set we will use articles from the COVID-19 Open Research Dataset (CORD-19) These articles are from the 10 April 2020 release of CORD-19 corpus. The size of the data is around 1.5GB.

We have relevance judgements. Each line contains the relevance judgement for a document. The first column is the topic-id, the second field is iteration (you will NOT be using this field), the third column is the document-id (cord-id), and the last column is the relevance judgement, where 0 means not-relevant, 1 means partially relevant and 2 means fully relevant.

I will use the documents of three topics, Topic 1: coronavirus origin, Topic 13: how does coronavirus spread, and Topic 18: mask prevent coronavirus. Topic 1 tries to answer the question “what is the origin of COVID-19?”, which can be described as “seeking range of information about the SARS-CoV-2 virus’s origin, including its evolution, animal source, and first transmission into humans”. Topic 13 tries to answer the question “what are the transmission routes of coronavirus?”, which can be described as “Looking for information on all possible ways to contract COVID-19 from people, animals and objects”. Topic 18 tries to answer the question “what are the best masks for preventing infection by Covid-19?”, which can be described as “What types of masks should or should not be used to prevent infection by Covid-19?”.

Given a topic, in the first step, the system will identify the most salient 10 documents for that topic. In the second step, the 10 selected documents will be summarized by selecting the most important 20 sentences from among all the sentences in these 10 documents. So, for each topic, I will obtain a 20 sentence summary.

To identify the most important documents in a given topic, I create a document graph, where each node corresponds to a document and each edge represents the TF-IDF weighted cosine similarity between the abstractcs of the corresponding two documents. I implement the PageRank algorithm and run it on this document graph to identify the 10 documents with the highest PageRank scores. For each topic, I use the documents that have a relevance score of 2 in the document-relevance file. After identifying the most important 10 documents for a topic, I create a sentence graph, where each node corresponds to a sentence (in the 10 documents) and each edge represents the TF-IDF weighted cosine similarity between the corresponding two sentences. I again implement the PageRank algorithm and run it on this sentence graph to identify the 20 sentences with the highest PageRank scores. These 20 sentences will be the summary of the given topic.

If the cosine similarity between two sentences/documents is less than 0.1, these sentences are not connected to each other, otherwise they are connected to each other in the graph. Also, the teleportation rate is set to 0.15 and the error tolerance in the power method is set to 0.00001.

Explanation of the Code

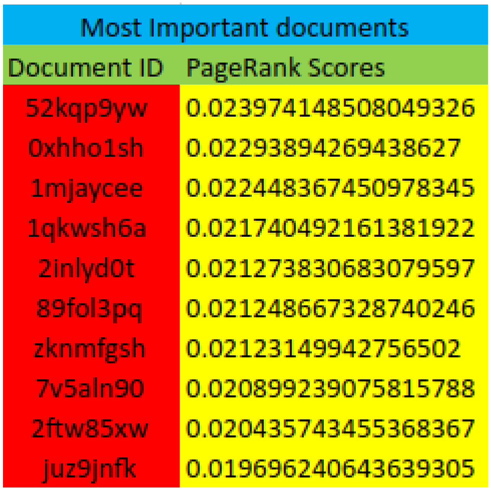
Firstly, I create my dictionary as a class which have add method and classic keys and values. The continuation, I ask the user question number for creating summary. After that in order, I read csv file and take 8. element from lines which reprsent abstracts of the documents. I create id list from “relationdoc.txt”. I take all full relevant document id’s according to entered question number. According to id list, I create “doc_list” which holds documents abstract. Lastly, I create “idfdict” and calculate idf scores of the words according the whole corpus.

In addition to these, I wrote 3 method which are: “tf”, “cosine_similarity” and “document_similarity”. Tf method takes 2 parameter. One of them word and second of them is document. This method find occurences of the word according to document. Cosine_similarity method takes 2 parameter which are 2 vector. With using these 2 vectors, it returns cosine_similarity of the 2 vectors according to cosine similarity formula. Document_similarity methods takes 2 parameters which are 2 documents. It creates tf-idf vectors of the documents and using cosine similarity methods, find cosine similarity between these 2 vectors.

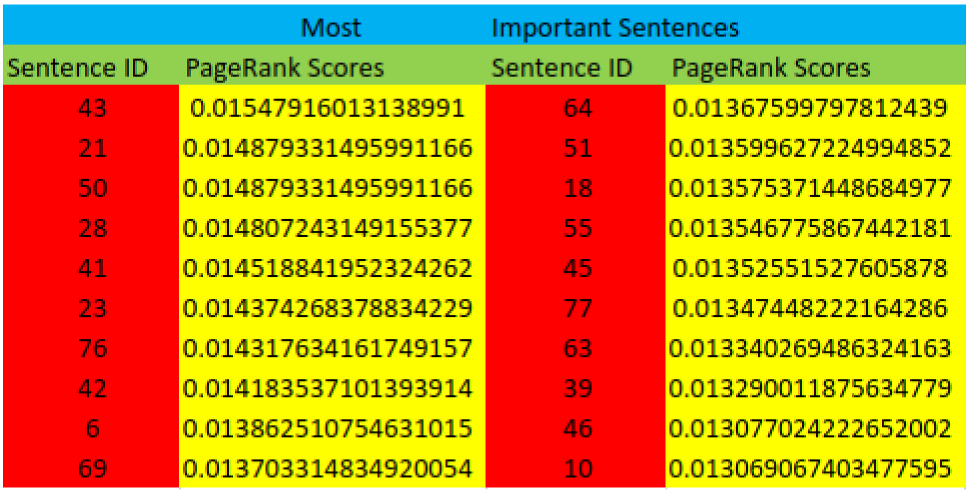
After these preparation parts, I create similarity array according to question number and relevant documents. For example, for question 1 we have 56 fully relevant documents. Therefore, I create 56x56 array and fill it with similarity using document similarity method between binary groups of the these 56 documents. However, I know that cosine similarity between document 1 and document 4 is equal to the document 4 and document 1. Therefore I thought an imaginary diagonal and I made half of the required calculations, then the rest was the reflection of the part which I calculate and filled.

In this part, I create a graph using networkx library. Then according to my similarity matrix, if the similarity rate between documents are less then 1 and greater than 0.1, I connect these 2 matrices in the graph. Then, I used my pagerank implementation and find 10 salient documents. After that, I repeat same process for sentences. According the 10 documents, I find abstract of the documents. For example, for the first question, I have 81 sentences. Again I create matrix which is 81x81. I filled it with using document similarity of the sentences. Using sentence similarity matrix, I create a graph and connect nodes if the similarity rate greater than 0.1. Lastly, again using pagerank method, I found highest 20 sentence pagerank scores. With this way, I create summary of the abstracts.

Question 1



Most Important documents	
Document ID	PageRank Scores
52kqp9yw	0.023974148508049326
0xhho1sh	0.02293894269438627
1mjaycee	0.022448367450978345
1qkwsh6a	0.021740492161381922
2inlyd0t	0.021273830683079597
89fol3pq	0.021248667328740246
zknmfgsh	0.02123149942756502
7v5aln90	0.020899239075815788
2ftw85xw	0.020435743455368367
juz9jnfk	0.019696240643639305



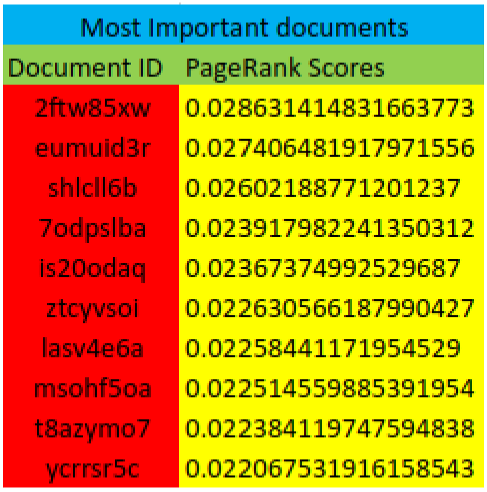
	Most	Important Sentences	
Sentence ID	PageRank Scores	Sentence ID	PageRank Scores
43	0.01547916013138991	64	0.01367599797812439
21	0.014879331495991166	51	0.013599627224994852
50	0.014879331495991166	18	0.013575371448684977
28	0.014807243149155377	55	0.013546775867442181
41	0.014518841952324262	45	0.01352551527605878
23	0.014374268378834229	77	0.01347448222164286
76	0.014317634161749157	63	0.013340269486324163
42	0.014183537101393914	39	0.013290011875634779
6	0.013862510754631015	46	0.013077024222652002
69	0.013703314834920054	10	0.013069067403477595

Summary for Question 1

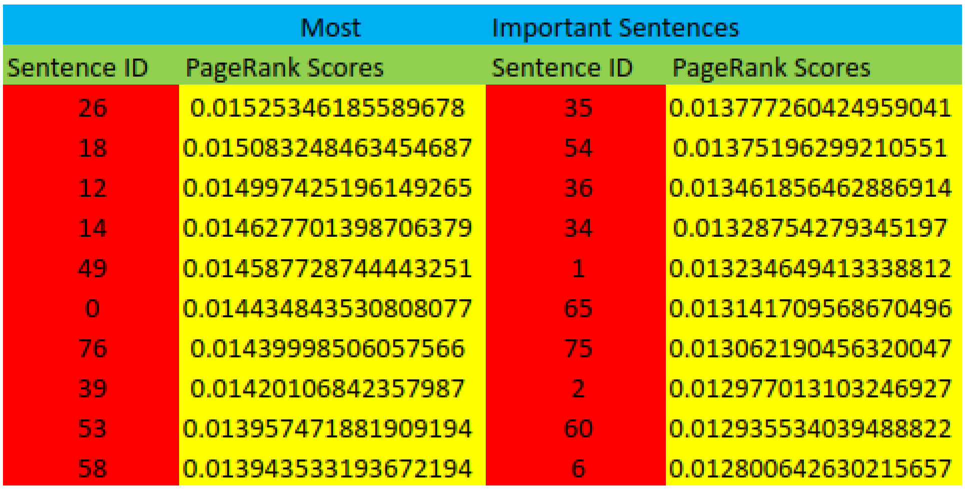
['Large surveillance of coronaviruses in pangolins could improve our understanding of the spectrum of coronaviruses in pangolins.', 'The outcome of SARS-CoV-2 infection is largely determined by virus-host interaction.', 'Currently, controlling infection to prevent the spread of SARS-CoV-2 is the primary intervention being used.', 'In the current review, we summarize and comparatively analyze the emergence and pathogenicity of COVID-19 infection and previous human coronaviruses severe acute respiratory syndrome coronavirus (SARS-CoV) and middle east respiratory syndrome coronavirus (MERS-CoV).', 'The molecular and phylogenetic analyses showed that pangolin Coronaviruses (pangolin-CoV) are genetically related to both the 2019-nCoV and bat Coronaviruses but do not support the 2019-nCoV arose directly from the pangolin-CoV.', 'The coronavirus disease 19 (COVID-19) is a highly transmittable and pathogenic viral infection caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), which emerged in Wuhan, China and spread around the world.', 'Results The genome of 2019-nCoV partially resembled SARS-CoV and MERS-CoV, and indicating a bat origin.', 'Our study also suggested that pangolin be natural host of Betacoronavirus, with a potential to infect humans.', 'In 2012, a novel human coronavirus, now called Middle East respiratory syndrome coronavirus (MERS-CoV), has emerged in the Middle East to cause fatal human infections in three continents.', 'A novel coronavirus, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), emerged in Wuhan, Hubei Province in

China in December 2019 and caused a serious type of pneumonia called coronavirus disease 2019 or COVID-19.', 'COVID-19 has become a global pandemic caused by a novel coronavirus SARS-CoV-2.', 'Here, we review the discovery, zoonotic origin, animal hosts, transmissibility and pathogenicity of SARS-CoV-2 in relation to its interplay with host antiviral defense.', 'Among patients with pneumonia caused by SARS-CoV-2 (novel coronavirus pneumonia or Wuhan pneumonia), fever was the most common symptom, followed by cough.', 'Abstract To investigate the evolutionary history of the recent outbreak of SARS-CoV-2 in China, a total of 70 genomes of virus strains from China and elsewhere with sampling dates between 24 December 2019 and 3 February 2020 were analyzed.', 'World Health Organization has declared the ongoing outbreak of coronavirus disease 2019 (COVID-19) a Public Health Emergency of International Concern.', 'The COVID-19 generally had a high reproductive number, a long incubation period, a short serial interval and a low case fatality rate (much higher in patients with comorbidities) than SARS and MERS.', 'Therefore, we concluded that the human SARS-CoV-2 virus, which is responsible for the recent outbreak of COVID-19, did not come directly from pangolins.', 'AbstractThe outbreak of 2019-nCoV pneumonia (COVID-19) in the city of Wuhan, China has resulted in more than 70,000 laboratory confirmed cases, and recent studies showed that 2019-nCoV (SARS-CoV-2) could be of bat origin but involve other potential intermediate hosts.', 'The virus was named severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) by the International Committee on Taxonomy of Viruses.', 'A multifaceted approach is necessary to control this evolving MERS-CoV outbreak.']

Question 13



Most Important documents	
Document ID	PageRank Scores
2ftw85xw	0.028631414831663773
eumuid3r	0.027406481917971556
shlcll6b	0.02602188771201237
7odpslba	0.023917982241350312
is20odaq	0.02367374992529687
ztcyvsoi	0.022630566187990427
lasv4e6a	0.02258441171954529
msohf5oa	0.022514559885391954
t8azymo7	0.022384119747594838
ycrrsr5c	0.022067531916158543



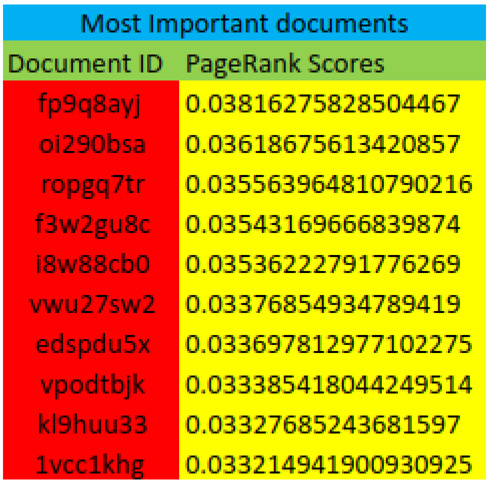
	Most	Important Sentences	
Sentence ID	PageRank Scores	Sentence ID	PageRank Scores
26	0.01525346185589678	35	0.013777260424959041
18	0.015083248463454687	54	0.01375196299210551
12	0.014997425196149265	36	0.013461856462886914
14	0.014627701398706379	34	0.01328754279345197
49	0.014587728744443251	1	0.013234649413338812
0	0.014434843530808077	65	0.013141709568670496
76	0.01439998506057566	75	0.013062190456320047
39	0.01420106842357987	2	0.012977013103246927
53	0.013957471881909194	60	0.012935534039488822
58	0.013943533193672194	6	0.012800642630215657

Summary for Question 13

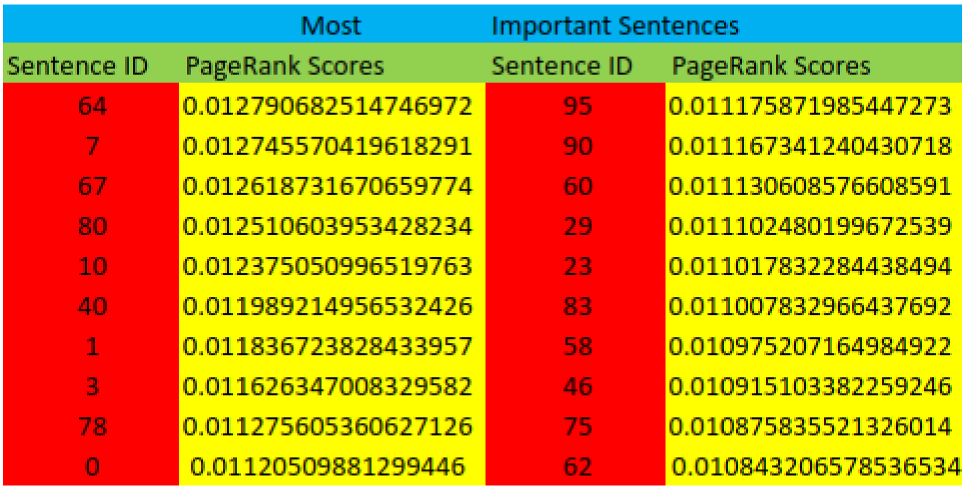
['53.42% of the patients tested positive in stool.', 'Routes of SARS-CoV-2 transmission are diversified and the main routes of transmission for COVID-19 are droplet transmission and close contact transmission.', 'The epidemic factors on the basis of knowledge of SARS-CoV-2 were discussed in this paper.', 'SARS-CoV-2 is a novel corona virus, the onset of COVID-19 is slow, and the pathogenesis of SARS-CoV-2 remains unclear and may lead to multiple organ damage.', 'The global impact of this new epidemic is yet uncertain.', 'A novel coronavirus, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), emerged in Wuhan, Hubei Province in

China in December 2019 and caused a serious type of pneumonia called coronavirus disease 2019 or COVID-19.', 'As of March 1, 2020, 79,968 patients in China and 7169 outside of China had tested positive for COVID19 and a mortality rate of 3.6% has been observed amongst Chinese patients.', 'The disease is transmitted by inhalation or contact with infected droplets and the incubation period ranges from 2 to 14 d. The symptoms are usually fever, cough, sore throat, breathlessness, fatigue, malaise among others.', 'The person-to-person transmission routes of 2019-nCoV included direct transmission, such as cough, sneeze, droplet inhalation transmission, and contact transmission, such as the contact with oral, nasal, and eye mucous membranes.', 'An acute respiratory disease, caused by a novel coronavirus (SARS-CoV-2, previously known as 2019-nCoV), the coronavirus disease 2019 (COVID-19) has spread throughout China and received worldwide attention.', 'Here, we summarize the known factors for the diverse transmission of MERS-CoV.', '2019-nCoV can also be transmitted through the saliva, and the fetalÃ¢â,¬â€œoral routes may also be a potential person-to-person transmission route.', 'There is a new public health crises threatening the world with the emergence and spread of 2019 novel coronavirus (2019-nCoV) or the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2).', 'The actual contributing factors to the spread of MERS-CoV are yet to be systematically studied, but data to date suggest viral, host and environmental factors play a major role.', 'This epidemic quickly spread across China and extended to more than 20 other countries.', 'The clinical symptoms of COVID-19 patients include fever, cough, fatigue and a small population of patients appeared gastrointestinal infection symptoms.', 'Coronavirus disease 2019 (COVID19) is an infectious disease caused by severe acute respiratory syndrome coronavirus 2 (SARSCoV2), it was first identified in 2019 in Wuhan, China and has resulted in the 2019-20 coronavirus pandemic.', 'This commentary discusses the reasons for the fast spread of SARS-CoV-2 in three aspects: the infectious sources, including the biological nature of the virus; the susceptible population; and the transmission routes.', 'The emergence of SARS-CoV-2, since the severe acute respiratory syndrome coronavirus (SARS-CoV) in 2002 and Middle East respiratory syndrome coronavirus (MERS-CoV) in 2012, marked the third introduction of a highly pathogenic and large-scale epidemic coronavirus into the human population in the twenty-first century.', 'In this study, we used rabbits to further characterize the transmission potential of MERS-CoV.']

Question 18



Most Important documents	
Document ID	PageRank Scores
fp9q8ayj	0.03816275828504467
oi290bsa	0.03618675613420857
ropgq7tr	0.035563964810790216
f3w2gu8c	0.03543169666839874
i8w88cb0	0.03536222791776269
vwu27sw2	0.03376854934789419
edspdu5x	0.033697812977102275
vpodtbjk	0.033385418044249514
kl9huu33	0.03327685243681597
1vcc1khg	0.033214941900930925



	Most	Important Sentences	
Sentence ID	PageRank Scores	Sentence ID	PageRank Scores
64	0.012790682514746972	95	0.011175871985447273
7	0.012745570419618291	90	0.011167341240430718
67	0.012618731670659774	60	0.011130608576608591
80	0.012510603953428234	29	0.011102480199672539
10	0.012375050996519763	23	0.011017832284438494
40	0.011989214956532426	83	0.011007832966437692
1	0.011836723828433957	58	0.010975207164984922
3	0.011626347008329582	46	0.010915103382259246
78	0.011275605360627126	75	0.010875835521326014
0	0.01120509881299446	62	0.010843206578536534

Summary for Question 18

['This literature review aims to presents accredited and the most current studies pertaining to the basic sciences of SARS-CoV-2, clinical presentation and disease course of COVID-19, public health interventions, and current epidemiological developments.', 'The molded N95 mask however tolerated only 1 cycle.', 'Additionally, the clinical and epidemiological differences between COVID-19 and other infections causing outbreaks (SARS, MERS, H1N1) are elucidated.', 'To alleviate this, many methods of N95 mask sterilization have been studied and proposed with the hope of being able to safely reuse masks.', 'In particular, single use disposable N95 face masks have been limited in supply.', 'Personal protective equipment (PPE), including surgical masks and N95 respirators, is crucially important to the safety of both patients and medical personnel, particularly in the event of infectious pandemics.', 'In particular, the supply of N95 respirator masks has become severely depleted with supplies having to be rationed and health care workers having to use masks for prolonged periods in many countries.', 'In addition, we sought to determine whether masks would tolerate repeated cycles of decontamination while maintaining structural and functional integrity.', 'The current COVID-19 pandemic has led to a dramatic shortage of masks and other personal protective equipment (PPE) in hospitals around the globe.', 'The response to the COVID19 epidemic is generating severe shortages of personal protective equipment around the world.', 'In particular, we examined the optimal deployment of face masks when resources are limited.', 'Even more importantly, they argue against using the qualitative fit test alone to assess mask integrity.', 'This suggests that reusable respirators are an acceptable alternative to N95 respirators in health care and offer 1 viable solution to prevent pandemic-generated respirator shortages.', 'Currently, controlling infection to prevent the spread of SARS-CoV-2 is the primary public healthcare intervention used.', 'The protective effect of wearing masks in Asia (OR = 0.31) appeared to be higher than that of Western countries (OR = 0.45).', 'Significant literature exists supporting the use of gamma radiation as a sterilization method, with viral inactivation of SARS-CoV reported at doses of at most 10 kGy, with other studies supporting 5 kGy for many types of viruses.', 'For all user groups, reusable respirators were significantly more likely (odds ratios 2.3-7.7) to be preferred over N95 filtering facepiece respirators in higher risk scenarios compared to Ã ¢â‚¬Å“usual circumstanceÃ¢â‚¬Â\x9d scenarios.', 'The primary obstacle to this approach is the possibility the UV radiation levels vary within BSCs.', 'Furthermore, the biochemistry of the major candidates for novel therapies is briefly reviewed and a summary of their current status in the clinical trials is presented.', 'There are multiple parameters of the clinical course and management of the COVID-19 that need optimization.']

How can you run my code ?

Just after downloading the necessary libraries, take a coffee and sit back. In the project folder terminal, write the “python3 pagerank.py” , then my program ask you to question number. Then write question number what you want and you will see page rank scores of the documents and sentences. Lastly, you will see summary for the question.

Notes

I run my code for all questions in the xml file and found 30 summaries I am adding these summaries between files. I just chose one of them randomly which is 18 and add to this report file.