Topic_Modelling

October 14, 2020

[1]:
```
from IPython.utils import io
with io.capture_output() as captured:
    !pip install scispacy
    !pip install https://s3-us-west-2.amazonaws.com/ai2-s2-scispacy/releases/v0.
 ↪2.4/en_core_sci_lg-0.2.4.tar.gz
```

[1]:
```
import numpy as np
import pandas as pd

from sklearn.feature_extraction import text
from sklearn.feature_extraction.text import CountVectorizer, TfidfVectorizer
from sklearn.decomposition import LatentDirichletAllocation

import scispacy
import spacy
import en_core_sci_lg

from scipy.spatial.distance import jensenshannon

import joblib

from IPython.display import HTML, display

from ipywidgets import interact, Layout, HBox, VBox, Box
import ipywidgets as widgets
from IPython.display import clear_output

from tqdm import tqdm
from os.path import isfile

import seaborn as sns
import matplotlib.pyplot as plt
plt.style.use("dark_background")
```

[2]:
```
df = pd.read_csv('../covid_data/Data/cord19_df/cord19_df.csv')
df.head()
```

```
C:\ProgramData\Anaconda3\lib\site-
```

```
packages\IPython\core\interactiveshell.py:3063: DtypeWarning: Columns (11) have
mixed types.Specify dtype option on import or set low_memory=False.
  interactivity=interactivity, compiler=compiler, result=result)
```

[2]:

	paper_id	body_text	methods	results	source	title	doi	abstract	publish_time
0	3cdc48bb9e40afd30a59463b7872761a726998c8	NDV (Roakin strain) was obtained from Dr. D. J…	NDV (Roakin strain) was obtained from Dr. D. J…	Adult house flies harbored Newcastle Disease v…	PMC	Experimental Evaluation of Musca domestica (Di…	10.1093/jmedent/44.4.666		
1	d99acb4e99be7852aa61a688c9fbd38d44b5a252	Live attenuated viruses have been developed an…	RSV A2 strain was obtained from ATCC (Manassas…	The reverse genetics system for measles Edmons…	PMC	Evaluation of Measles Vaccine Virus as a Vecto…	10.2174/1874357901206010012		
2	748d4c57fe1acc8d9d97cf574f7dea5296f9386c	Ebola virus (EBOV) and other members of the fa…	U2OS human osteosarcoma cells were cultured in…	For evaluating EBOV GP triggering under biosaf…	PMC	Direct Visualization of Ebola Virus Fusion Tri…	10.1128/mbio.01857-15		
3	b891efc6e1419713b05ff7d89b26d260478c28df	To the Editor:\nChina has the world's second l…	NaN	NaN	PMC	Tuberculosis prevention in healthcare workers …	10.1183/23120541.00015-2015		
4	353852971069ad5794445e5c1ab6077ce23da75d	Coronavirus disease 2019 (COVID-19) has spread…	NaN	NaN	NaN	NaN	NaN		

```
0  House flies, Musca domestica L. (Diptera: Musc…   2007-07-01
1  Live attenuated recombinant measles vaccine vi…   2012-02-16
2  Ebola virus (EBOV) makes extensive and intrica…   2016-02-09
3  BSL3 and respiratory isolation wards protect h…   2015-08-21
4                                                NaN          NaN

                                             authors        journal arxiv_id  \
0  Watson, D. Wes; Niño, Elina L.; Rochon, Katery…  J Med Entomol      NaN
1  Mok, Hoyin; Cheng, Xing; Xu, Qi; Zengel, James…   Open Virol J      NaN
2  Spence, Jennifer S.; Krause, Tyler B.; Mittler…           mBio      NaN
3  Deng, Yunfeng; Li, Yan; Wang, Fengtian; Gao, D…   ERJ Open Res      NaN
4                                                NaN            NaN      NaN

                                                 url  publish_year  \
0  https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7…          2007
1  https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3…          2012
2  https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4…          2016
3  https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5…          2015
4                                                NaN           -1

   is_covid19                       study_design
0       False                                 []
1       False  ['truncated', 'gamma', 'protocol']
2       False      ['truncated', 'heterogeneity']
3       False                                 []
4        True                                 []
```

[3]:
```
all_texts = df.body_text
print(all_texts[0][:500])
```

```
NDV (Roakin strain) was obtained from Dr. D. J. King, Southeast Poultry Research
Laboratory, Athens, GA. The virus was propagated by inoculation of 10-d-old
embryonated chicken eggs by the allantoic route, 0.1 ml per egg (SPAFAS, Charles
River Laboratories Inc., Wilmington, MA). Allantoic fluid was harvested from
eggs that died 3-4 d postinoculation. Titration of the virus was accomplished by
preparation of 10-fold dilutions of allantoic fluid in Dulbecco's minimal
essential medium (DMEM) and in
```

[4]:
```
nlp = en_core_sci_lg.load(disable=['tagger','parser','ner'])
nlp.max_length = 3000000
```

[5]:
```
# pip install spacy==2.2.4
```

```
Requirement already satisfied: spacy==2.2.4 in
c:\programdata\anaconda3\lib\site-packages (2.2.4)
Requirement already satisfied: thinc==7.4.0 in
c:\programdata\anaconda3\lib\site-packages (from spacy==2.2.4) (7.4.0)
```

```
Requirement already satisfied: catalogue<1.1.0,>=0.0.7 in
c:\programdata\anaconda3\lib\site-packages (from spacy==2.2.4) (1.0.0)
Requirement already satisfied: numpy>=1.15.0 in
c:\programdata\anaconda3\lib\site-packages (from spacy==2.2.4) (1.18.1)
Requirement already satisfied: tqdm<5.0.0,>=4.38.0 in
c:\programdata\anaconda3\lib\site-packages (from spacy==2.2.4) (4.42.0)
Requirement already satisfied: murmurhash<1.1.0,>=0.28.0 in
c:\programdata\anaconda3\lib\site-packages (from spacy==2.2.4) (1.0.2)
Requirement already satisfied: srsly<1.1.0,>=1.0.2 in
c:\programdata\anaconda3\lib\site-packages (from spacy==2.2.4) (1.0.2)
Requirement already satisfied: plac<1.2.0,>=0.9.6 in
c:\programdata\anaconda3\lib\site-packages (from spacy==2.2.4) (1.1.3)
Requirement already satisfied: preshed<3.1.0,>=3.0.2 in
c:\programdata\anaconda3\lib\site-packages (from spacy==2.2.4) (3.0.2)
Requirement already satisfied: cymem<2.1.0,>=2.0.2 in
c:\programdata\anaconda3\lib\site-packages (from spacy==2.2.4) (2.0.3)
Requirement already satisfied: wasabi<1.1.0,>=0.4.0 in
c:\programdata\anaconda3\lib\site-packages (from spacy==2.2.4) (0.8.0)
Requirement already satisfied: setuptools in c:\programdata\anaconda3\lib\site-
packages (from spacy==2.2.4) (45.1.0.post20200127)
Requirement already satisfied: blis<0.5.0,>=0.4.0 in
c:\programdata\anaconda3\lib\site-packages (from spacy==2.2.4) (0.4.1)
Requirement already satisfied: requests<3.0.0,>=2.13.0 in
c:\programdata\anaconda3\lib\site-packages (from spacy==2.2.4) (2.23.0)
Requirement already satisfied: importlib-metadata>=0.20; python_version < "3.8"
in c:\programdata\anaconda3\lib\site-packages (from
catalogue<1.1.0,>=0.0.7->spacy==2.2.4) (1.5.0)
Requirement already satisfied: chardet<4,>=3.0.2 in
c:\programdata\anaconda3\lib\site-packages (from
requests<3.0.0,>=2.13.0->spacy==2.2.4) (3.0.4)
Requirement already satisfied: certifi>=2017.4.17 in
c:\programdata\anaconda3\lib\site-packages (from
requests<3.0.0,>=2.13.0->spacy==2.2.4) (2020.6.20)
Requirement already satisfied: urllib3!=1.25.0,!=1.25.1,<1.26,>=1.21.1 in
c:\programdata\anaconda3\lib\site-packages (from
requests<3.0.0,>=2.13.0->spacy==2.2.4) (1.25.8)
Requirement already satisfied: idna<3,>=2.5 in
c:\programdata\anaconda3\lib\site-packages (from
requests<3.0.0,>=2.13.0->spacy==2.2.4) (2.8)
Requirement already satisfied: zipp>=0.5 in c:\programdata\anaconda3\lib\site-
packages (from importlib-metadata>=0.20; python_version <
"3.8"->catalogue<1.1.0,>=0.0.7->spacy==2.2.4) (2.1.0)
Note: you may need to restart the kernel to use updated packages.
```

[6]:
```
def spacy_tokenizer(sentence):
    return [word.lemma_ for word in nlp(sentence) if not (word.like_num or word.
 ↪is_stop or
```

```
                                                        word.is_punct or word.
↪is_space or len(word)==1)]
```

[7]:
```
# Here we are downloading Wordcloud to create wordcloud based on the column␣
↪values using textmining

from PIL import Image
from wordcloud import WordCloud, STOPWORDS, ImageColorGenerator
```

[8]:
```
body_text = " ".join(text for text in df['body_text'])
```

[19]:
```
(print(len(body_text[:50000000])))
```

```
50000000
```

[13]:
```
# generating Wordcloud based on the frequency of word.

# Create stopword list:
stopwords = set(STOPWORDS)
stopwords.update([    'doi', 'preprint', 'copyright', 'org', 'https', 'et',␣
↪'al', 'author', 'figure', 'table',
    'rights', 'reserved', 'permission', 'use', 'used', 'using', 'biorxiv',␣
↪'medrxiv', 'license', 'fig', 'fig.', 'al.', 'Elsevier', 'PMC', 'CZI',
    '-PRON-', 'usually',
    r'\usepackage{amsbsy', r'\usepackage{amsfonts', r'\usepackage{mathrsfs',␣
↪r'\usepackage{amssymb', r'\usepackage{wasysym',
    r'\setlength{\oddsidemargin}{-69pt',  r'\usepackage{upgreek',␣
↪r'\documentclass[12pt]{minimal','although','within'])

wc = WordCloud(background_color="white", max_words=2000, stopwords=stopwords,␣
↪max_font_size=50,
            contour_width=3, contour_color='firebrick')
wc.generate(body_text[:5000000])
plt.figure(figsize=(20,15))
plt.imshow(wc, interpolation='bilinear')
plt.axis("off")
plt.show()
```

[14]:
```
# New stop words list
customize_stop_words = [
    'doi', 'preprint', 'copyright', 'org', 'https', 'et', 'al', 'author',␣
 ↪'figure', 'table',
    'rights', 'reserved', 'permission', 'use', 'used', 'using', 'biorxiv',␣
 ↪'medrxiv', 'license', 'fig', 'fig.', 'al.', 'Elsevier', 'PMC', 'CZI',
    '-PRON-', 'usually',
    r'\usepackage{amsbsy', r'\usepackage{amsfonts', r'\usepackage{mathrsfs',␣
 ↪r'\usepackage{amssymb', r'\usepackage{wasysym',
    r'\setlength{\oddsidemargin}{-69pt',  r'\usepackage{upgreek',␣
 ↪r'\documentclass[12pt]{minimal'
]

# Mark them as stop words
for w in customize_stop_words:
    nlp.vocab[w].is_stop = True
```

[16]:
```
filepath = '../input/topic-modeling-finding-related-articles/'
```

[17]:
```
vectorizer = CountVectorizer(tokenizer = spacy_tokenizer, min_df=2)
data_vectorized = vectorizer.fit_transform(tqdm(all_texts))
```

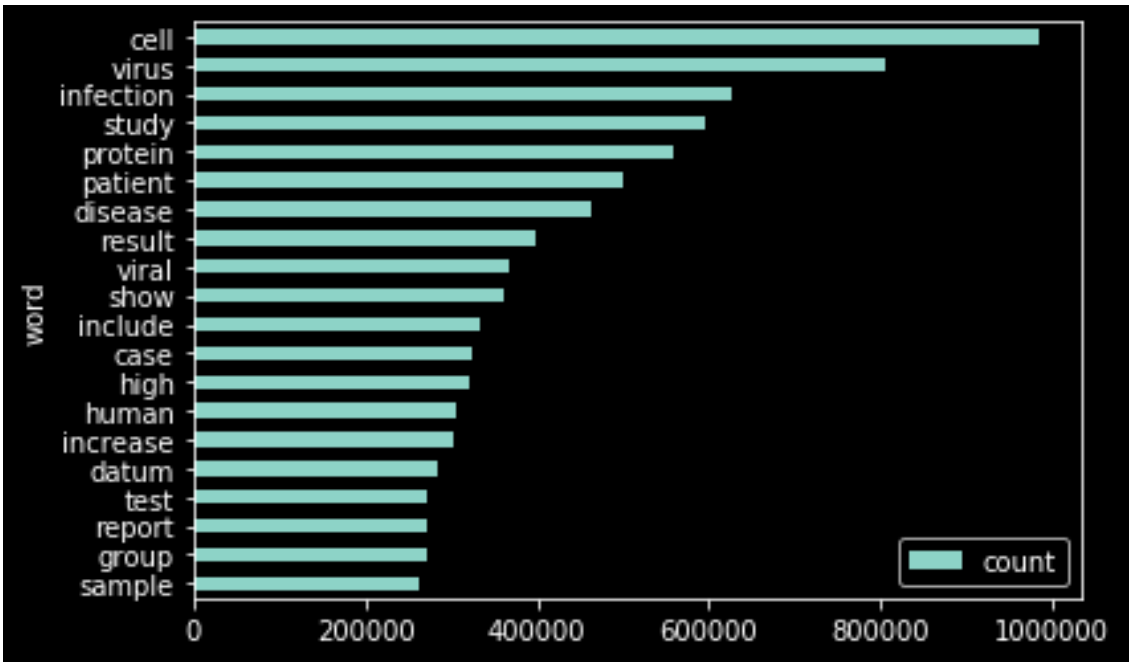
```
100%|          | 47110/47110 [29:00<00:00, 27.06it/s]
```

[18]:
```
data_vectorized.shape
```

[18]:
```
(47110, 541125)
```

[19]:
```
# most frequent words
word_count = pd.DataFrame({'word': vectorizer.get_feature_names(), 'count': np.
 ↪asarray(data_vectorized.sum(axis=0))[0]})

word_count.sort_values('count', ascending=False).set_index('word')[:20].
 ↪sort_values('count', ascending=True).plot(kind='barh')
```

[19]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x204cf0b0f08>
```



[21]:
```
joblib.dump(vectorizer, '../topic-modeling-finding-related-articles/vectorizer.
 ↪csv')
joblib.dump(data_vectorized, '../topic-modeling-finding-related-articles/
 ↪data_vectorized.csv')
```

[21]:
```
['../topic-modeling-finding-related-articles/data_vectorized.csv']
```

[26]:
```
lda = LatentDirichletAllocation(n_components = 50,␣
 ↪random_state=0,learning_method='online',verbose=1)
lda.fit(data_vectorized)
joblib.dump(lda,'../topic-modeling-finding-related-articles/lda.csv')
```

```
iteration: 1 of max_iter: 10
iteration: 2 of max_iter: 10
iteration: 3 of max_iter: 10
iteration: 4 of max_iter: 10
iteration: 5 of max_iter: 10
```

```
iteration: 6 of max_iter: 10
iteration: 7 of max_iter: 10
iteration: 8 of max_iter: 10
iteration: 9 of max_iter: 10
iteration: 10 of max_iter: 10
```

[26]:
```
['../topic-modeling-finding-related-articles/lda.csv']
```

[27]:
```
def print_top_words(model, vectorizer, n_top_words):
    feature_names = vectorizer.get_feature_names()
    for topic_idx, topic in enumerate(model.components_):
        message = "\nTopic #%d: " % topic_idx
        message += " ".join([feature_names[i]
                             for i in topic.argsort()[:-n_top_words - 1:-1]])
        print(message)
    print()
```

[28]:
```
print_top_words(lda, vectorizer, n_top_words=25)
```

```
Topic #0: blood test plasma donor product transfusion platelet sample unit p.
process result red inactivation screen study pcv2 reaction pool antibody safety
positive donation storage type

Topic #1: protein membrane fusion domain cell acid mutant amino contain site
process sequence cleavage protease form activity show er region function express
suggest residue complex encode

Topic #2: sars-cov mers-cov coronavirus sars human cov respiratory camel mers
spike animal infection study covs rbd bat severe dpp4 viral syndrome s1 middle
virus report patient

Topic #3: gene expression target sirna mrna vector express cell promoter
transcript dna plasmid level protein rna transcription control sirnas
transfection silence transfect genome reporter pathway rnai

Topic #4: health public disease research need system provide include information
risk country work new global service state care development response government
issue plan national people international

Topic #5: model time numb rate value datum individual parameter population
result estimate network different effect distribution increase case give show
study probability change epidemic follow mean

Topic #6: cell antibody serum culture assay control protein incubate medium
plate show wash pbs min result determine perform level time day test sample
stain describe datum
```

Topic #7: dengue infection zikv denv mosquito fever pregnancy wnv woman chikv pregnant zika maternal mother fetal birth pdcs encephalitis flavivirus ae fetus neonatal newborn flaviviruses serotype

Topic #8: \usepackage{amsmath methylation mt pei ppi \left interaction ppis y2h aps asl vips gav wd rg ami \beta sch covid-19 \lambda bsr f0 \in isr pscnv

Topic #9: detection assay dna method target system sensitivity probe technique application reaction develope high detect nanoparticle specific technology test specificity nucleic surface amplification design rapid diagnostic

Topic #10: hiv hiv-1 env macaque cd4 gag aid viral monkey siv retrovirus rhesus tat primate target antiretroviral retroviral cq immunodeficiency envelope rev mlv load tetherin p24

Topic #11: datum sequence analysis sample method identify set base result test approach numb database information different cluster provide include perform value read select study dataset score

Topic #12: ace2 rat lung heart ii increase receptor kidney injury level cardiac tissue ang effect ace mouse study diabetes activity fibrosis renal blood kd human enzyme

Topic #13: mm ph cell concentration nm show membrane solution particle min contain increase observe buff experiment temperature result obtain ml presence fraction water condition study low

Topic #14: cause disease occur intestinal sign animal small include result clinical increase live horse day tissue fluid blood infection common body treatment affect skin intestine case

Topic #15: influenza virus ha h1n1 pandemic h5n1 autophagy human avian iav subtype na infection np ifitm3 antiviral flu strain oseltamivir h3n2 seasonal neuraminidase ferret mdck viral

Topic #16: cell response study expression increase role induce receptor activation immune effect human level function factor mechanism show protein pathway signal cytokine suggest activate result disease

Topic #17: covid-19 sars-cov-2 author/funder display grant available peer-reviewed work perpetuity international post allow report april test cc-by-nc-nd reuse version datum submit section receive review form present

Topic #18: sperm bfa arm microtubule cilium pcd hunov mhv-infected dynein ciliary fertility kir mif -/mice 17cl-1 spermatozoon defect pmc mpg possum transport acanthamoeba pbb hpev1 nocodazole

Topic #19: ifn signal response type activation pathway innate rig-i protein

activity phosphorylation induction nf- b activate immune antiviral ifn- ubiquitin degradation inhibit induce production kinase dsrna isg15

Topic #20: der apod icv int lattice und 1998a morphine 1996a iec s-layer 1996b sdab 1998b valence ad3 ling eqa cff method(s dbd som jong ad2 hy

Topic #21: surveillance country vaccination influenza datum disease travel year report malaria region vaccine system season ili health africa seasonal fever cdc site national international pilgrim hajj

Topic #22: patient pneumonia lung treatment clinical disease infection respiratory severe antibiotic therapy pulmonary acute case cause include airway diagnosis increase day symptom asthma bacterial associate result

Topic #23: bind structure residue domain interaction site complex form structural model molecule region conformation bond chain position crystal loop ligand substrate pro affinity energy dock motif

Topic #24: rna sequence codon translation genome structure synthesis mrna nucleotide site rnas region replication polymerase mrnas end strand contain ribosome element mutant initiation dna helicase frameshifting

Topic #25: lipid wang chen li zhang liu van lee app raft lin smith pm zhao yu su brown ma jiang williams arf phospholipid martin 2006a wilson

Topic #26: animal pig calve disease infection farm cattle diarrhea milk herd water rotavirus food sample pathogen bovine high cause infect swine c. cow coli calf e.

Topic #27: bind receptor acid surface cell glycoprotein lectin sialic human glycosylation carbohydrate glycans attachment glycan affinity dc-sign oligosaccharide type recognize interaction specificity site entry recognition molecule

Topic #28: gene mutation genetic strain genotype variant resistance allele population frequency phenotype susceptibility polymorphism region variation study association locus associate snps individual resistant difference snp chromosome

Topic #29: virus human species host disease bat pathogen animal population transmission new include cause infection study find family know wild genus high occur genome year example

Topic #30: air room temperature droplet exposure mask particle hand tb surface aerosol control care environment procedure water area hospital transmission respiratory staff clean airborne concentration facility

Topic #31: cat dog study de feline test sample university canine fip positive

group disease result serum clinical usa fcov la perform report t. concentration fipv healthy

Topic #32: plant production extract e. produce coli product system growth yeast culture recombinant a. leave seed high food bacterium expression yield level grow insect tobacco n.

Topic #33: patient infection treatment therapy disease recipient cancer transplant risk clinical cmv transplantation hepatitis live include ebv receive treat donor chronic hbv day month dose cell

Topic #34: sequence virus strain rna sample primer gene pcr genome viral isolate show analysis pedv region detect study acid dna nucleotide rt-pcr result perform positive contain

Topic #35: infant nec preterm rst emodin identifi premature ep signifi nicu diffi benefi confi il-35 cient bioactive fq phytochemical feed unigenes rmed myricetin ndings glycoside cantly

Topic #36: strain ibv virus chicken group bird isolate poultry egg day inoculate dpi avian infect show study inoculation ndv infection tissue titre s1 vaccine duck challenge

Topic #37: exon splice cftr isoform proteasome intron ifn- pbm macrodomain lbms mhv-1 degradation macrodomains uv-c enac b mutagenic pers aav2 merlin proteasomal lpm skp2 ifnlr1 mutator

Topic #38: case outbreak infection disease transmission epidemic contact day report china infect sars spread numb control patient death country infectious people health symptom period time rate

Topic #39: virus respiratory infection child study viral test sample rsv detect patient specimen positive clinical pathogen pneumonia detection influenza pcr age year symptom tract case illness

Topic #40: peptide disorder zinc conjugate muscle amino toxin antimicrobial delivery specifi charge uptake ppmo dystrophin macro wssv pmo pip pna conjugation defensins nsp11 dmd cationic skeletal

Topic #41: cell lesion disease brain tissue ms cns neuron occur cause present area spinal rat nerve csf normal case change include result cord affect microglia associate

Topic #42: virus viral cell infection replication infect host rna antiviral hcv entry cellular particle virion inhibit target protein mechanism require infectious ebov genome effect replicate hepatitis

Topic #43: activity compound drug inhibitor effect inhibit antiviral treatment

```
inhibition acid group concentration show active derivative enzyme agent
inhibitory 1h hz treat vitro effective value protease

Topic #44: ab a1 a2 aa ph1n1 ptb hnrnp ddx1 gt cg nf-b nf hsr nucleolin snv rha
ddx5 com igr  t snvs fm rrs ka llc-pk1

Topic #45: wu 2005a tan li 2005b 2002a fork rosa sharma 2002b suppl nanotrap ber
qa p.t qu antarctic mao roe vo cog shimizu efflux hammond yoo

Topic #46: train student learn fi resident teach medical education skill course
video hcp bms telemedicine lecture classroom rfid erent programme ective fellow
content basic cod sps

Topic #47: vaccine antibody response antigen immune epitope vaccination virus
neutralize challenge immunization immunity protection induce study human serum
vector development igg recombinant animal high vaccinate dose

Topic #48: study patient group age high datum hospital year risk report analysis
result child include compare significant level increase care factor control
difference participant rate effect

Topic #49: mouse cell infection day response cd8 lung cd4 animal infect immune
cytokine model lymphocyte level macrophage tissue control compare observe show
expression increase follow group
```

[31]:
```
doc_topic_dist = pd.DataFrame(lda.transform(data_vectorized))
doc_topic_dist.to_csv('../topic-modeling-finding-related-articles/
 ↪doc_topic_dist.csv', index=False)
```

[32]:
```
doc_topic_dist.head()
```

[32]:

	0	1	2	3	4	5	6	\
0	0.009452	0.000014	0.000014	0.010177	0.001276	0.017239	0.140493	
1	0.000007	0.000007	0.000007	0.026468	0.000007	0.000007	0.177363	
2	0.000007	0.203549	0.000007	0.010034	0.003828	0.015899	0.095299	
3	0.000032	0.010201	0.000032	0.000032	0.045060	0.000032	0.009011	
4	0.000039	0.000039	0.000039	0.000039	0.342898	0.094446	0.000039	

	7	8	9	…	40	41	42	43	\
0	0.000014	0.000014	0.000014	…	0.000014	0.000014	0.000014	0.000014	
1	0.000007	0.000007	0.000007	…	0.000007	0.000007	0.042212	0.000007	
2	0.000007	0.000007	0.010066	…	0.000007	0.000007	0.282044	0.000007	
3	0.000032	0.000032	0.016270	…	0.000032	0.000032	0.000032	0.003720	
4	0.000039	0.000039	0.062805	…	0.000039	0.000039	0.000039	0.000039	

	44	45	46	47	48	49

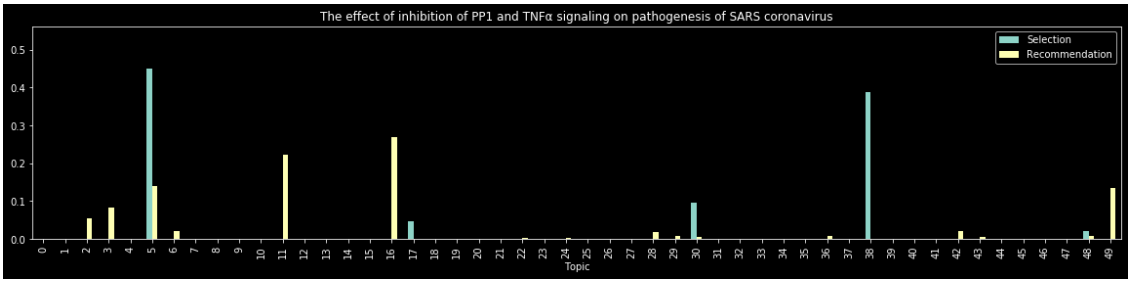
```
0  0.000014  0.000014  0.000014  0.000014  0.000014  0.000014
1  0.009549  0.000007  0.000007  0.422698  0.002827  0.056158
2  0.000007  0.000007  0.000007  0.005721  0.003749  0.000007
3  0.000032  0.000032  0.000032  0.007538  0.276761  0.014606
4  0.000039  0.000039  0.000039  0.000039  0.006979  0.000039

[5 rows x 50 columns]
```

[34]:
```
is_covid19_article = df.body_text.str.
↪contains('COVID-19|SARS-CoV-2|2019-nCov|SARS Coronavirus 2|2019 Novel␣
↪Coronavirus|Corona|Corona virus')
```

[35]:
```
def get_k_nearest_docs(doc_dist, k=5, lower=1950, upper=2020,␣
↪only_covid19=False, get_dist=False):
    '''
    doc_dist: topic distribution (sums to 1) of one article

    Returns the index of the k nearest articles (as by Jensen-Shannon␣
↪divergence in topic space).
    '''

    relevant_time = df.publish_year.between(lower, upper)

    if only_covid19:
        temp = doc_topic_dist[relevant_time & is_covid19_article]

    else:
        temp = doc_topic_dist[relevant_time]

    distances = temp.apply(lambda x: jensenshannon(x, doc_dist), axis=1)
    k_nearest = distances[distances != 0].nsmallest(n=k).index

    if get_dist:
        k_distances = distances[distances != 0].nsmallest(n=k)
        return k_nearest, k_distances
    else:
        return k_nearest
```

[50]:
```
# def plot_article_dna(paper_id, width=20):
#     t = df[df.paper_id == paper_id].title.values[0]
#     doc_topic_dist[df.paper_id == paper_id].T.plot(kind='bar', legend=None,␣
↪title=t, figsize=(width, 4))
#     plt.xlabel('Topic')

def compare_dnas(paper_id, recommendation_id, width=20):
    t = df[df.paper_id == recommendation_id].title.values[0]
    temp = doc_topic_dist[df.paper_id == paper_id]
```

```
    ymax = temp.max(axis=1).values[0]*1.25
    temp = pd.concat([temp, doc_topic_dist[df.paper_id == recommendation_id]])
    temp.T.plot(kind='bar', title=t, figsize=(width, 4), ylim= [0, ymax])
    plt.xlabel('Topic')
    plt.legend(['Selection', 'Recommendation'])

compare_dnas('90b5ecf991032f3918ad43b252e17d1171b4ea63',␣
 ↪'a137eb51461b4a4ed3980aa5b9cb2f2c1cf0292a')

# def dna_tabs(paper_ids):
#     k = len(paper_ids)
#     outs = [widgets.Output() for i in range(k)]

#     tab = widgets.Tab(children = outs)
#     tab_titles = ['Paper ' + str(i+1) for i in range(k)]
#     for i, t in enumerate(tab_titles):
#         tab.set_title(i, t)
#     display(tab)

#     for i, t in enumerate(tab_titles):
#         with outs[i]:
#             ax = plot_article_dna(paper_ids[i])
#             plt.show(ax)

# def compare_tabs(paper_id, recommendation_ids):
#     k = len(recommendation_ids)
#     outs = [widgets.Output() for i in range(k)]

#     tab = widgets.Tab(children = outs)
#     tab_titles = ['Paper ' + str(i+1) for i in range(k)]
#     for i, t in enumerate(tab_titles):
#         tab.set_title(i, t)
#     display(tab)

#     for i, t in enumerate(tab_titles):
#         with outs[i]:
#             ax = compare_dnas(paper_id, recommendation_ids[i])
#             plt.show(ax)
```



[45]:
```
def relevant_articles(tasks, k=3, lower=1950, upper=2020, only_covid19=False):
    tasks = [tasks] if type(tasks) is str else tasks

    tasks_vectorized = vectorizer.transform(tasks)
    tasks_topic_dist = pd.DataFrame(lda.transform(tasks_vectorized))

    for index, bullet in enumerate(tasks):
        print(bullet)
        recommended = get_k_nearest_docs(tasks_topic_dist.iloc[index], k,␣
↪lower, upper, only_covid19)
        recommended = df.iloc[recommended]

        h = '<br/>'.join(['<a href="' + l + '" target="_blank">'+ n + '</a>'␣
↪for l, n in recommended[['url','title']].values])
        display(HTML(h))
```

[46]:
```
def relevant_articles_for_text():
    textW = widgets.Textarea(
        value='',
        placeholder='Type something',
        description='',
        disabled=False,
        layout=Layout(width='90%', height='200px')
    )

    yearW = widgets.IntRangeSlider(min=1950, max=2020, value=[2010, 2020],␣
↪description='Year Range',
                                   continuous_update=False,␣
↪layout=Layout(width='40%'))
    covidW = widgets.Checkbox(value=True,description='Only␣
↪COVID-19-Papers',disabled=False, indent=False, layout=Layout(width='25%'))
    kWidget = widgets.IntSlider(value=10, description='k', max=50, min=1,␣
↪layout=Layout(width='25%'))

    button = widgets.Button(description="Search")

    display(VBox([HBox([kWidget, yearW, covidW], layout=Layout(width='90%',␣
↪justify_content='space-around')),
        textW, button], layout=Layout(align_items='center')))

    def on_button_clicked(b):
        clear_output()
        display(VBox([HBox([kWidget, yearW, covidW], layout=Layout(width='90%',␣
↪justify_content='space-around')),
```

```
            textW, button], layout=Layout(align_items='center')))
        relevant_articles(textW.value, kWidget.value, yearW.value[0], yearW.
↪value[1], covidW.value)

    button.on_click(on_button_clicked)
```

[47]:
```
relevant_articles_for_text()
```

```
VBox(children=(HBox(children=(IntSlider(value=10, description='k', layout=Layout(width='25%'),
```