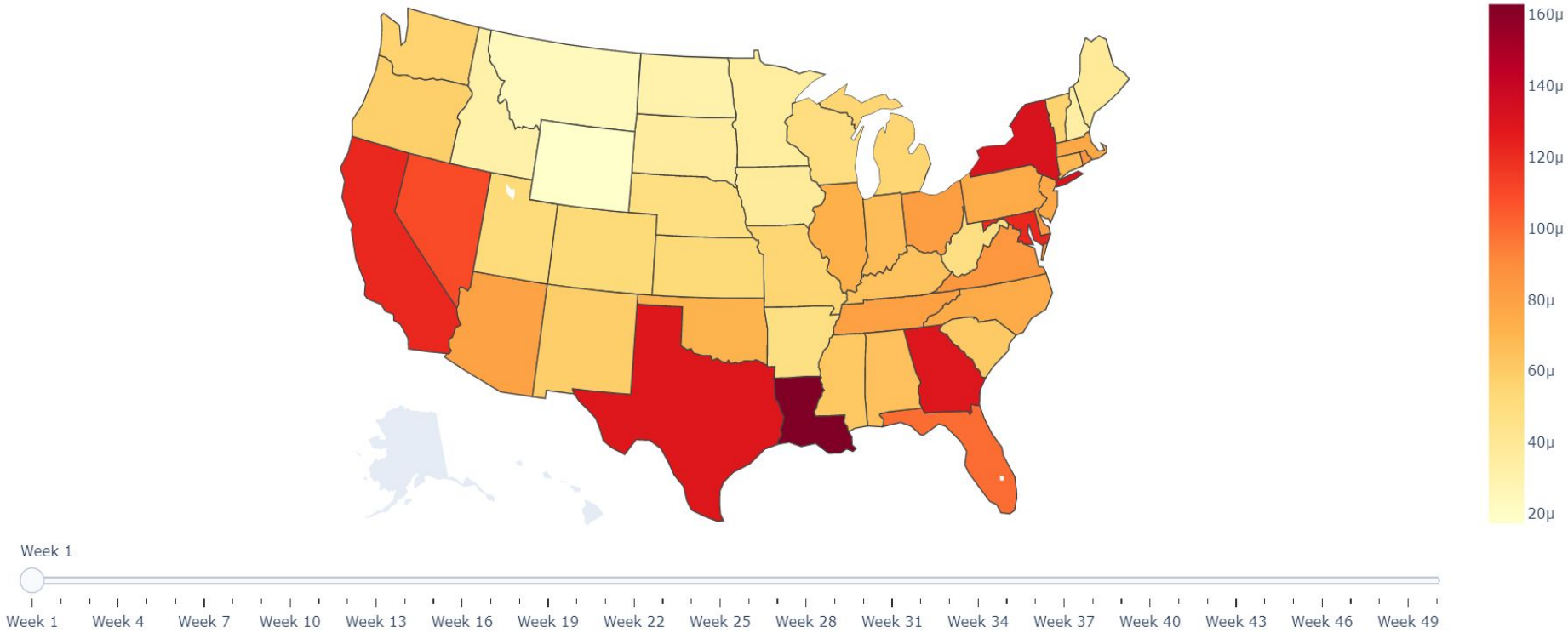
Covid19 Twitter Project

```
Summary
=======

This Twitter dataset is crawled from ids provided in the TweetsCOV19 dataset <https://data.gesis.org/tweetscov19/>.


The data are contained in the files `tweets.txt`, `users.txt`, `places.txt` and `media.txt`. More details about the contents and use of
all these files follows.


For questions, please contact Mohamed Reda Bouadjenek <rbouadjenek@gmail.com>.

Content and Use of Files
========================

Formatting and Encoding
-----------------------

The dataset files are formatted using the JSON file format (https://en.wikipedia.org/wiki/JSON) where each row is one instance. These
files are encoded as UTF-8.


Tweets File Structure (tweets.txt)
----------------------------------------

All tweets are contained in the file `tweets.txt`. Each line of this file represents the information crawled for one tweet id, and has
**mainly** the following JSON keys:

    id: the tweets id
    author_id: the id of the user who posted the tweet
    created_at: creation date of the tweet
    text: the content of the tweet
    public_metrics: public metrics like retweet count, reply count, like count, and quote count
    entities: entities mentionned in the text including mentions and hashtags
    geo: which gives the id of the place that can be retrieved from 'places.txt'.
```

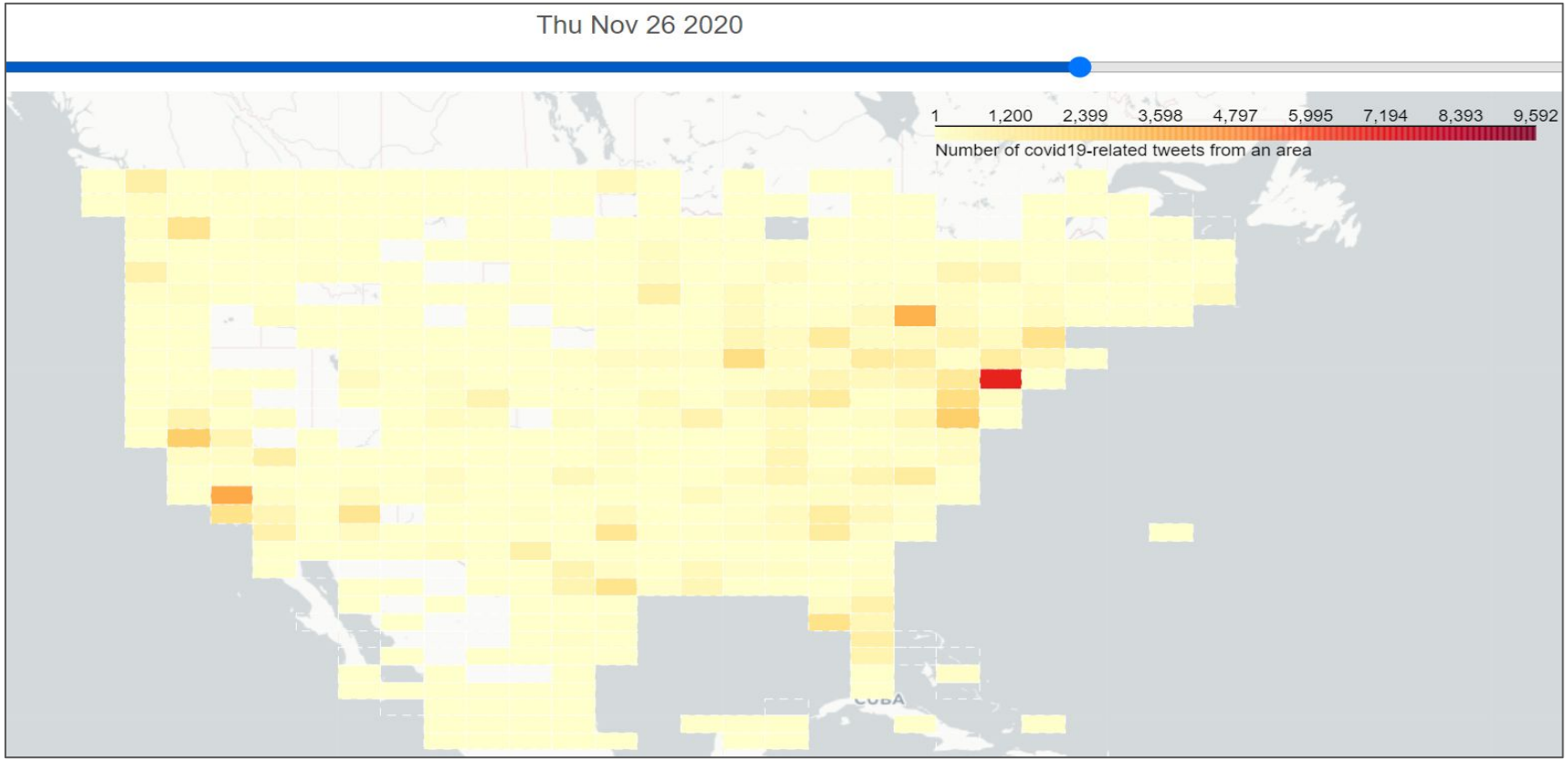
Hshmat Sahak
Litao (John) Zhou

Time Series Visualizations



Weekly Twitter Count by State (Normalized by Population)

Grid Map with Time Slider



Same data as previous slide, but different visualization approach. This is grid-wise; not state-wise

Sentiment Analysis

Current Strategy:

Remove election-related tweets:

- Trump|Biden|Election|democratic|republican|party|President|campaign|elector|candidate

Then divide into:

Family-related Tweets

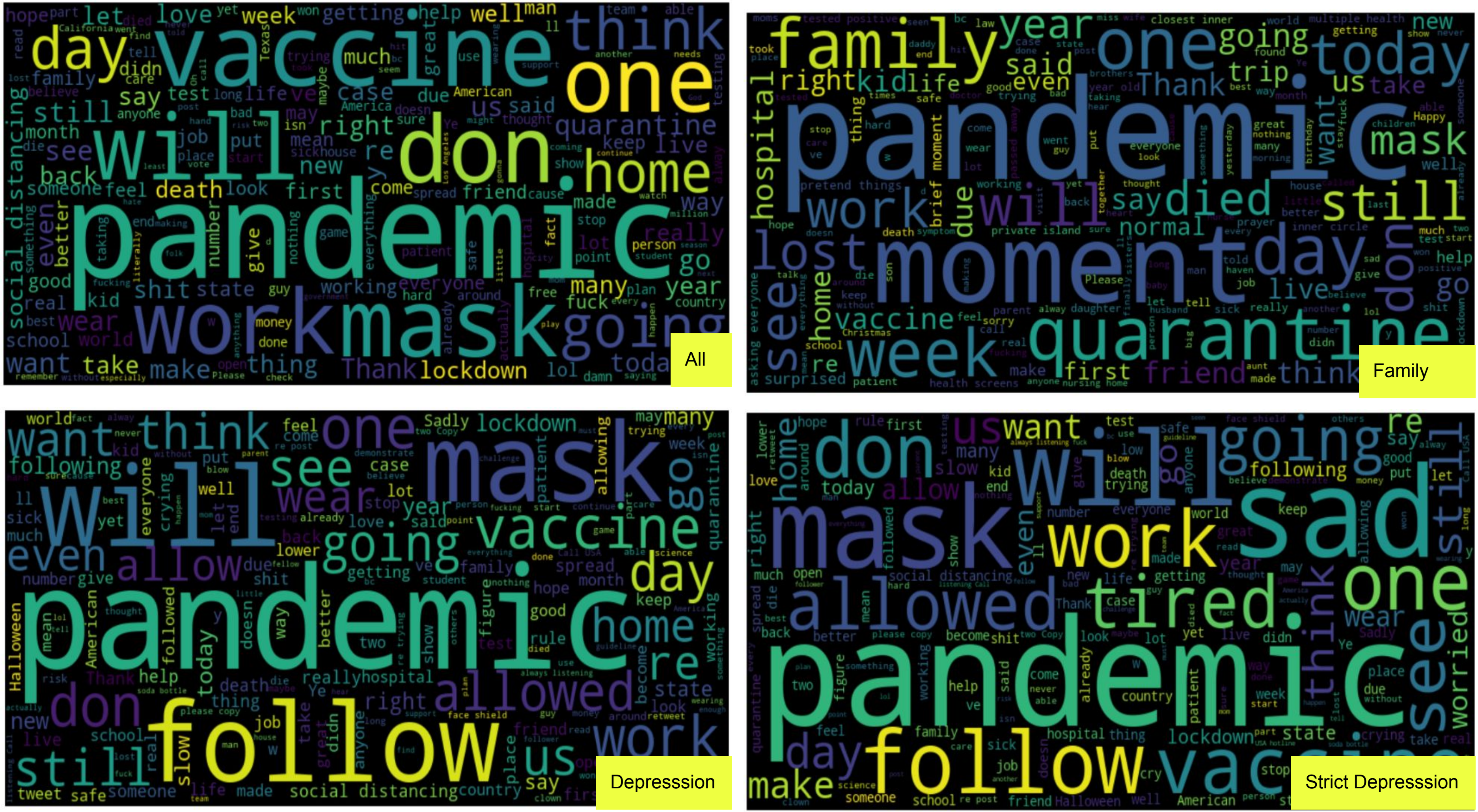
- sister|brother|mother|father|grandpa|grandma|grandparents|grandmother|grandfather|cousin|dad|mom

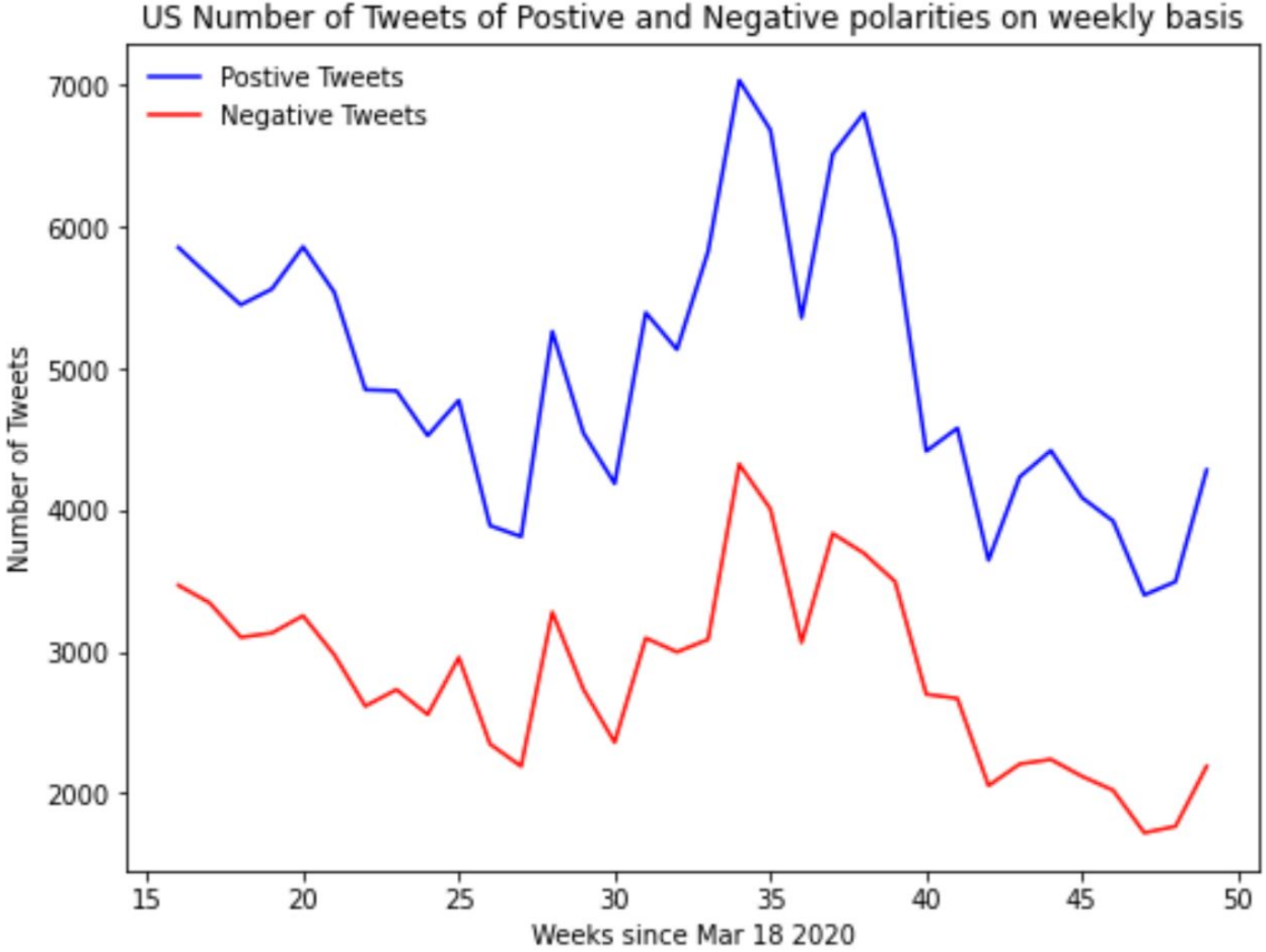
Depression-related Tweets

- overwhelmed|exhausted|distressed|anxiety|anxious|tired|low|depression|discouraged|desperate|demotivated|insomnia|cry|nervous|worried|lonely|sad|empty|suicide|antidepressant|hopeless

Strict Depression-related Tweets

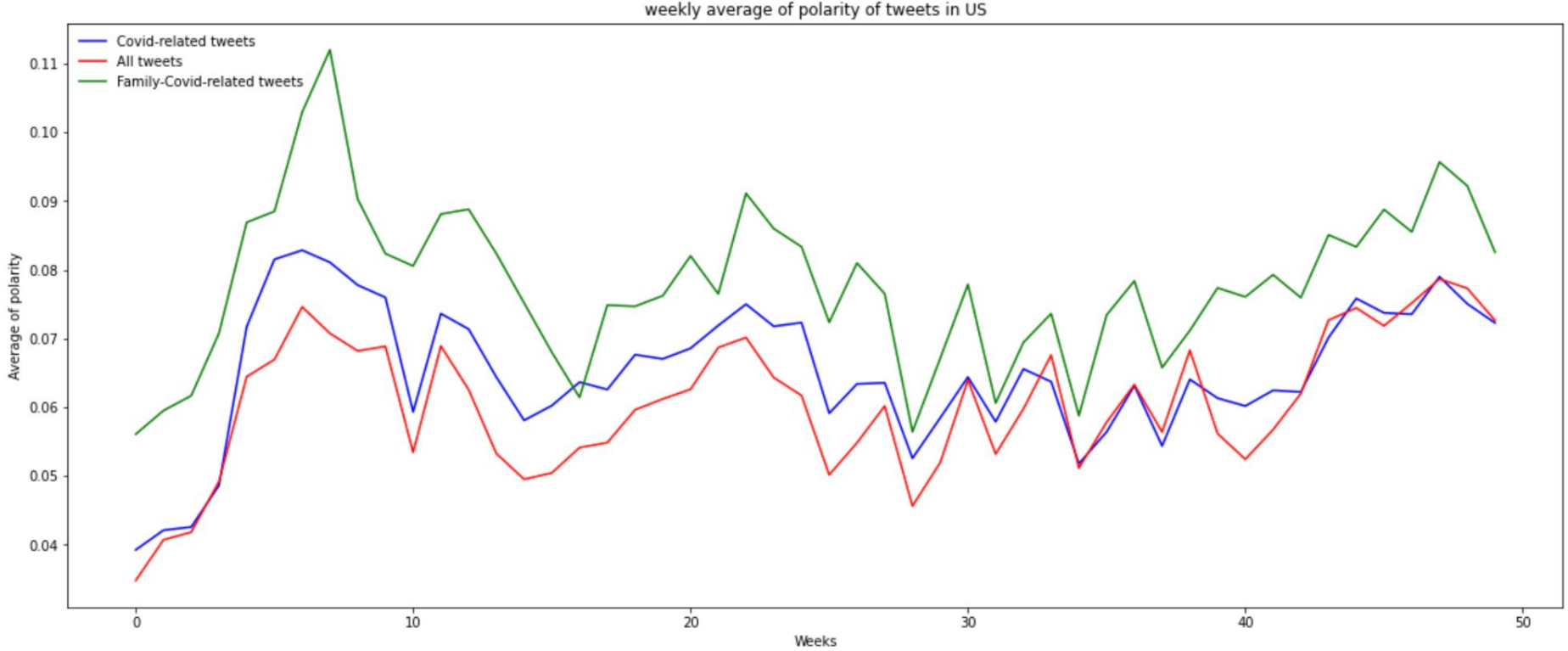
- overwhelmed|exhausted|distressed|anxiety|anxious|depression|discouraged|demotivated|insomnia|lonely|empty|suicide|antidepressant|hopeless


All
Family
Depresssion
Strict Depresssion

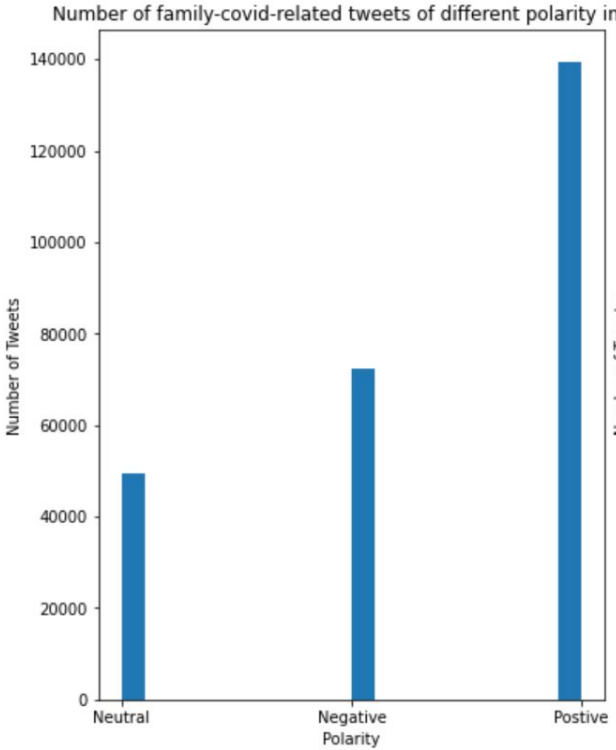

US Number of Tweets of Postive and Negative polarities on weekly basis
Postive Tweets
Negative Tweets
Number of Tweets
7000
6000
5000
4000
3000
2000
15
20
25
30
35
40
45
50
Weeks since Mar 18 2020

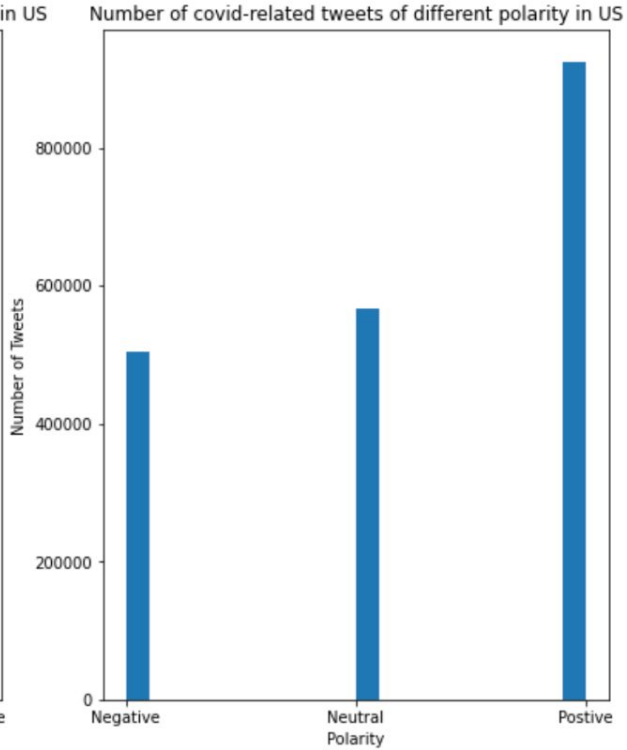
Filters for topics of family,covid and us election

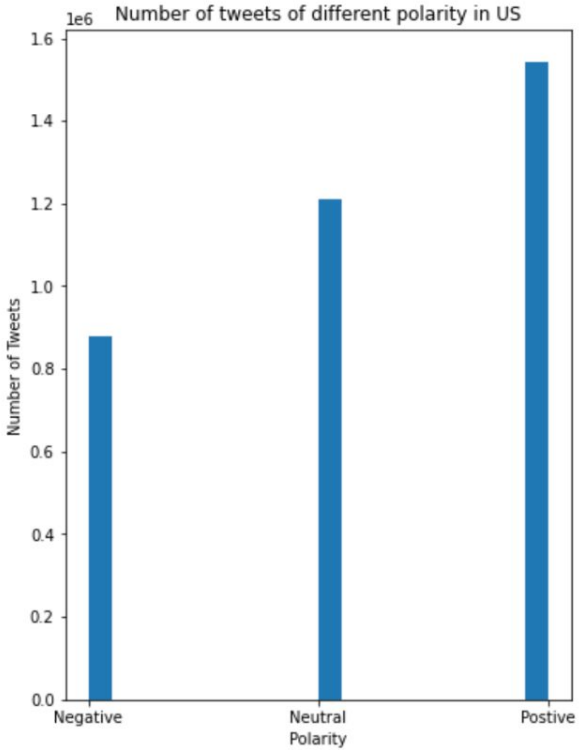
[3]:
```
#use regex to clean and filter out the useful info from tweets related to family
#also filter out the word related to us election
regex_election = re.compile(r'(?i)Trump|Biden|Election|democratic|republican|party|President|campaign|elector|candidate')
regex_family = re.compile(r'(?i)sister|brother|mother|father|grandpa|grandma|grandparents|grandmother|grandfather|cousin|child|dad|mom|uncle|aunt|nephew
regex_covid = re.compile(r'(?i)corona|covid|virus|pandemic|epidemic|quarantine|lockdown|social distancing|isolation|mask|infectious|formite|vanccine|vac


def covid_related_text(text):
    if regex_covid.search(text):
        return True
    return False

def family_related_text(text):
    if regex_family.search(text):
        return True
    return False

def election_related_text(text):
    if regex_election.search(text):
        return True
    return False

total_df['election_related'] = total_df['processed_text'].apply(election_related_text)
total_df['covid_related'] = total_df['processed_text'].apply(covid_related_text)
total_df['family_related'] = total_df['processed_text'].apply(family_related_text)
```


weekly average of polarity of tweets in US
Covid-related tweets
All tweets
Family-Covid-related tweets
Average of polarity
0.11
0.10
0.09
0.08
0.07
0.06
0.05
0.04
0
10
20
30
40
50
Weeks

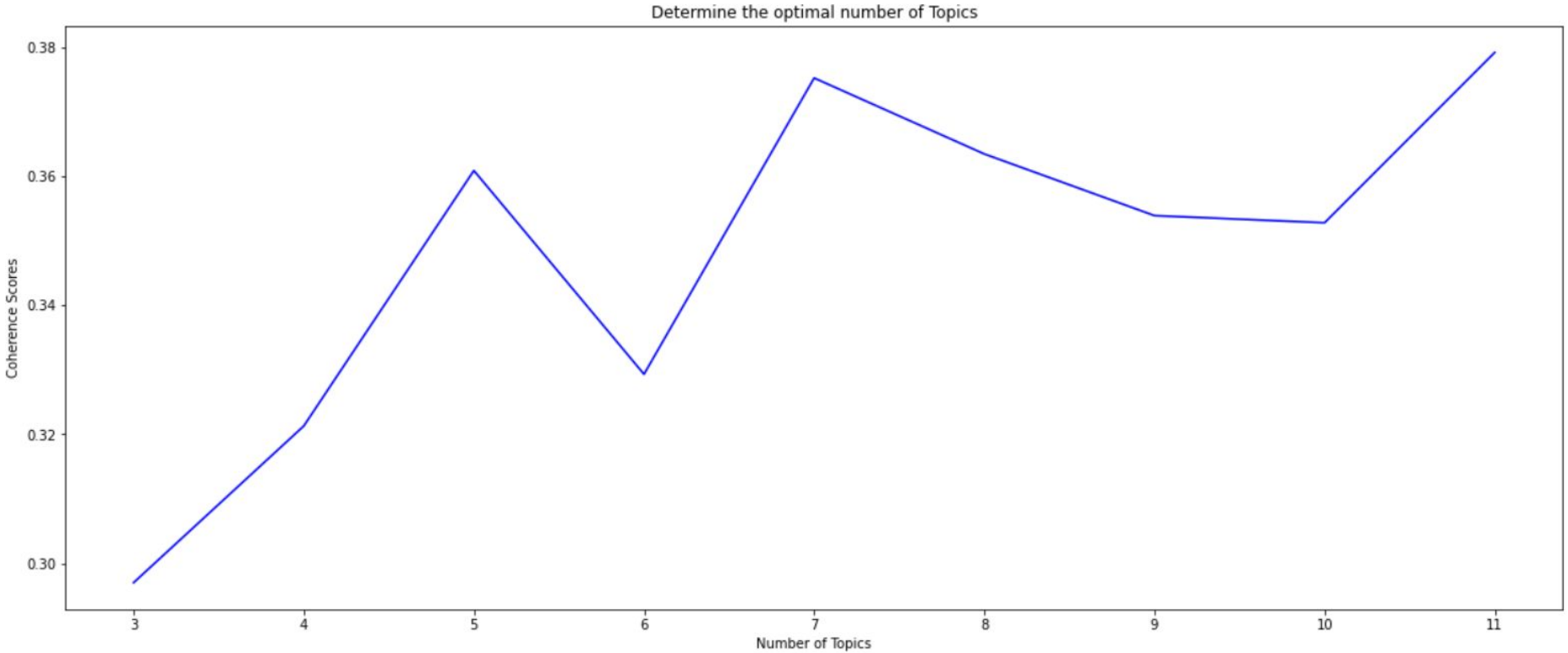
Textblob toolkit for tweets polarity analysis





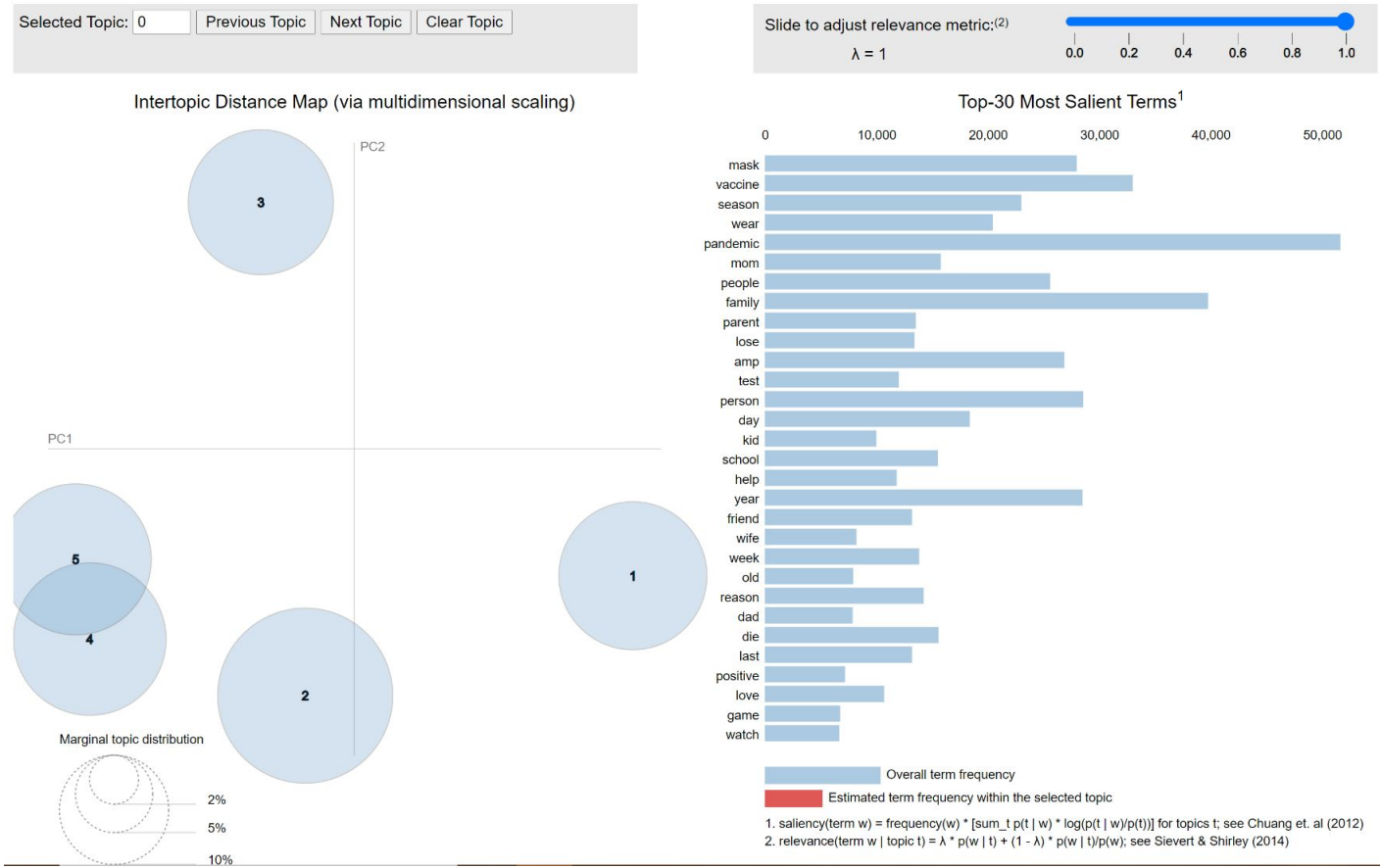


Optimal Number of Topics

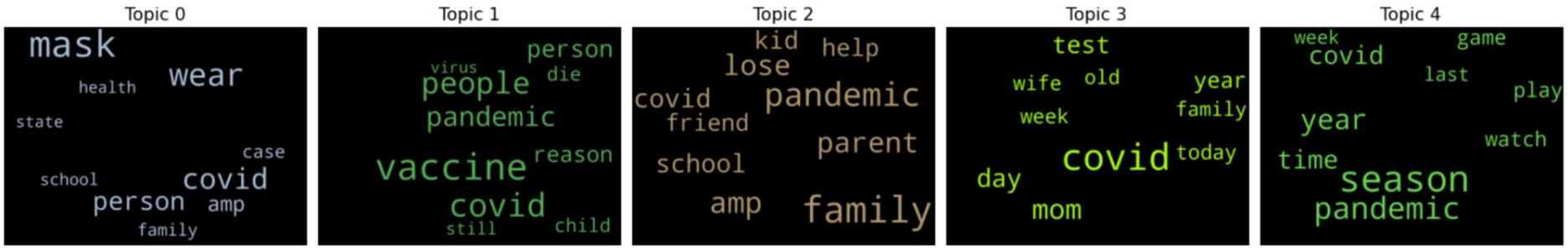


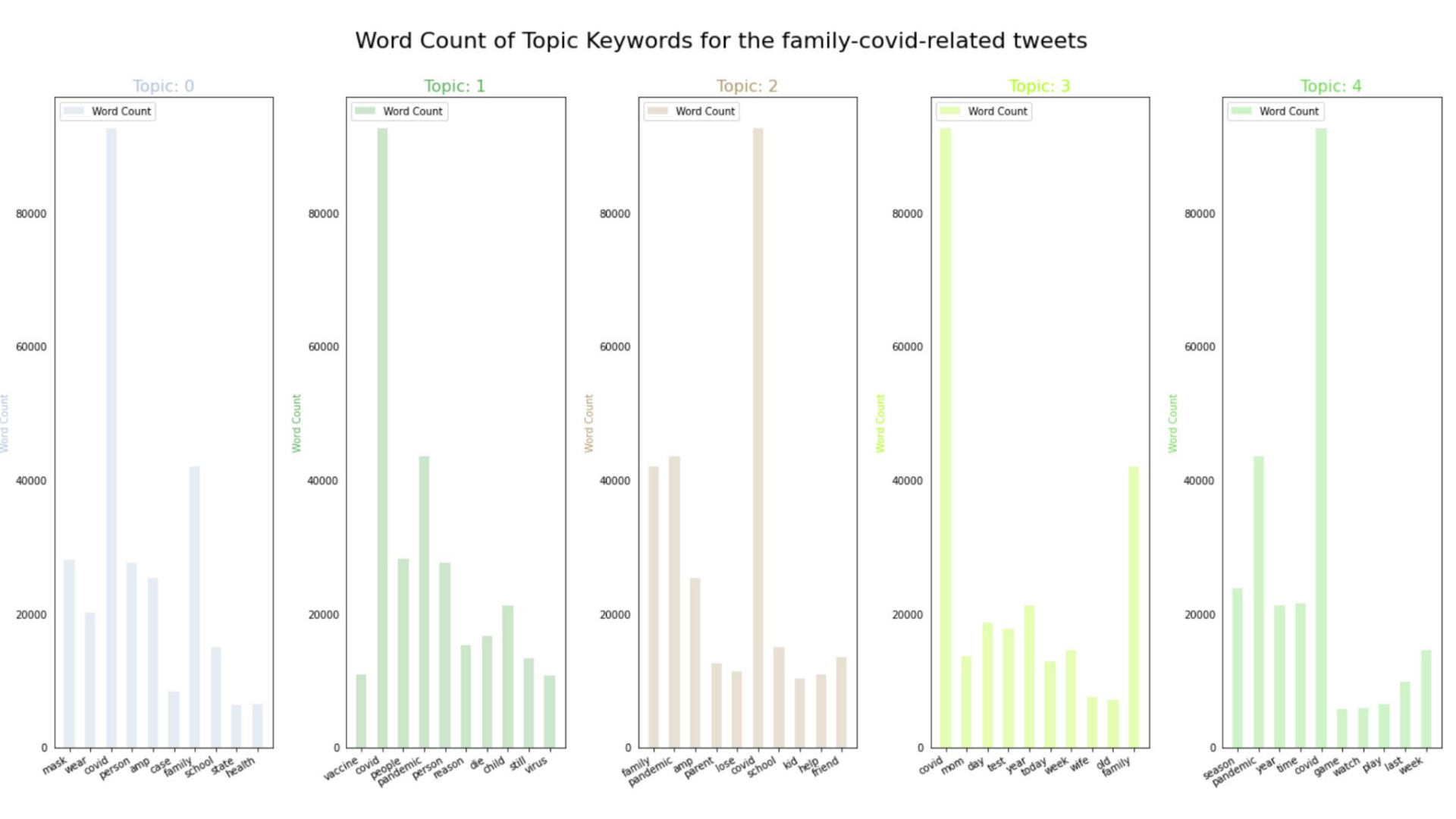
Pick 7 as optimal number of topics

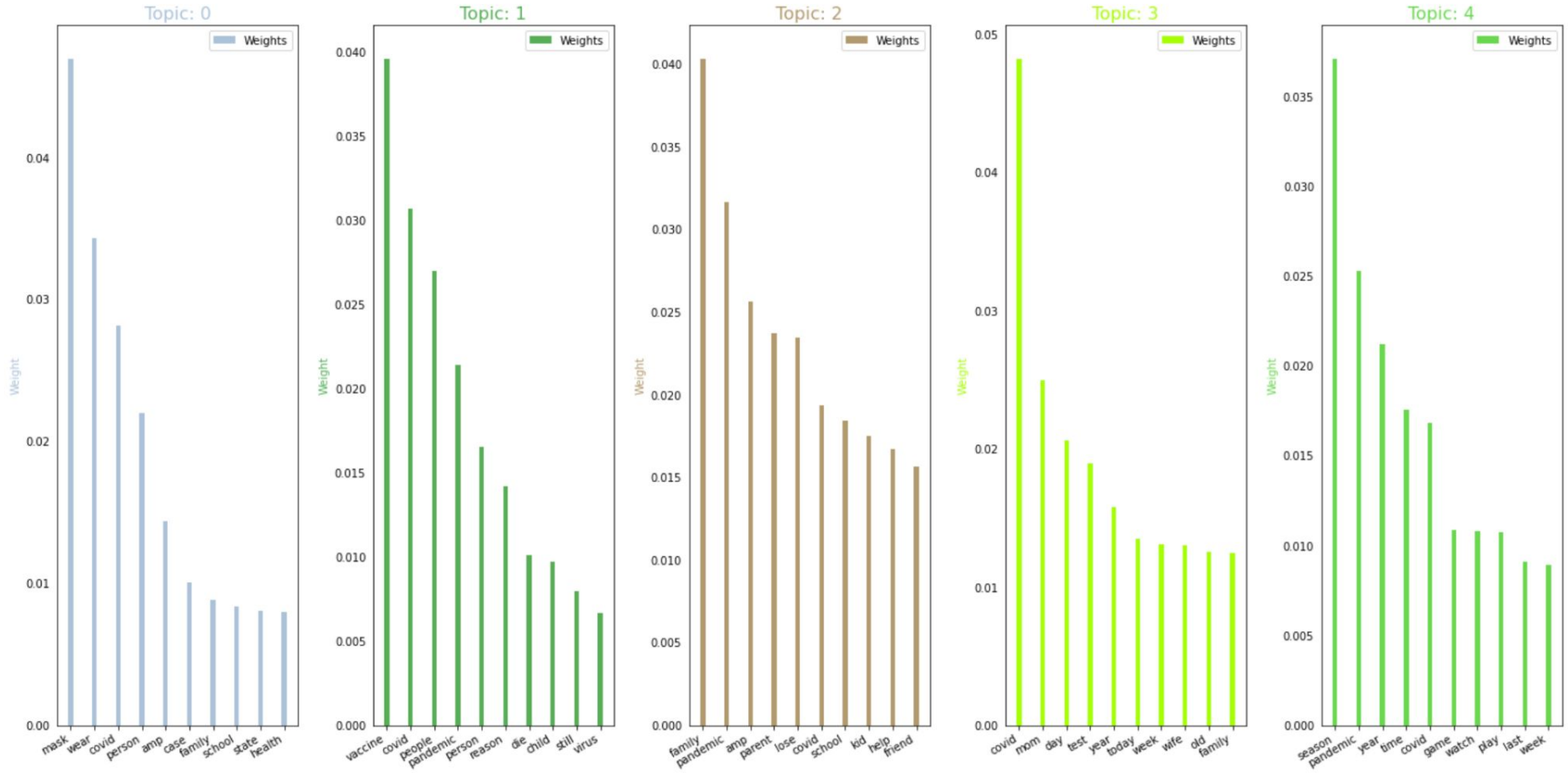
Topic Modelling - Family covid-related



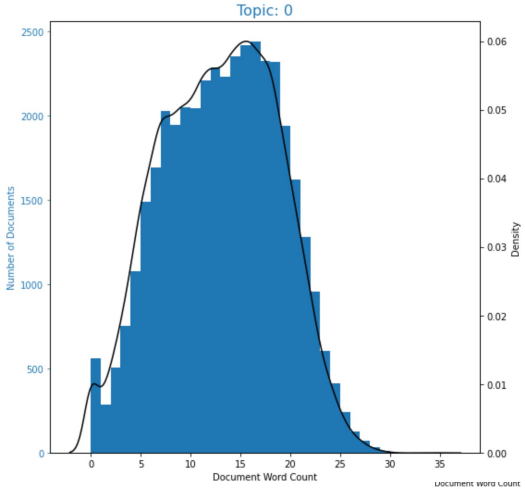
Word Cloud for Topic Modelling - Family-covid-related

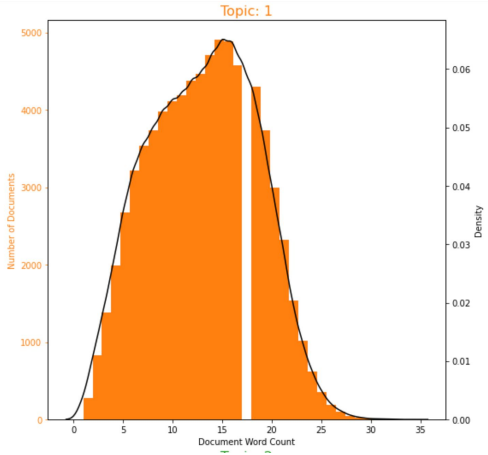


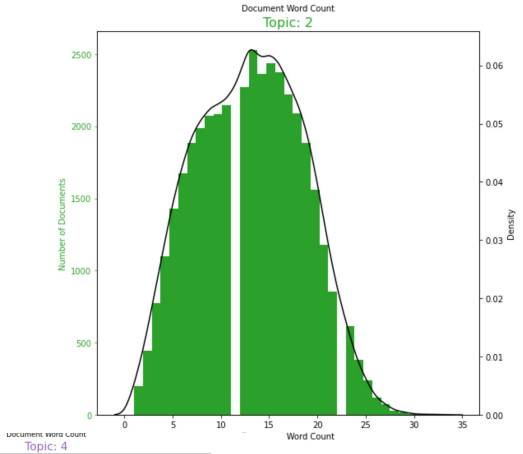

Word Count of Topic Keywords for the family-covid-related tweets
Topic: 0
Topic: 1
Topic: 2
Topic: 3
Topic: 4
Word Count
Word Count
0
20000
40000
60000
80000
mask wear covid person amp case family school state health
vaccine covid people pandemic person reason die child still virus
family pandemic amp parent lose covid school kid help friend
covid mom day test year today week wife old family
season pandemic year time covid game watch play last week

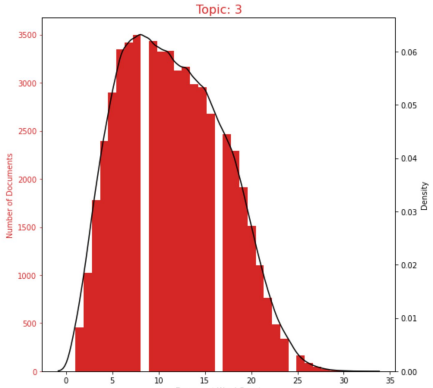

Word weight of Topic Keywords for the family-covid-related tweets
Topic: 0
Weights
Weight
mask wear covid person amp case family school state health
Topic: 1
vaccine covid people pandemic person reason die child still virus
Topic: 2
family pandemic amp parent lose covid school kid help friend
Topic: 3
covid mom day test year today week wife old family
Topic: 4
season pandemic year time covid game watch play last week

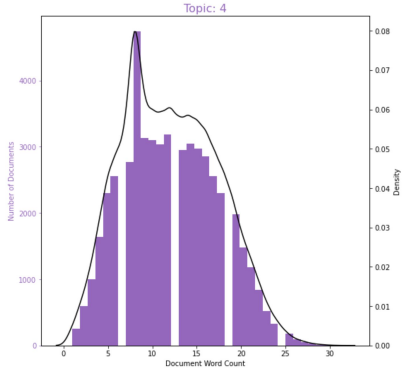
Frequency of word count in each documents



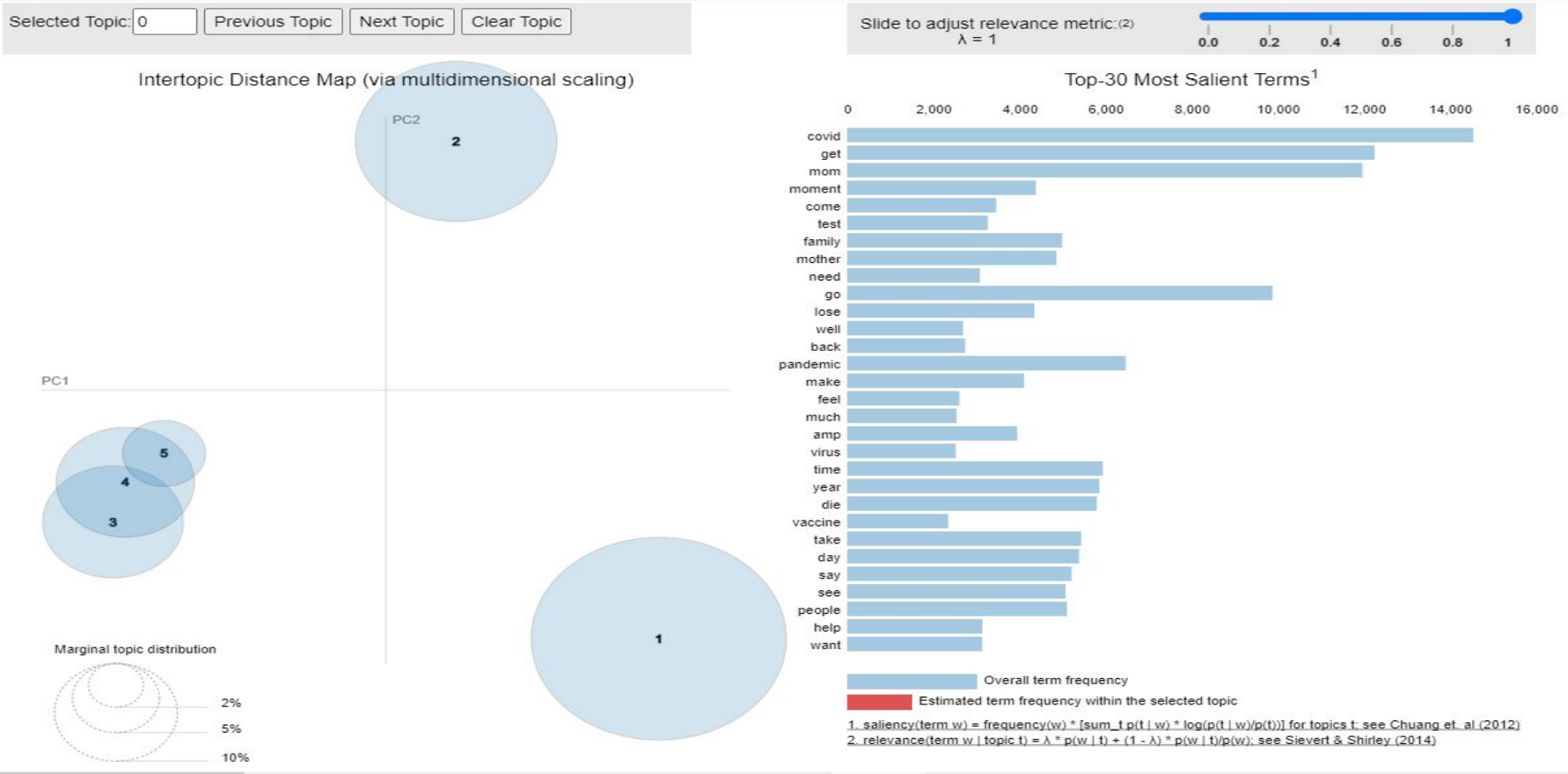








Topic Modelling- Strict Depression



What we need help with

- What particular problem should we try to approach that hasn’t been covered by others? (or, that we can do better)
  - Publication is a goal here…
- What better ways can we detect depression
  - Something better than just keywords…
  - Reference papers use keywords or train ML algorithms. We could get advice from a mental health expert...