CMPE 493

Introduction to Information Retrieval

Extractive Text Summarization for COVID-19

REPORT

Şadi Uysal

2015400162

a) Describe any assumptions or choices that you made while implementing your summarization system.

I assumed following libraries can be imported succesfully:

Please read README file for more information.

```
import nltk

from nltk.tokenize import sent_tokenize

import string

from nltk.tokenize import word_tokenize

import json

import os

import math

import re

import numpy as np

from scipy import spatial

import pandas   #used for only reading csv file
```

b) For each of the selected topics: Provide the IDs and the PageRank scores of the top 10 documents.

Topic 1: coronavirus origin

Top 10 DocIds for topic id : 1

['1mjaycee', '0xhho1sh', 'zknmfgsh', 'dnxhtbxn', 'k9yus2sv', '2inlyd0t', 'm6akijzn', '7v5aln90', 'hzuo6pwx', '52kqp9yw']

Top 10 Document's corresponding PageRanks for topic : 1
[0.02343556694065839, 0.024124688204806465, 0.02401134242468868, 0.025202598164612136, 0.025875923666231676, 0.027545294365254387, 0.030009908773439523, 0.03078271017698758, 0.031110903130130912, 0.04134307586960971]

Topic 13: how does coronavirus spread

Top 10 DocIds for topic id : 13
['vk4lt83x', 'vpcx2t3w', 'yr1dq258', 'msohf5oa', 'lasv4e6a', 'xuczplaf', 'eumuid3r', 'o9z6rdim', 'awp1nck0', 'xs7vkm19']

Top 10 Document's corresponding PageRanks for topic : 13
[0.020408163265306006, 0.02040816326530601, 0.02040816326530601, 0.025085029118579946, 0.022450257013543348, 0.020426774180595735, 0.02164163778139707, 0.025640944046552174, 0.03362366479073527, 0.02611301966242625]

Topic 17: coronavirus clinical trials

Top 10 DocIds for topic id : 17

['g8h6cpzr', 't1wpujpm', 'g0iwfpkg', 'vxqdfiel', 'o8j17zzs', 'nrnc8u28', 'jc9ugexn', '4as055wh', 'iwsa760n', 'yth3t2cf']

Top 10 Document's corresponding PageRanks for topic : 17
[0.023863323661622672, 0.047770171146285255, 0.033897476755093525, 0.03200293847941042, 0.025589258725214004, 0.02979096319002936, 0.029666506712363764, 0.026680852511709208, 0.044291548570459714, 0.037741225367560007]

c)For each of the selected three topics: Provide the top 20 sentences as well as their PageRank scores.

Topic 1: coronavirus origin

Top 20 Sentence for topic : 1

['To explore the potential intermediate animal host of the SARS-CoV-2 virus, we re-analyzed virome datasets from pangolins and representative SARS-related coronaviruses isolates from bats, with particular attention paid to the spike glycoprotein gene.', 'Among patients with pneumonia caused by SARS-CoV-2 (novel coronavirus pneumonia or Wuhan pneumonia), fever was the most common symptom, followed by cough.', 'Source identification requires detailed epidemiological studies of the infected patients and enhanced surveillance of MERS-CoV or similar coronaviruses in humans and animals.', 'We show evidence of strong purifying selection around the receptor binding motif (RBM) in the spike gene and in other genes among bat, pangolin and human coronaviruses, indicating similar strong evolutionary constraints in different host species.', 'Moreover, the presence of SARSr-CoV ORF7a-like protein in Rs-BatCoV HKU32 suggests a common evolutionary origin of this accessory protein with SARS-CoV, also from Chinese horseshoe bats, an apparent reservoir for coronavirus epidemics.', 'Genome analyses showed that Rs-BatCoV HKU32 is closely related to BatCoV HKU10 and related viruses from diverse bat families, whereas Tr-BatCoV HKU33 is closely related to BtNv-AlphaCoV and similar viruses exclusively from bats of Vespertilionidae family.', '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.', 'Two novel alphacoronaviruses, Rhinolophus sinicus bat coronavirus HKU32 (Rs-BatCoV HKU32) and Tylonycteris robustula bat coronavirus HKU33 (Tr-BatCoV HKU33), were discovered from Chinese horseshoe bats in Hong Kong and greater bamboo bats in Guizhou Province, respectively.', 'SUMMARYA novel coronavirus (nCoV-2019) was the cause of an outbreak of respiratory illness detected in Wuhan, Hubei Province, China in December of 2019.', 'Firstly, coronaviruses, in addition to influenza viruses, can cause severe and rapidly spreading human infections.', 'While bats are increasingly recognized as a source of coronavirus epidemics, the diversity and emergence potential of bat coronaviruses remains to be fully understood.', 'Since then, researchers around the world, especially those in Asia where SARS-CoV was first identified, have turned their focus to find novel coronaviruses infecting humans, bats, and other animals.', 'Two human coronaviruses, HCoV-HKU1 and HCoV-NL63, were identified shortly after the SARS-CoV epidemic as common causes of human respiratory tract infections.', 'Similar purifying selection in different host species and frequent recombination among coronaviruses suggest a common evolutionary mechanism that could lead to new emerging human coronaviruses.', 'Genomic analysis revealed that SARS-CoV-2 is phylogenetically related to severe acute respiratory syndrome-like (SARS-like) bat viruses, therefore bats could be the possible primary reservoir.', 'Both viruses also have close relationships with bat coronaviruses.', 'COVID-19 has become a global pandemic caused by a novel coronavirus SARS-CoV-2.', 'Large surveillance of coronaviruses in pangolins could improve our understanding of the spectrum of coronaviruses in pangolins.', '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.', 'Among 1779 bat samples collected in China, diverse coronaviruses were detected in 32 samples from five different bat species by RT-PCR.']

Top 20 Sentence's corresponding PageRanks for topic : 1
[0.014161423515978913, 0.014462073797071748, 0.014171740974289096, 0.014538678547431984, 0.014683985803946973, 0.016934995770836247, 0.015056852505016819, 0.014722690286482636, 0.018948276551109602, 0.016390913854874633, 0.026079625059759814, 0.018437957488013055, 0.015015823798952884, 0.019769522374615812, 0.01640373677520713, 0.019749296230893013, 0.015454211046261905, 0.014989005299196292, 0.015534370553782182, 0.01675372256711865]

Topic 13: how does coronavirus spread

Top 20 Sentence for topic : 13

['In this review, we highlights the symptoms, epidemiology, transmission, pathogenesis, phylogenetic analysis and future directions to control the spread of this fatal disease.', 'For MHV, log10 reduction factors were 3.9 for 70% ethanol, 1.3 for phenolic, 1.7 for OPA, 0.62 for 1:100 hypochlorite, 2.7 for 62% ethanol, and 2.0 for 71% ethanol.', 'Results After 1-minute contact time, for TGEV, there was a log10 reduction factor of 3.2 for 70% ethanol, 2.0 for phenolic, 2.3 for OPA, 0.35 for 1:100 hypochlorite, 4.0 for 62% ethanol, and 3.5 for 71% ethanol.', 'It is commonly recognized that droplet transmission is the main route.', 'Here, we focus on the potential transmission routes that have been investigated in the SARSÃ¢â‚¬Â\x90CoVÃ¢â‚¬Â\x902 epidemic recently.', 'Conclusion Only ethanol reduced infectivity of the 2 coronaviruses by >3-log10 after 1 minute.', 'However, unlike MERS-CoV-infected dromedaries, these rabbits did not develop clinical manifestations including nasal discharge and did shed only limited amounts of infectious virus from the nose.', 'Epidemiological experts, as well as the WHO, consider more evidence is needed to confirm.3 Besides, there are other routes except respiratory transmission.', 'AbstractSince December 2019, a newly identified coronavirus (2019 novel coronavirus, 2019-nCov) is causing outbreak of pneumonia in one of largest cities, Wuhan, in Hubei province of China and has draw significant public health attention.', 'Disease spread through both direct (droplet and person-to-person) as well as indirect contact (contaminated objects and airborne transmission) are indicated, supporting the use of airborne isolation precautions.', 'At present, the origin, susceptible population, and infection sources already have been clear.1, 2 However, the transmission routes, a key step to the epidemic control, have not yet been fully ascertained.', 'The routes of transmission are direct contact, and droplet and possible aerosol transmissions.', 'Consistently, no transmission by contact or airborne routes was observed in rabbits.', 'Abstract Coronavirus disease (COVID-19) is caused by SARS-COV2 and represents the causative agent of a potentially fatal disease that is of great global public health concern.', 'Here we recommend the infection control measures during dental practice to block the person-to-person transmission routes in dental clinics and hospitals.', 'Lack of evidence on SARS-CoV-2 transmission dynamics has led to shifting isolation guidelines between airborne and droplet isolation precautions.', '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.', '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.', 'These results indicated along with respiratory systems, digestive system is a potential routes for 2019-nCov infection.', '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.']

Top 20 Sentence's corresponding PageRanks for topic : 13

[0.01585193985677509, 0.015874193539261568, 0.015874193539261568, 0.01588280073067776, 0.0159216230403013, 0.01652950218085588, 0.017617966637861977, 0.0159216230403013, 0.016503965355799483, 0.018360077812796624, 0.018871780049734162, 0.023390077367057147, 0.02636982624896766, 0.020548683965643154, 0.023242414725256625, 0.019055230608871133, 0.02265838257853824, 0.023526263360599913, 0.02053706423553472, 0.022632634275593297]

Topic 17: coronavirus clinical trials

Top 20 Sentence for topic : 17

['DATA COLLECTION AND ANALYSIS: Two review authors independently screened and selected trials, assessed risk of bias and extracted data.', 'But, there is still a lack of systematic review of registered clinical trials.', 'Conclusions: Disorderly and intensive clinical trials of COVID-19 using traditional Chinese medicine and western medicine are ongoing or will being carried out in China.', 'Results: 97 eligible study protocols were identified from 160 clinical trials.', 'Compared with that of during SARS period in 2003, China have the stronger capability to carry out clinical trials of new drugs in emergency period.', 'Conclusion: A COS for COVID-19 may improve consistency of outcome reporting in clinical trials, which may help identify valued interventions after comparing different trials when the researchers report the same outcomes.', 'Only 11 trials have begun to recruit patients, and none of the registered clinical trials had been completed; 34 trials were early clinical exploratory trials or in a pre-experiment stage, 15 trials belonged to phrase Ã¢â€¦Â¢ and 4 trials were phrase Ã¢â€¦Â£.', 'Background: There are a large number of clinical trials for COVID-19.', 'None of the trials reported on infection during ventilation or quality of life after discharge.', 'Therefore, we conducted a systematic review of the clinical trials of COVID-19 to summarize the characteristics of the COVID-19 registered clinical trials.', 'Findings: Out of the 353 studies identified, 115 clinical trials were selected for data extraction.', 'The median sample size of the trials was 100 (IQR: 60 - 200), and the median execute time of the trials was 179 d (IQR: 94 - 366 d).', 'The methods of intervention included traditional Chinese medicine involving 26 trials, Western medicine involving 30 trials, and integrated traditional Chinese medicine and Western medicine involving 19 trials.', '76 outcomes were identified from TCM clinical trials, 126 outcomes were identified from Western medicine clinical trials.', 'However, 62 trials (54%) did not describe the phase of the study.', 'In the end, a COS was developed for clinical trials of TCM and Western medicine was developed.', 'Strategies to facilitate future clinical trials during outbreaks of unknown or novel pathogens are also presented.', 'Interpretation: Numerous clinical trials have been registered since the beginning of the COVID-19 outbreak, however, a number of information regarding drugs or trial design were lacking.', 'Despite this, no controlled clinical trials assessing the efficacy of these agents were conducted.', 'It is emergency to develop a core outcome set (COS) for clinical trials.']

Top 20 Sentence's corresponding PageRanks for topic : 17

[0.013813188750978757, 0.013868586908000091, 0.014341866036315353, 0.014550953033118463, 0.013906426569692544, 0.016507815623507848, 0.017530676834909015, 0.021377151375200093, 0.018189483258878835, 0.017530676834909015, 0.018649961511444493, 0.017409675050399594, 0.017089219681905548, 0.022716444191158643, 0.017837648210052282, 0.021133763121048726, 0.017627256755757724, 0.019091588346600716, 0.017502684040608556, 0.017035072421603285]