

SGanguli_Assignment5

June 11, 2024

1 Assignment 5: Topic Modeling (NLP)

1.1 Import Libraries

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

import json
import sys
import os
import ast
import random
pd.set_option('display.max_columns', 40)
```

```
[2]: import nltk
import gensim
import wordcloud
import faiss
from nltk.corpus import stopwords
from helper_functions import *
```

1.2 Data Processing

```
[3]: root_location = 'assignment_5_data' # Insert Root Location here

#file_locations = [f'{root_location}/biorxiv_medrxiv/biorxiv_medrxiv',
#                  f'{root_location}/noncomm_use_subset/noncomm_use_subset',
#                  f'{root_location}/comm_use_subset/comm_use_subset',
#                  f'{root_location}/custom_license/custom_license']

file_locations = [f'{root_location}/biorxiv_medrxiv/biorxiv_medrxiv']

## Set up Stop Words
## Add Any other relevant options manually

stop_words = stopwords.words('english') + ['et', 'al', 'fig', 'etal', 'et al', '↵',
↵ 'et-al']
```

```

processed_articles = process_articles(file_locations, stop_words)
processed_articles.read_files()

print('Example File Name')
print(processed_articles.root_files[0])
print('Number of files')
print(len(processed_articles.root_files))
print('Example Article Information')
print(processed_articles.title_text[2])

```

Processing Files at the requested location

There are 177 files to process

There were 885 files in the dataset

Example File Name

160ecaaa5a766c289fc2f6b5499f0dfe7aab971c.json

Number of files

177

Example Article Information

['biorxiv_medrxiv', 'The effectiveness of full and partial travel bans against COVID-19 spread in Australia for travellers from China', ['In response to the epidemic of COVID-19, (1) Australia implemented a travel ban from China on February 1 st 2020, adding Iran and then South Korea to the ban on February 29 th and March 5 th respectively. In addition, Australians evacuated from Wuhan and from the Diamond Princess cruise ship were quarantined for two weeks in dedicated quarantine facilities. The ban on travel from China has been periodically reviewed, with lifting of restrictions announced on February 23 rd for high school students, who number less than 800. In contrast, over 120000 university students are unable to enter Australia to commence or resume their studies, and a booming tourism industry has ceased. Travel bans and social distancing measures are effective public health tools to control epidemic diseases (2), and Australia successfully delayed the introduction of the 1918 pandemic by 1 year and reduced the total mortality compared to other countries (3) . However, travel bans are not sustainable indefinitely, and a careful risk analysis needs to be done comparing the health and economic consequences of alternative scenarios. The epidemic in China peaked on February 5 th and has declined since (4) . The risk of importation of COVID-19 cases through travel from an affected country is proportional to the volume of travel from that country and their prevalence of infection at that time point. We aimed to estimate the impact of the implementation of the travel ban on China from February 1 st 2020 on the epidemic trajectory in Australia, as well as the impact of lifting the ban completely or partially from the 8 of March. Three scenarios were considered. 1. No travel ban -the epidemic curve if the travel ban was never placed 2. Complete travel ban from February 1st to March 8st, followed by complete lifting ban 3. Complete travel ban from February 1st to March 8st, followed by partial lifting ban (allowing university students, but not tourists, to enter the country) The evacuations from Wuhan and the Diamond Princess Cruise ship are not considered in this model, which only examines

regular travel between China and Australia. In order to estimate the effectiveness of the travel ban that has been implemented in Australia for travellers from China, we did not consider bans to other countries. We assumed that the chance of cases coming into Australia from China depends from the number of cases in China and the number of travellers to Australia. To estimate the number of people infected that are predicted to enter Australia every two weeks from 20/01 to April, we utilised 2019 air travel passenger movements between China and Australia, derived from incoming passenger arrival cards, with data aggregated monthly and published by the Australia Bureau of Statistics (5). For the purpose of this analysis air movements of passengers between China and Australia were derived from 2019 data. A baseline level of entries into Australia from China was calculated from the total number of entries over the April -Jun 2019 time period and was assumed to represent the baseline arrivals for the purpose of tourism and other business. The seasonal excess of travellers was then calculated by deducting this baseline from the January-March 2019 data. The seasonal excess arrivals were assumed to represent the arrival of international students starting the 2019 study year, which begins in February to March each year. Where travel bans were instituted in this analysis, or lifted, it was assumed that international students unable to enter Australia would return to Australia following the lifting of the ban, 60% in the rest of March and the remaining 40% over the month of April. However, tourists not able to travel during a travel ban were not assumed to enter Australia at a later date. Tourism activity was assumed to recover to baseline levels immediately after the lifting of a travel ban. The daily number of travellers from China to Australia in each month and for each scenario is showed in Table S2 of the supplementary material. To then calculate the probable number of those that could be infected we used an epidemiological dataset of confirmed cases of COVID-19 in China collected from WHO situation reports (6) and available in our supplementary materials (Table S1). The dataset includes all confirmed cases in China reported from 31/12/2019 to 23/02/2020. We then assumed that notified cases reflect only 10% of the real new infections per day, due to under-reporting, mild cases and asymptomatic infections. This assumption is based on data from Japan (7), which estimated that only 9.2% of cases in China were notified or detected. This estimate is based on testing of all evacuees from Wuhan to Japan and the documented cases in China at the time (7). Furthermore it has been showed that a high proportion of infected people will have very mild symptoms (8) which are unlikely to be reported. We then estimated the possible true epidemic curve. In order to project the future incidence cases in China we used a Poisson regression model to fit data from the 5th (start of the incidence declining) to 23rd of February and estimated the decreasing rate per day (z) as: Where $I(t)$ is the number of new infected at time t and I_0 is the initial value at time $t=0$ (Incidence at day 5 of February). Once the decreasing rate z was estimated and the incidence forecasted from 23 of February onwards, we then calculated the number of infected people coming from China every two weeks period ($t, t+14$) as: Where N is the total population of China and ($t, t+14$) is the number of people travelling from China to Australia in every two weeks period. When calculating the prevalence of infection in China, we started from two weeks before the period travelling in order to include the people that could

be infected and in a latent state. In scenarios 2 and 3, we assumed a linear declining distribution in time of travel for university students waiting to enter the country after lifting of the travel ban. A full and partial lifting of the ban was examined. In the partial ban, over 150,000 university students can enter Australia, but the just over 80,000 expected tourists not. The cases of COVID-19 occurring over time in Australia due to imported cases from China were estimated for each scenario. We used an age specific deterministic model, with 8 mutually exclusive compartments: susceptible (S), Latent traced (LT), Latent untraced (LU), Infectious (I), Isolated (I), Recovered (R) and dead (D). Each of those compartments is divided in 18 age stratified groups each of 5 years duration, ranging from 0 to 84 years old plus an additional age group of 85+ years. The entire Australian population was considered susceptible. The duration of each model run is 400 days. The initial infected cohort is assumed to be generated from cases arriving from China by air. After arrival of an infected case, it is assumed that, if and when they become symptomatic, they are isolated, and a designated portion of their contacts will also be quarantined. Cases transition between epidemiological compartments in accordance with transition rates determined by their duration of stay in each compartment. Model parameters are shown in Table 1 . Further details of the model (diagram and differential equations) are described in the supplementary material. We conducted a sensitivity analyses on the proportion of asymptomatic people, and based on growing evidence of equal viral loads in symptomatic and asymptomatic cases(9-12), we considered the latent period to be equally infectious as the symptomatic period. The proportion of asymptomatic infections, being the main source of community transmission in scenarios where the travel bans are lifted, was assumed to be 34.6% based on testing of passengers aboard the Diamond Princess cruise ship (13, 14) . The model uses an optimistic assumption that 80% of contacts are identified and quarantined, and 90% of symptomatic cases are isolated after 5 days (17) . Studies show a long, mild prodrome of several days before people feel unwell enough to seek medical attention, which is also considered in the model (15) . to the 23 of February. Figure 2 shows the modelled epidemic curve fitting the incidence data from 5 to 23 of February and then forecasted until the 4 of April, which is the time we expect the incidence decreasing to almost zero should the current trend in China continue. . CC-BY-NC-ND 4.0 International license It is made available under a is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. The copyright holder for this preprint .

<https://doi.org/10.1101/2020.03.09.20032045> doi: medRxiv preprint We found that following the peak on February 5 th and decline of the epidemic in China, the probability that an infected traveller can arrive in the partial ban scenario (allowing university students only) is low. The complete removal of travel restrictions on 8 th of March results in an estimated arrival of 5 cases in the first two weeks and 1 in the following two weeks. However, if we compare a 5 week ban scenario (scenario 2) with the scenario without a travel ban (scenario 1), we estimate that 32, 43 and 36 infected coming every two weeks from the 26 of January would have been averted. Due to a surge of students coming in the first two weeks following the lifting of the ban in the second scenario, an additional 2 more infected are estimated to enter from 8 to 21 of March (Table 2

). In Figure 3 we show the epidemic curve without and with the ban implemented for 5 weeks followed by a full lifting (scenario 1 and 2) and we show a large impact on averting an epidemic in Australia. In both cases, the model reproduces the 15 notified imported cases reported in Australia between the 20 of January and 8 of February. The modelled epidemic in scenario 2, with the full ban, predicts 57 cases in Australia by the 6th of March. The notified cases by 6th of March were 66, however we did not account for imported cases from other countries. In the epidemic curve for scenario 2, when travel resumes there will be a small surge in cases followed by a decrease and the epidemic can be controlled, with a total of less than 300 cases and about 8 deaths. If the ban was never in place (scenario 1), the epidemic would continue for more than a year resulting in more than 2000 cases and about 400 deaths. Varying the proportion of asymptomatic cases to 20% or 50% keeps about the same percentages of cases and deaths prevented by the ban in the two scenarios, however the total number of cases and deaths are almost 6 times higher and 2 times lower in the case of the proportion of asymptomatic being 50% and 20% respectively (results not shown). We estimated that the travel ban implemented on 1 of February by Australia has been very effective, reducing the number of cases and deaths from COVID-19 by about 87%. Studies have been published on effectiveness of domestic and international travel restrictions on COVID-19 (19, 20). However, this study is the first one to show the effectiveness of travel ban in Australia, and can inform a phased approach of partial lifting of bans when cases in the source country decline. This allows monitoring of the ongoing situation in China, which may yet see a second wave of the epidemic. Our estimate of the true epidemic curve is supported by other studies (7, 8, 21), and projected case numbers would change with any change in this estimate. Even if the true number of cases in China is 10 or 100 times that reported, only a fraction of the entire population of China has been infected, which leaves a possibility of a subsequent wave of the epidemic. If cases increase in China, the model can provide estimates of risk based on daily new case numbers. We do not consider cases coming in from other countries -however, this study illustrates the principle of travel bans and public health impact on epidemic control using China as a case study. A further limitation is the uncertainty of parameters used, particularly the proportion of asymptomatic cases. We have used a conservative estimate, but if the rate is higher than 40%, the outcomes would be worse. While it has been showed that distancing measures are highly effective (2, 22) a systematic review looking at the effectiveness of travel restrictions (23), shows that international travel restriction are effective in delaying the epidemic but may not contain it. We also assumed a very optimistic scenario of 80% of contacts being identified, which may not occur with high case numbers, if a high proportion of asymptomatic transmission is occurring, or if self-quarantine is ineffective. In this study we assumed voluntary home quarantine, which is showed to be about 50% effective in R_0 reduction (16), however there could be an increased risk of intra-household transmission infected people to contacts (24), which is not considered in this model. We showed that the ban implemented for travellers from China, when the epidemic was almost at its peak, substantially delayed the spread into Australia. There is now evidence of community transmission in Australia, but the epidemic is still in the early

stages, and this study provide evidence to support the new travel bans that have been implemented on Iran and South Korea, in order to delay the epidemic. The model predicted 57 cases by March 6 th in Australia, which is slightly less than the notified number of 66, which suggests the model assumptions were reasonable, given we did not account for cases coming in from other countries. Community transmission in Australia in early March is likely linked to imported cases from China, given the fairly long incubation period and less than 3 incubation periods since the first evacuation of Australians from Wuhan on February 3rd. The model fit to observed data was good, also suggesting the epidemic is still possible to contain, if adequate resources are available for thorough contact tracing. This analyses is a first insight into the effectiveness of travel restrictions for COVID-19 outbreak, supports the effectiveness of the Australian response, informs gradual lifting of the bans or placing of new bans on other countries, and could inform other countries in reducing the burden of importations and resulting domestic transmission of COVID-19. Valentina Costantino: Methodology and modelling construction, parameterisation of model, writing and revision David Heslop: Manipulation of travelling data, writing and revision Raina MacIntyre: Conception of the study and scenarios, parametrisation of model, writing and revision. All: Designing and conceptualising the model']]

```
[4]: processed_articles.process_text()
```

```
Cleaning out Junk
Tokenizing words
Converting to list of words and removing stop words
Creating word stems
```

```
[8]: processed_articles.trigrams([art[3] for art in processed_articles.
    ↪processed_article])
processed_articles.processed_trigrams[2][4]
```

```
Creating Bigrams
Creating Trigrams from Bigrams
```

```
[8]: 'respons epidem covid australia implement travel_ban china februari st ad iran
south_korea ban februari_th march_th respect addit australian evacu wuhan
diamond_princess cruiss_ship quarantin two_week dedic quarantin facil ban travel
china period review lift restrict announc februari rd high school student number
less contrast univers_student unabl enter_australia commenc resum studi boom
tourism industri ceas travel_ban social_distanc_measur effect public_health tool
control epidem diseas australia success delay introduct pandem year reduc total
mortal compar countri howev travel_ban sustain indefinit care risk analysi need
done compar health econom consequ altern scenario epidem china peak februari_th
declin sinc risk import covid case travel affect_countri proport volum travel
countri preval infect time_point aim estim impact implement travel_ban china
februari st epidem trajectori australia well impact lift_ban complet partial
march three scenario consid travel_ban epidem_curv travel_ban never place
complet travel_ban februari st march_st follow complet lift_ban complet
```

travel_ban februari st march_st follow partial lift_ban allow univers_student
 tourist enter countri evacu wuhan diamond_princess cruiseship consid model
 examin regular travel china australia order estim effect travel_ban implement
 australia travel china consid ban countri assum chanc case come australia china
 depend number case china number travel australia estim number_peopl infect
 predict enter_australia everi two_week april utilis air_travel passeng movement
 china australia deriv incom passeng arriv card data aggreg monthli publish
 australia bureau statist purpos analysi air movement passeng china australia
 deriv data baselin level entri australia china calcul total_number entri april
 jun time_period assum repres baselin arriv purpos tourism busi season excess
 travel calcul deduct baselin januari_march data season excess arriv assum repres
 arriv intern student start studi year begin februari march year travel_ban
 institut analysi lift assum intern student unabl enter_australia would return
 australia follow lift_ban rest march remain month april howev tourist abl travel
 travel_ban assum enter_australia later date tourism activ assum recov baselin
 level immedi lift travel_ban daili_number travel china australia month scenario
 show tabl supplementari_materi calcul probabl number could infect use
 epidemiolog dataset confirm_case covid china collect situat_report avail
 supplementari_materi tabl dataset includ confirm_case china report assum notifi
 case reflect real new infect per_day due report mild case asymptomat_infect
 assumpt base data japan estim case china notifi detect estim base test evacue
 wuhan japan document case china time furthermor show high proport infect_peopl
 mild_symptom unlik report estim possibl true epidem_curv order project futur
 incid case china use poisson regress model fit data th start incid declin rd
 februari estim decreas rate per_day number new infect time initi valu time incid
 day februari decreas rate estim incid_forecast februari onward calcul number
 infect_peopl come china everi two_week period total popul china number_peopl
 travel china australia everi two_week period calcul preval infect china start
 two_week period travel order includ peopl could infect latent state scenario
 assum linear declin distribut time travel univers_student wait enter countri
 lift travel_ban full partial lift_ban examin partial ban univers_student
 enter_australia expect tourist case covid occur time australia due import_case
 china estim scenario use age_specif determinist_model mutual_exclus compart
 suscept latent trace lt latent untrac lu infecti isol recov dead compart divid
 age_stratifi group year durat rang year_old plu addit age_group year entir
 australian popul consid suscept durat model run day initi infect cohort assum
 gener case arriv china air arriv infect case assum becom symptomat isol design
 portion contact also quarantin case transit epidemiolog compart accord
 transit_rate determin durat stay compart model paramet shown_tabl detail model
 diagram differenti_equat describ supplementari_materi conduct sensit analys
 proport_asymptomat peopl base grow evid equal viral_load symptomat_asymptomat
 case consid latent_period equal infecti symptomat period proport_asymptomat
 infect main sourc commun transmiss scenario travel_ban lift assum base test
 passeng aboard diamond_princess cruiseship model use optimist assumpt contact
 identifi quarantin symptomat case isol day studi show long mild prodrom sever
 day peopl feel unwell enough seek medic attent also consid model februari
 figur_show model epidem_curv fit incid data februari forecast april time expect

incid decreas almost zero current trend china continu cc_nc_nd_intern
 licens_made_avail_author funder_grant_medrxiv_licens display_preprint
 perpetu_copyright_holder_preprint doi_medrxiv_preprint found follow peak
 februari_th declin epidem china probabl infect travel arriv partial ban scenario
 allow univers_student low complet remov travel_restrict th_march result estim
 arriv case first two_week follow two_week howev compar week ban
 scenario_scenario scenario without travel_ban scenario estim infect come everi
 two_week januari would avert due surg student come first two_week follow
 lift_ban second scenario addit infect estim enter march tabl figur_show
 epidem_curv without ban_implement week follow full lift scenario show larg
 impact avert epidem australia case model reproduc notifi import_case report
 australia januari_februari model epidem scenario full ban predict case australia
 th_march notifi case th_march howev account import_case countri epidem_curv
 scenario travel resum small surg case follow decreas epidem control total less
 case death ban never place scenario epidem would continu year result case death
 vari proport_asymptomat case keep percentag case death prevent ban two scenario
 howev total_number case death almost time higher time lower case
 proport_asymptomat respect result_shown estim travel_ban implement februari
 australia effect reduc number case death covid studi publish effect domest
 intern travel_restrict covid howev studi first one show effect travel_ban
 australia inform phase approach partial lift_ban case sourc countri declin allow
 monitor ongo situat china may yet see second wave epidem estim true epidem_curv
 support studi project case number would chang chang estim even true number case
 china time report fraction entir popul china infect leav possibl subsequ wave
 epidem case increas china model provid estim risk base daili new case number
 consid case come countri howev studi illustr principl travel_ban public_health
 impact epidem control use china case studi limit uncertainti paramet use
 particularli proport_asymptomat case use conserv estim rate higher outcom would
 wors show distanc measur highli effect systemat_review look effect
 travel_restrict show intern travel_restrict effect delay epidem may contain also
 assum optimist scenario contact identifi may occur high case number high
 proport_asymptomat transmiss occur self quarantin ineffect studi assum voluntari
 home quarantin show effect reduct howev could increas risk intra household
 transmiss infect_peopl contact consid model show ban_implement travel china
 epidem almost peak substanti delay spread australia evid commun transmiss
 australia epidem still earli_stage studi provid_evid support new travel_ban
 implement iran south_korea order delay epidem model predict case march_th
 australia slightli less notifi number suggest model assumpt reason given account
 case come countri commun transmiss australia earli march like link import_case
 china given fairli long incub_period less incub_period sinc first evacu
 australian wuhan februari rd model fit observ data good also suggest epidem
 still possibl contain adequ resourc avail thorough contact_trace analys first
 insight effect travel_restrict covid_outbreak support effect australian respons
 inform gradual lift_ban place new ban countri could inform countri reduc burden
 import result domest transmiss covid valentina costantino methodolog model
 construct model write revis david heslop manipul travel data write revis raina
 macintyr concept studi scenario parametris model write revis design conceptualis

model'

1.3 Create train and test data

```
[9]: from sklearn.feature_extraction.text import CountVectorizer

train, test = train_test_splitter(processed_articles.processed_trigrams, 0.90)

vectorizer = CountVectorizer(min_df = 50, max_df = 0.8, max_features = 50000)
tf = vectorizer.fit_transform([t[4] for t in train]) ## Vectorize training set
tf_feature_names = vectorizer.get_feature_names_out() ## Pull out words for use
↳ in eval

# Transform test data for perplexity eval

tf_test = vectorizer.transform([t[4] for t in test])
```

1.4 Build LDA models and display topics

```
[10]: from sklearn.decomposition import LatentDirichletAllocation

# Function to build and evaluate LDA models
def build_lda_model(tf, n_topics):
    lda = LatentDirichletAllocation(n_components=n_topics,
                                    learning_offset = 15.,
                                    learning_decay = 0.75,
                                    random_state=42)

    lda.fit(tf)
    return lda

def display_topics(model, feature_names, no_top_words):
    for topic_idx, topic in enumerate(model.components_):
        print(f"Topic {topic_idx}:")
        print(" ".join([feature_names[i] for i in topic.argsort()[:-no_top_words - 1:-1]]))
    ↳ -no_top_words - 1:-1]]))

# Build and display topics for 8, 9, and 10 topics
for n_topics in [8, 9, 10]:
    print(f"\nBuilding LDA model with {n_topics} topics")
    lda_model = build_lda_model(tf, n_topics)
    print(f"Topics in LDA model with {n_topics} topics:")
    display_topics(lda_model, tf_feature_names, 10)
```

Building LDA model with 8 topics

Topics in LDA model with 8 topics:

Topic 0:

cell express protein figur preprint_peer_review_author gene activ level increas

observ
Topic 1:
model epidem estim number case paramet valu effect figur day
Topic 2:
patient diseas sever effect method within sampl level increas rate
Topic 3:
sar_cov viru preprint_peer_review_doi patient human coronaviru virus may sampl
pathogen
Topic 4:
case outbreak transmiss estim model number import report individu distribut
Topic 5:
case covid report day estim china number confirm patient detect
Topic 6:
protein genom host virus structur viru complex cluster rna gene
Topic 7:
sequenc gene model genom identifi base one tabl test detect

Building LDA model with 9 topics

Topics in LDA model with 9 topics:

Topic 0:
cell express protein figur preprint_peer_review_author level increas activ viral
control
Topic 1:
model epidem estim number case valu paramet china day countri
Topic 2:
diseas effect rate express increas measur within patient level individu
Topic 3:
patient sar_cov viru preprint_peer_review_doi coronaviru sever human group may
covid
Topic 4:
outbreak case transmiss model number import individu estim local distribut
Topic 5:
case report covid estim day number detect china confirm sever
Topic 6:
genom protein rna virus viru structur host sequenc complex one
Topic 7:
gene sampl identifi compar specif base tabl detect method genom
Topic 8:
sequenc model report test one genom base tabl design set

Building LDA model with 10 topics

Topics in LDA model with 10 topics:

Topic 0:
cell express protein figur preprint_peer_review_author level increas activ viru
control
Topic 1:
model epidem estim number case paramet day valu china figur
Topic 2:

diseas effect rate method increas individu within measur level indic

Topic 3:
 patient sar_cov viru sever preprint_peer_review_doi human may diseas virus
 coronaviru

Topic 4:
 case outbreak transmiss number import estim individu model local distribut

Topic 5:
 case report covid estim day detect number china confirm sever

Topic 6:
 protein structur virus complex viru genom interact cluster form rna

Topic 7:
 gene genom identifi compar sampl host one tabl specif associ

Topic 8:
 sequenc preprint_peer_review_doi sar_cov target base method sampl genom human
 rna

Topic 9:
 sequenc model report test one design tabl may viru preprint_peer_review_author

1.5 Perplexity for topics

```
[11]: lda_8 = build_lda_model(tf, 8)
lda_9 = build_lda_model(tf, 9)
lda_10 = build_lda_model(tf, 10)

# Calculate Perplexity
perplexity_8 = lda_8.perplexity(tf_test)
perplexity_9 = lda_9.perplexity(tf_test)
perplexity_10 = lda_10.perplexity(tf_test)

print(f"Perplexity for 8 topics: {perplexity_8}")
print(f"Perplexity for 9 topics: {perplexity_9}")
print(f"Perplexity for 10 topics: {perplexity_10}")
```

Perplexity for 8 topics: 248.7512807049442

Perplexity for 9 topics: 251.69924714678154

Perplexity for 10 topics: 259.95994935443525

```
[12]: # Check and print elements in processed_articles.processed_trigrams that are
      ↪ lists instead of strings
for i, doc in enumerate(processed_articles.processed_trigrams):
    if not isinstance(doc[4], str):
        print(f"Document at index {i} is not a string:")
        print(doc)
```

```
[13]: from gensim.corpora.dictionary import Dictionary
from gensim.models.coherencemodel import CoherenceModel
from gensim.models.ldamodel import LdaModel
```

```

def prepare_gensim_input(processed_docs):
    # Recursive function to flatten nested lists
    def flatten(nested_list):
        return [item for sublist in nested_list for item in (flatten(sublist)
↪if isinstance(sublist, list) else [sublist])]

    # Flatten and tokenize documents
    flattened_docs = [' '.join(flatten(doc)) if isinstance(doc, list) else doc
↪for doc in processed_docs]
    tokenized_docs = [doc.split() for doc in flattened_docs]

    # Create dictionary and corpus
    dictionary = Dictionary(tokenized_docs)
    corpus = [dictionary.doc2bow(doc) for doc in tokenized_docs]

    return tokenized_docs, dictionary, corpus

# Prepare corpus and dictionary for Gensim
tokenized_docs, dictionary, corpus = prepare_gensim_input([t[4] for t in
↪processed_articles.processed_trigrams])

def calculate_coherence_score(lda_model, texts, dictionary, corpus):
    coherence_model_lda = CoherenceModel(model=lda_model, texts=texts,
↪dictionary=dictionary, coherence='c_v')
    coherence_lda = coherence_model_lda.get_coherence()
    return coherence_lda

def build_gensim_lda_model(corpus, dictionary, num_topics):
    lda_model = LdaModel(corpus=corpus, id2word=dictionary,
↪num_topics=num_topics, random_state=42)
    return lda_model

# Build Gensim LDA models
lda_8 = build_gensim_lda_model(corpus, dictionary, 8)
lda_9 = build_gensim_lda_model(corpus, dictionary, 9)
lda_10 = build_gensim_lda_model(corpus, dictionary, 10)

# Calculate coherence scores for Gensim LDA models
coherence_8 = calculate_coherence_score(lda_8, tokenized_docs, dictionary,
↪corpus)
coherence_9 = calculate_coherence_score(lda_9, tokenized_docs, dictionary,
↪corpus)
coherence_10 = calculate_coherence_score(lda_10, tokenized_docs, dictionary,
↪corpus)

print(f"Coherence Score for 8 topics: {coherence_8}")

```

```
print(f"Coherence Score for 9 topics: {coherence_9}")
print(f"Coherence Score for 10 topics: {coherence_10}")
```

Coherence Score for 8 topics: 0.22442731580962746
 Coherence Score for 9 topics: 0.2373412069291813
 Coherence Score for 10 topics: 0.24264137239001266

1.6 Interpretation

Here the model `lda_8` has the lowest perplexity score of 248.75 and highest coherence score of 0.242. Hence, this is the best model.

1.7 Most common word in each topic

```
[15]: import matplotlib.pyplot as plt
      from wordcloud import WordCloud

      # Extract the most common word in each topic
      def get_most_common_words(lda_model, num_words):
          topics = lda_model.show_topics(formatted=False, num_words=num_words)
          common_words = {}
          for topic_num, topic in topics:
              common_words[topic_num] = [word for word, _ in topic]
          return common_words

      # Get the most common words for each topic
      num_words = 10
      common_words = get_most_common_words(lda_8, num_words)

      # Print the most common words for each topic
      for topic_num, words in common_words.items():
          print(f"Topic {topic_num}: {' '.join(words)}")
```

Topic 0: use, copyright_holder, sequenc, case, infect, report, model, patient, studi, includ
 Topic 1: use, copyright_holder, infect, case, model, data, time, sequenc, studi, epidem
 Topic 2: use, case, copyright_holder, sequenc, model, gene, data, studi, time, cell
 Topic 3: use, copyright_holder, infect, gene, also, model, case, outbreak, studi, data
 Topic 4: use, copyright_holder, infect, data, case, model, gene, report, also, studi
 Topic 5: use, case, infect, data, model, copyright_holder, gene, also, number, cell
 Topic 6: infect, use, copyright_holder, case, epidem, cell, differ, data, viru, model
 Topic 7: use, case, data, copyright_holder, gene, infect, report, model, also,

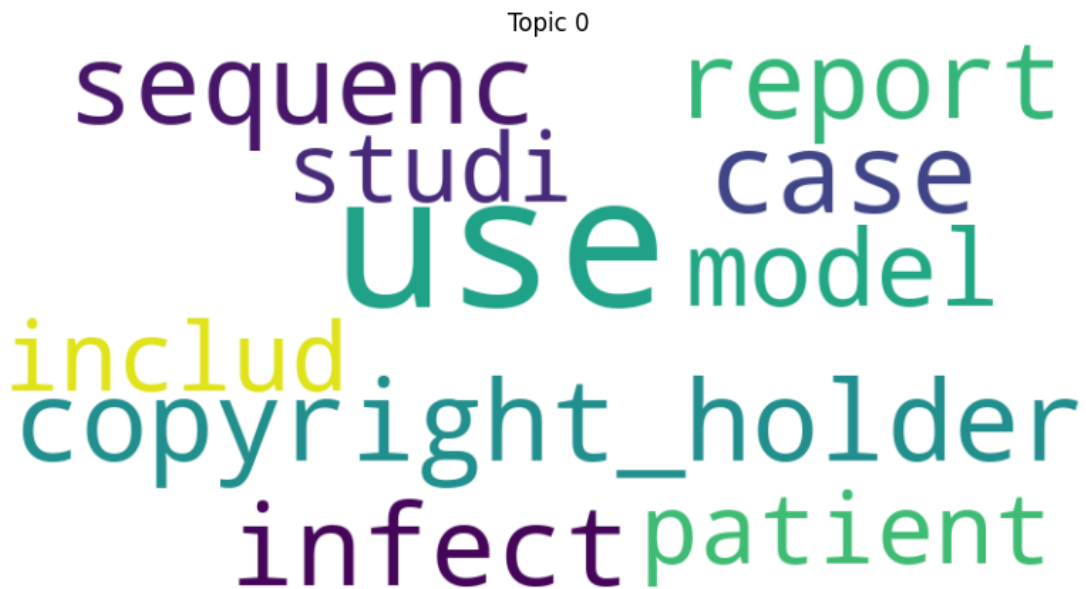
result

1.8 Word Cloud

```
[17]: # Create word clouds for each topic
def create_word_clouds(lda_model, num_words):
    topics = lda_model.show_topics(formatted=False, num_words=num_words)
    for topic_num, topic in topics:
        word_freq = {word: freq for word, freq in topic}
        wordcloud = WordCloud(width=800, height=400, background_color='white').
            generate_from_frequencies(word_freq)

        plt.figure(figsize=(10, 5))
        plt.imshow(wordcloud, interpolation='bilinear')
        plt.title(f"Topic {topic_num}")
        plt.axis("off")
        plt.show()

create_word_clouds(lda_8, num_words)
```



Topic 1



A word cloud for Topic 1 featuring various terms in different colors and sizes. The words are: time, infect, sequenc, copyright_holder, use, epidem, case, model, data, and studi.

time
infect
sequenc
copyright_holder
use
epidem
case
model
data
studi

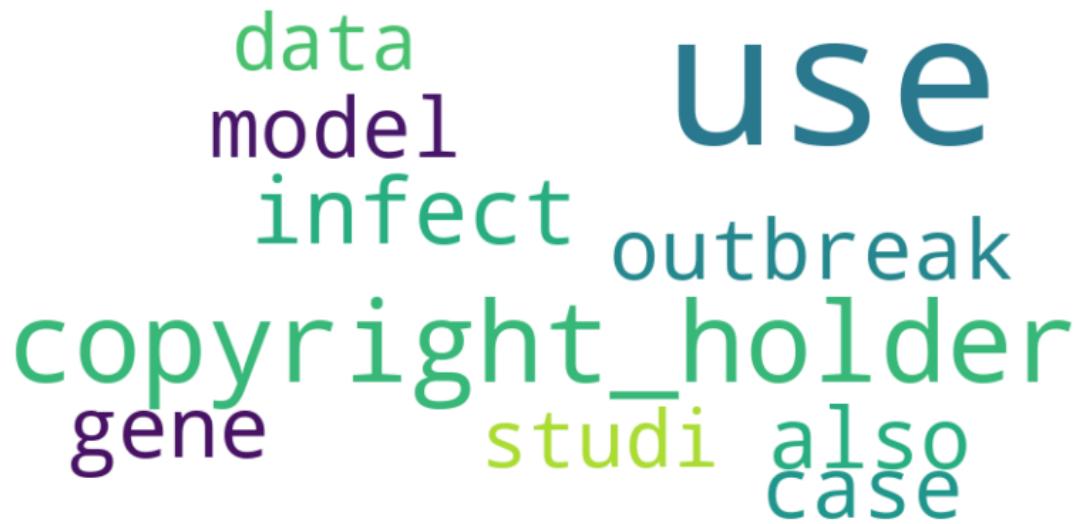
Topic 2



A word cloud for Topic 2 featuring various terms in different colors and sizes. The words are: model, use, gene, sequenc, time, cell, copyright_holder, studi, case, and data.

model
use
gene
sequenc
time
cell
copyright_holder
studi
case
data

Topic 3



A word cloud for Topic 3. The words are arranged in a cluster. 'use' is the largest word in dark blue. 'copyright_holder' is a large green word. 'outbreak' is a medium-sized teal word. 'infect' is a medium-sized green word. 'model' is a medium-sized purple word. 'data' is a medium-sized green word. 'gene' is a medium-sized purple word. 'studi' is a medium-sized yellow-green word. 'also' is a medium-sized teal word. 'case' is a medium-sized teal word.

data
model
infect
outbreak
copyright_holder
gene
studi
also
case
use

Topic 4



A word cloud for Topic 4. The words are arranged in a cluster. 'use' is the largest word in dark blue. 'copyright_holder' is a large purple word. 'infect' is a medium-sized green word. 'gene' is a medium-sized yellow word. 'data' is a medium-sized purple word. 'report' is a medium-sized teal word. 'also' is a medium-sized yellow word. 'studi' is a medium-sized blue word. 'case' is a medium-sized green word. 'model' is a medium-sized purple word.

report
gene
infect
data
use
copyright_holder
also
studi
case
model

Topic 5



A word cloud for Topic 5. The words are arranged in a circular pattern. The largest words are 'use' in teal and 'case' in yellow. Other words include 'infect' in teal, 'model' in blue, 'data' in purple, 'number' in purple, 'copyright_holder' in green, 'also' in blue, 'cell' in green, and 'gene' in purple.

use
case
infect
model
data
number
copyright_holder
also
cell
gene

Topic 6



A word cloud for Topic 6. The words are arranged in a circular pattern. The largest words are 'use' in teal and 'infect' in blue. Other words include 'case' in purple, 'data' in blue, 'epidem' in green, 'viro' in green, 'cell' in yellow, 'differ' in blue, 'model' in blue, and 'copyright_holder' in purple.

use
infect
case
data
epidem
viro
cell
differ
model
copyright_holder



1.9 Summary of Results

1.9.1 Topic Description:

- 0) Medical and Clinical Research: focuses on patient modeling and patient study, infectious diseases, clinical report generation.
- 1) Epidemiology: focuses on epidemiology and infectious diseases, discussing aspects such as infection cases, modeling data over time, sequencing studies, and the spread of epidemics.
- 2) Genetics: focuses on study of genetic modeling and sequencing. cellular aspect of gene study and modeling over time
- 3) Epidemiology: focuses on study and modeling of infectious diseases and their outbreaks, and genetic factor impacting infectious diseases.
- 4) Infectious diseases: focuses on study and reporting of infectious diseases, influencing of genetic factors, case studies and modeling, report generation
- 5) Infectious diseases: focuses on modeling and analysis of infectious disease cases based genetic and cellular data
- 6) Epidemiology: focuses on study of epidemics and viral infections, based on cellular differences and data modeling
- 7) Infectious Diseases: focuses on the reporting and modeling of infection cases, with an emphasis on genetic data and study results

Based on the individual studies, the common title can be **Influence of Genetics on Infectious Diseases**

[]: