In [2]:
```
import pandas as pd
import numpy as np
import json
import math
import glob
import nltk
```

In [3]:
```
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.feature_extraction.text import TfidfTransformer
from nltk.stem.snowball import SnowballStemmer
from scipy.spatial import distance
from matplotlib import pyplot as plt
from nltk.tokenize import PunktSentenceTokenizer,sent_tokenize, word_tokenize
```

In [4]:
```
df = pd.read_csv('C:\\Users\\asus\\Covid Data\\metadata.csv')
doc_paths = 'C:\\Users\\asus\\Covid Data\\pdf_json.json'
df.sha.fillna("", inplace=True)

#get text for articles that are available
```

```
C:\Users\asus\anaconda3\lib\site-packages\IPython\core\interactiveshell.py:
3165: DtypeWarning: Columns (1,4,5,6,13,14,15,16) have mixed types.Specify
dtype option on import or set low_memory=False.
  has_raised = await self.run_ast_nodes(code_ast.body, cell_name,
```

In [5]:
```
df.head()
```

Out[5]:

cord_uid	sha	source_x	title	doi	
ug7v899j	d1aafb70c066a2068b02786f8929fd9c900897fb	PMC	Clinical features of culture-proven Mycoplasma...	10.1186/1471-2334-1-6	F
02tnwd4m	6b0567729c2143a66d737eb0a2f63f2dce2e5a7d	PMC	Nitric oxide: a pro-inflammatory mediator in l...	10.1186/rr14	F
ejv2xln0	06ced00a5fc04215949aa72528f2eeaae1d58927	PMC	Surfactant protein-D and pulmonary host defense	10.1186/rr19	F
2b73a28n	348055649b6b8cf2b9a376498df9bf41f7123605	PMC	Role of endothelin-1 in lung disease	10.1186/rr44	F
9785vg6d	5f48792a5fa08bed9f56016f4981ae2ca6031b32	PMC	Gene expression in epithelial cells in respons...	10.1186/rr61	F

In [6]:
```
df.shape
```

Out[6]:
```
(522159, 19)
```

In [7]:
```
def get_text(sha):
    if sha == "":
        return ""
    document_path = [x for x in doc_paths if sha in x]
    if not document_path:
        return ""
    with open(document_path[0]) as f:
        file = json.load(f)
        full_text = []
        #iterate over abstract and body part
        for part in ['abstract', 'body_text']:
            # iterate over each paragraph
            for text_part in file[part]:
                text = text_part['text']
                # remove citations from each paragraph
                for citation in text_part['cite_spans']:
                    text = text.replace(citation['text'], "")
                full_text.append(text)

        return str.join(' ', full_text)
```

In [8]:
```
%time df['text'] = df.apply(lambda x: get_text(x.sha), axis=1)
```

```
Wall time: 8.99 s
```

In [9]:
```
from nltk.stem import WordNetLemmatizer
analyzer = CountVectorizer().build_analyzer()
lemmatizer = WordNetLemmatizer()
def preprocess1(doc):
    doc=doc.lower()
    return str.join(" ", [lemmatizer.lemmatize(w) for w in analyzer(doc)])

def preprocess_row1(row):
    text = str.join(' ', [str(row.title), str(row.abstract), str(row.text)])
    return preprocess1(text)
```

In [10]:
```
%time df['lempreprocessed'] = df.apply(lambda x: preprocess_row1(x), axis=1)
```

```
Wall time: 9min 34s
```

In [11]:
```
cv = CountVectorizer(max_df=0.95, stop_words='english')
%time word_count = cv.fit_transform(df.lempreprocessed)
tfidf_tr = TfidfTransformer(smooth_idf=True, use_idf=True)
%time tfidf_tr.fit(word_count)
```

```
Wall time: 1min 35s
Wall time: 272 ms
```

Out[11]:
```
TfidfTransformer()
```

In [12]:
```
def get_word_vector(document):
    w_vector = tfidf_tr.transform(cv.transform([document]))
    return w_vector
```

sample_tfidfvectorizer= get_word_vector(df.iloc[0].lempreprocessed) df1 = pd.DataFrame(sample_tfidfvectorizer.T.todense(), index=cv.get_feature_names(), columns= ["tfidf"]) df1.sort_values(by=["tfidf"],ascending=False)

In [13]:
```
%time df['word_vector'] = df.lempreprocessed.apply(get_word_vector)
```

```
Wall time: 1h 2min 21s
```

feature_names = cv.get_feature_names() def get_words_with_value(w_vector): return sorted([(feature_names[ind], val) for ind, val in zip(w_vector.indices, w_vector.data)], key=lambda x: x[1], reverse=True)

In [14]:
```
def calculate_distance_between_words_vectors(search_words_indices, search_vec
    document_vec = document_vector[0, search_words_indices].toarray()
    return distance.euclidean([search_vec], document_vec)
```

def get_rel_doc search_vector = get_word_vector(preprocess(topic)) search_words_indices = search_vector.indices search_vec = search_vector.data distance_idx = df.apply(lambda x: calculate_distance_between_words_vectors(search_words_indices, search_vec, x.word_vector), axis=1) relevant_indexes = distance_idx.sort_values().head(10).index result_columns = ["title", "doi", "pmcid", "license", "authors"] result = df[result_columns].iloc[relevant_indexes].fillna("") return result

In [15]:
```
def display_friendly_results(df_result):
    display_columns = [
 'title',
 'doi',
 'abstract','authors',
 'lempreprocessed',
 ]
    display(df_result[display_columns].reset_index(drop=True))
```

In [27]:

```
def get_related_documents(text, number_of_documents):
    search_vector = get_word_vector(preprocess1(text))
    search_words_indices = search_vector.indices
    search_vec = search_vector[0, search_words_indices].toarray()
    distance_idx = df.apply(lambda x: calculate_distance_between_words_vector
    relevant_indexes = distance_idx.sort_values().head(number_of_documents).i
    result_columns = [
 'title',
 'doi',
 'abstract','authors',
 'lempreprocessed',
]
    result = df[result_columns].iloc[relevant_indexes].fillna("")
    return result
```

In [23]:

```
topics = {
    "What is known about transmission, incubation, and environmental stabilit
    [
        "Range of incubation periods for the disease in humans (and how this
        "Prevalence of asymptomatic shedding and transmission (e.g., particul

    ],
    "What do we know about COVID-19 risk factors?":
    [
        "Data on potential risks factors",

        "Transmission dynamics of the virus, including the basic reproductive

    ],
    "What do we know about virus genetics, origin, and evolution?":
    [
        "Real-time tracking of whole genomes and a mechanism for coordinating
        "Access to geographic and temporal diverse sample sets to understand
        "Evidence that livestock could be infected (e.g., field surveillance,
        "Evidence of whether farmers are infected, and whether farmers could
        "Surveillance of mixed wildlife- livestock farms for SARS-CoV-2 and o
        "Experimental infections to test host range for this pathogen.",
        "Animal host(s) and any evidence of continued spill-over to humans",
        "Socioeconomic and behavioral risk factors for this spill-over",
        "Sustainable risk reduction strategies"
    ],
    "What do we know about vaccines and therapeutics?":
    [
        "Effectiveness of drugs being developed and tried to treat COVID-19 p
        "Clinical and bench trials to investigate less common viral inhibitor
        "Methods evaluating potential complication of Antibody-Dependent Enha
        "Exploration of use of best animal models and their predictive value
        "Capabilities to discover a therapeutic (not vaccine) for the disease
        "Alternative models to aid decision makers in determining how to prio
        "Efforts targeted at a universal coronavirus vaccine.",
        "Efforts to develop animal models and standardize challenge studies",
        "Efforts to develop prophylaxis clinical studies and prioritize in he
        "Approaches to evaluate risk for enhanced disease after vaccination",
        "Assays to evaluate vaccine immune response and process development f
    ],
    "What do we know about diagnostics and surveillance?":
    [
        "How widespread current exposure is to be able to make immediate poli
        "Efforts to increase capacity on existing diagnostic platforms and ta
        "Recruitment, support, and coordination of local expertise and capaci
        "National guidance and guidelines about best practices to states (e.g
        "Development of a point-of-care test (like a rapid influenza test) an
        "Rapid design and execution of targeted surveillance experiments call
        "Separation of assay development issues from instruments, and the rol
        "Efforts to track the evolution of the virus (i.e., genetic drift or
        "Latency issues and when there is sufficient viral load to detect the
        "Use of diagnostics such as host response markers (e.g., cytokines) t
        "Policies and protocols for screening and testing.",
        "Policies to mitigate the effects on supplies associated with mass te
        "Technology roadmap for diagnostics.",
        "Barriers to developing and scaling up new diagnostic tests (e.g., ma
```

```
        "New platforms and technology (e.g., CRISPR) to improve response time
        "Coupling genomics and diagnostic testing on a large scale.",
        "Enhance capabilities for rapid sequencing and bioinformatics to targ
        "Enhance capacity (people, technology, data) for sequencing with adva
        "One Health surveillance of humans and potential sources of future sp
    ],
    "What do we know about non-pharmaceutical interventions?":
    [
        "Guidance on ways to scale up NPIs in a more coordinated way (e.g., e
        "Rapid design and execution of experiments to examine and compare NPI
        "Rapid assessment of the likely efficacy of school closures, travel b
        "Methods to control the spread in communities, barriers to compliance
        "Models of potential interventions to predict costs and benefits that
        "Policy changes necessary to enable the compliance of individuals wit
        "Research on why people fail to comply with public health advice, eve
        "Research on the economic impact of this or any pandemic. This would
    ],
    "What has been published about medical care?":
    [
        "Resources to support skilled nursing facilities and long term care f
        "Mobilization of surge medical staff to address shortages in overwhel
        "Age-adjusted mortality data for Acute Respiratory Distress Syndrome
        "Extracorporeal membrane oxygenation (ECMO) outcomes data of COVID-19
        "Outcomes data for COVID-19 after mechanical ventilation adjusted for
        "Knowledge of the frequency, manifestations, and course of extrapulmo
        "Application of regulatory standards (e.g., EUA, CLIA) and ability to
        "Approaches for encouraging and facilitating the production of elasto
        "Best telemedicine practices, barriers and faciitators, and specific
        "Guidance on the simple things people can do at home to take care of
        "Oral medications that might potentially work.",
        "Use of AI in real-time health care delivery to evaluate intervention
        "Best practices and critical challenges and innovative solutions and
        "Efforts to define the natural history of disease to inform clinical
        "Efforts to develop a core clinical outcome set to maximize usability
        "Efforts to determine adjunctive and supportive interventions that ca
    ],
    "What has been published about information sharing and inter-sectoral col
    [
        "Methods for coordinating data-gathering with standardized nomenclatu
        "Sharing response information among planners, providers, and others."
        "Understanding and mitigating barriers to information-sharing.",
        "How to recruit, support, and coordinate local (non-Federal) expertis
        "Integration of federal/state/local public health surveillance system
        "Value of investments in baseline public health response infrastructu
        "Modes of communicating with target high-risk populations (elderly, h
        "Risk communication and guidelines that are easy to understand and fo
        "Communication that indicates potential risk of disease to all popula
        "Misunderstanding around containment and mitigation.",
        "Action plan to mitigate gaps and problems of inequity in the Nation’
        "Measures to reach marginalized and disadvantaged populations.",
        "Data systems and research priorities and agendas incorporate attenti
        "Mitigating threats to incarcerated people from COVID-19, assuring ac
        "Understanding coverage policies (barriers and opportunities) related
    ],
    "What has been published about ethical and social science considerations?
    [
        "Efforts to articulate and translate existing ethical principles and
```

```
        "Efforts to embed ethics across all thematic areas, engage with novel
        "Efforts to support sustained education, access, and capacity buildin
        "Efforts to establish a team at WHO that will be integrated within mu
        "Efforts to develop qualitative assessment frameworks to systematical
        "Efforts to identify how the burden of responding to the outbreak and
        "Efforts to identify the underlying drivers of fear, anxiety and stig
    ]
}
```

In [24]:
```
l=[]
```

In [25]:
```
def display_topics_results(question):
    for topic in topics[question]:
        res=get_related_documents(topic,6)
        for i in res.index:
            l.append(res['lempreprocessed'][i])
        display_friendly_results(res)
```

In [29]:
```
%time display_topics_results("What is known about transmission, incubation, a
```

0	urgent urological surgery during th...		Since first reported in December 2019, the nov...	Louise; Yang, Bob; Abdelmotagly, Yeh...	delivering urgent urological surgery during th...
1	Delivering urgent urological surgery during th...	10.1111/bju.15110	Since first reported in December 2019, the nov...	Paramore, Louise; Yang, Bob; Abdelmotagly, Yeh...	delivering urgent urological surgery during th...
2	Association between the Severity of Influenza ...	10.1371/journal.pone.0148506	BACKGROUND: In early 2013, a novel avian-origi...	Virlogeux, Victor; Yang, Juan; Fang, Vicky J.;...	association between the severity of influenza ...
3	The new coronavirus-cOvid-19 in Uzbekistan		The article includes the latest researches on ...	Matnazarova, Gulbahor; Mirtazaev, Omonturdi; B...	the new coronavirus covid 19 in uzbekistan the...
4	Pre-symptomatic		We used contact tracing to	Kong, Dechuan;	pre symptomatic

In [32]:
```
from transformers import BertForQuestionAnswering
from transformers import BertTokenizer
import torch
```

In [35]:
```
model = BertForQuestionAnswering.from_pretrained('bert-large-uncased-whole-wo

#Tokenizer
tokenizer = BertTokenizer.from_pretrained('bert-large-uncased-whole-word-mask
```

In [36]:
```
question = "What is known about transmission, incubation, and environmental s
```

In [37]:
```
def info(abstract,question):
    paragraph = abstract
    encoding = tokenizer.encode_plus(text=question,text_pair=paragraph, add_s
    inputs = encoding['input_ids']  #Token embeddings
    sentence_embedding = encoding['token_type_ids']  #Segment embeddings
    tokens = tokenizer.convert_ids_to_tokens(inputs) #input tokens
    start_scores, end_scores = model(input_ids=torch.tensor([inputs]), token_
    start_index = torch.argmax(start_scores)
    end_index = torch.argmax(end_scores)
    if end_index>start_index:
        answer = tokens[start_index:end_index]

    else:
        answer = tokens[start_index:]
    return answer
```

In [39]:
```
a=[]
```

In [40]:
```
for i in list(set(l)):
    print(len(i))
    answer=info(i,question)
    a.append(answer)
```

```
893

Keyword arguments {'add_special': True} not recognized.
Keyword arguments {'add_special': True} not recognized.
Keyword arguments {'add_special': True} not recognized.
Keyword arguments {'add_special': True} not recognized.

424

Keyword arguments {'add_special': True} not recognized.
Keyword arguments {'add_special': True} not recognized.

1379

Keyword arguments {'add_special': True} not recognized.
Keyword arguments {'add_special': True} not recognized.

669

Keyword arguments {'add_special': True} not recognized.
Keyword arguments {'add_special': True} not recognized.

1260

Keyword arguments {'add_special': True} not recognized.
Keyword arguments {'add_special': True} not recognized.

473

Keyword arguments {'add_special': True} not recognized.
Keyword arguments {'add_special': True} not recognized.

493

Keyword arguments {'add_special': True} not recognized.
Keyword arguments {'add_special': True} not recognized.

1170
```

In [41]:
```
for i in a:
    corrected_answer = ''
    for word in i:
        if word[0:2] == '##':
            corrected_answer += word[2:]
        else:
            corrected_answer += ' ' + word
    print(len(corrected_answer),corrected_answer)
```

```
82  [CLS] what is known about transmission , incubation , and environmental
stability
35  risk of transmission to healthcare
1474  [CLS] what is known about transmission , incubation , and environment
al stability [SEP] association between the severity of influenza h7n9 virus
infection and length of the incubation period background in early 2013 nove
l avian origin influenza h7n9 virus emerged in china and ha caused sporadic
human infection the incubation period is the delay from infection until ons
et of symptom and varies from person to person few previous study have exam
ined whether the duration of the incubation period correlate with subsequen
t disease severity method and finding we analyzed data of period of exposur
e on 395 human case of laboratory confirmed influenza h7n9 virus infection
in china in bayesian framework using weibull distribution we found longer i
ncubation period for the 173 fatal case with mean of day 95 credibility int
erval cri compared to mean of day 95 cri for the 222 non fatal case and the
difference in mean wa marginally significant at 47 day 95 cri 04 99 there w
a statistically significant correlation between longer incubation period an
d an increased risk of death after adjustment for age sex geographical loca
tion and underlying medical condition adjusted odds ratio 70 per day increa
se in incubation period 95 credibility interval 47 97 conclusion we found s
ignificant association between longer incubation period and greater risk of
death among human h7n9 case the underlying biological mechanism leading to
this association deserve further exploration [SEP]
63  pre symptomatic transmission of novel coronavirus in community
71  incubation period last from to 14 day an infected person can be contag
36  viral rna shedding and transmission
49  viral ribonucleic acid shedding and transmission
80  current evidence show an incubation period of up to 14 day post exposur
e to the
```