Emotional Analysis Via Content Creation on YouTube During Coronavirus

Nguyen Ha, Computer Science

Cat Mai, Computer Science

Clark University

Abstract

In the recent years, YouTube has acquired an enormous viewership that warrants being one of the busiest sites on the Internet. Such content diversity gives rise to many intriguing research questions. In this project, we are interested in the emotional response and coping mechanism of content creators on YouTube. During the data collection phase, we have collected over 4700 and analyzed over 1600 YouTube videos related to the COVID-19 pandemic. The approach is to analyze the description and hashtags of the video and use a classifier to create a predictive tagging model. We use supervised machine learning algorithm to classify the data and run analysis on them. Natural language processing ideas are also utilized during the project. We conclude insights about content creators' response during the pandemic and the statistics of them.

Background

In goal 3 of UN sustainability goals, it has committed to “reduce by one third premature mortality from noncommunicable diseases through prevention and treatment and goa promote mental health and well-being" (United Nations). That means to improve mental health of people around the word. As The COVID-19 pandemic is causing fright across the globe and worsening mental health issues, it is even more important to take care of mental well-being. While studying some techniques of dealing with crisis in the pandemic, this project collects and studies data from an online video-sharing platform, YouTube. This platform is valuable in gaining public opinions and reactions of the crisis. Also, people who are dealing with mental issues might find some comforting approaches to distract themselves or to put their head at ease. Hence, YouTube provides these contents and even with a greater quantity and diversity during this pandemic. Thus, it is useful for analysis studying what kind of coping mechanism people have in the time of the pandemic.

Introduction

During this coronavirus pandemic, it is common to find people around the world in fear and stress, especially adolescents who are inexperienced to face such disruption. Feeling of financial despair, loneliness, boredom, and anxiety can all have effects on the state of mental health. In response to this situation, it is common for anyone to have fight-or- fly response. Fight-or- fly response is a psychological reaction happens when people are facing life-threatening situations. This biological mechanism triggers to help people to fight the threat off or flee to safety (Cherry 2019). Thus, this project tries to understand how adolescents fight-or- fly in this situation. Isolation method to prevent the spreading of the virus is important during these times is crucial; therefore, people are shifting to the Internet and social media to connect with

others and enhance their relationships in this solitude state. Especially, YouTube has seen a dramatic increase in viewership (Romero 2020). Thus, in adapting to this novel time, YouTube creators are creating more contents to continue entertain their fans and new viewers (Alexander 2020). As YouTube has seen a thrive in quantity and diversity from content creators, this platform fulfills all aspect required in data of this study.

The initial conjecture before data analysis suggests that there would be more fight videos than fly videos. This hypothesis is formed on the ground that only creators' contents will be analyzed, which means all news channels and other informative channels, such as health organizations, will be filtered out. Thus, this leads to another hypothesis that amateur creators will dominate in term of quantity. Creators, including amateur, professional, and celebrity, are categorized based on number of subscribers. To become a professional, a YouTube channels should have over 10,000 subscriber and should surpass 1 million subscribers to become a celebrity. Based on the earlier assumption, it also predicts that fly videos mainly serve as emotional supports for mental well-being, as opposed to instruction or appeal to action to help the community.

In order to categorize videos into fight or fly, data analysis runs classification algorithm called Naive Bayes classifiers, which is a supervised machine learning model. Thus, the classifier is trained by 144 data input tagged manually. The analysis also includes ANOVA models to find the difference in the means of likes and views between the two predicted category groups, namely, "fight" and "fly." The result suggests that there are indeed more fly videos than fight videos. In addition, most videos creators post fly videos are content creators, which agrees with the hypothesis. However, as the hypothesis predicts that videos do not have contents

instruction or appeal to action to help the community as much as emotional support, analysis suggests that most popular contents are related to some practical activities that can do at home.

Capture

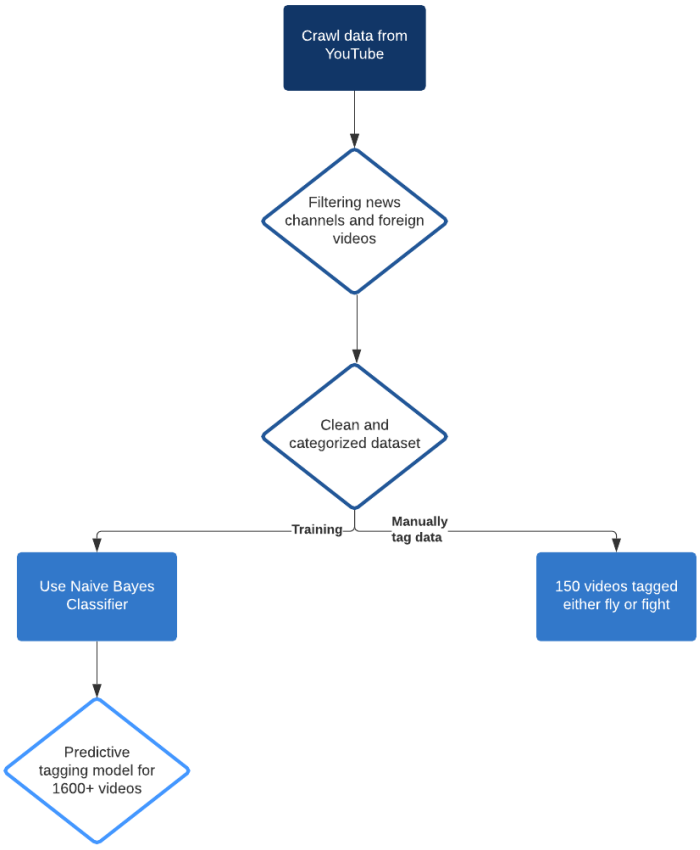
This project is looking into a platform where users can share their creativities and quickly adapt to the times. Since YouTube contents the requirements and popular around the globe, we are using YouTube data to analyze. Data is crawled from YouTube Data API using keywords “stayhome withme”, “pandemic” which are unbiased to either “fight” or “fly”, but also specific enough to find videos created during this pandemic. Data stored in three different json files for each keyword are list of dictionaries. We collect a total of more than 4700 videos and perform data analysis on more than 1600 videos. Filtered out content are those with non-ASCII description or title, have foreign language based on language detection package by Anaconda, and are from news channels.

Data retrieved is stored in json files per YouTube’s data retrieval object specification. Below is an example figure of YouTube’s API response in json.


```
{
  "kind": "youtube#searchListResponse",
  "etag": "\"m2yskBQFythfE4irbTIeOgYYfBU/PaiEDiVxOyCWelLPuuwa9LKz3Gk\"",
  "nextPageToken": "CAUQAA",
  "regionCode": "KE",
  "pageInfo": {
    "totalResults": 4249,
    "resultsPerPage": 5
  },
  "items": [
    {
      "kind": "youtube#searchResult",
      "etag": "\"m2yskBQFythfE4irbTIeOgYYfBU/QpOIr3QKlV5EUlzfFcVvDiJT0hw\"",
      "id": {
        "kind": "youtube#channel",
        "channelId": "UCJowOS1R0FnhipXVqEnYU1A"
      }
    },
```

We then use chiefly the pandas package to pickle data and run analysis.

Process



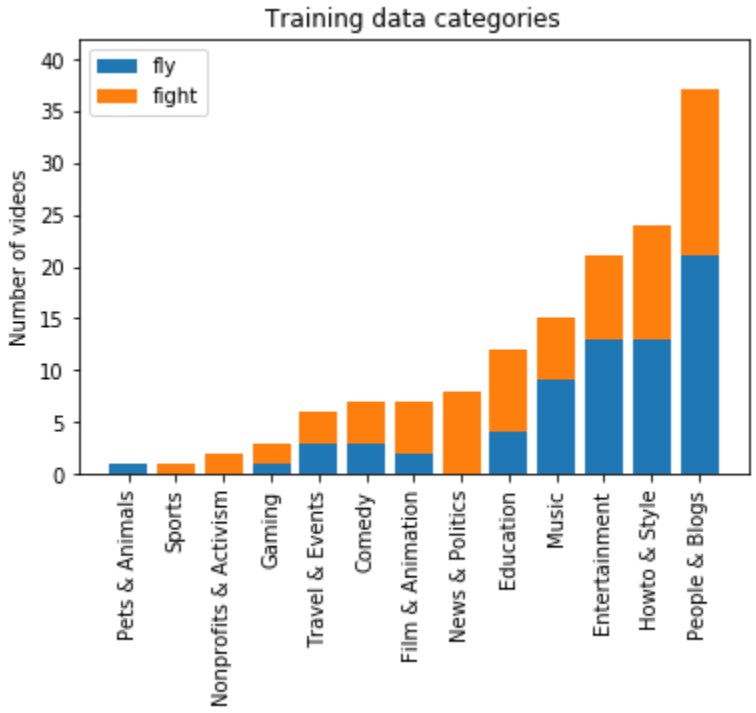
The flowchart above is shown to demonstrate our pipeline from crawling data, filtering videos, to creating a machine learning predictive model.

We use Gaussian naïve Bayes classifier to approach tagging the data, where each x a keyword associated with the description. We perform the training on a 150 manually tagged dataset with 2 labels, namely, fight and fly. From there, we use apply the following formula to create the classifier. The smoothing factor is $k = 1$.

$$P(y|x_1,...,x_n) = \frac{P(x_1|y)P(x_2|y)...P(x_n|y)P(y)}{P(x_1)P(x_2)...P(x_n)}$$

We choose Naïve Bayes model instead of k-Means or KNN algorithms because firstly, for a relatively small dataset like ours, we believe a supervised machine learning algorithm would be more fitting. Secondly, the analysis is heavily text-based and Naïve Bayes is undoubtedly a better tool for this.

The figure below shows the statistics on our preliminary training dataset i.e. the manually tagged videos and their responding categories.

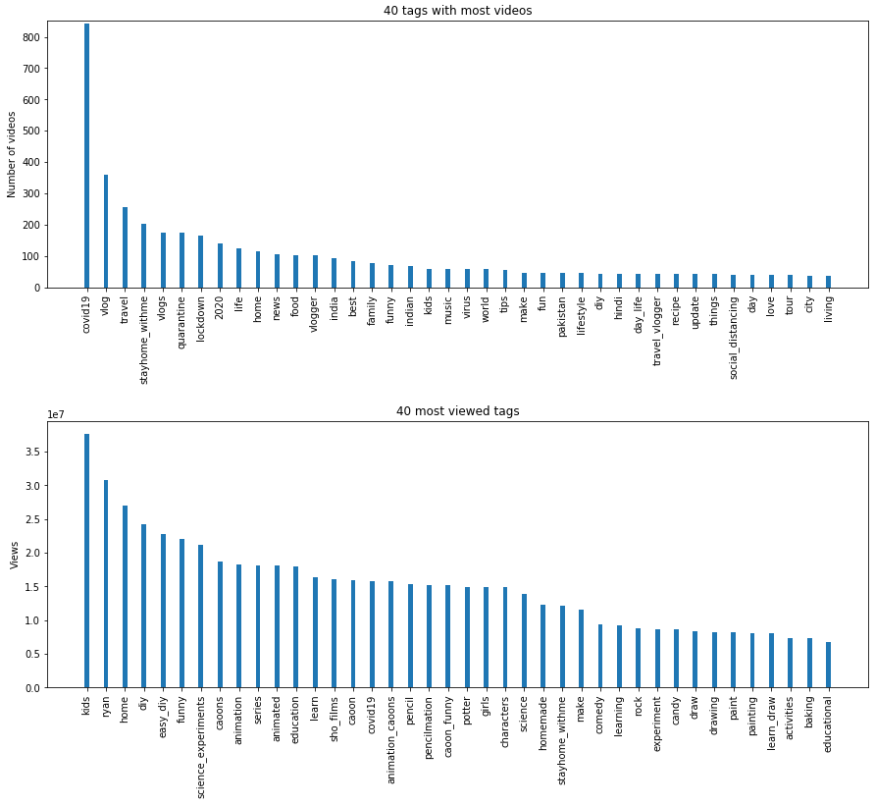


Another important part of our project is data analysis based on both clean data and predicted tags. For hashtag-specific analysis, we decide to aggregate data for two hashtags, “Covid-19” and “Stayhome withme.” Tags that are closely related or are typos or typing variants of those two, such as “Coronavirus” and “coronavirus” or “Stayhome” and “Stayhome_withme” are

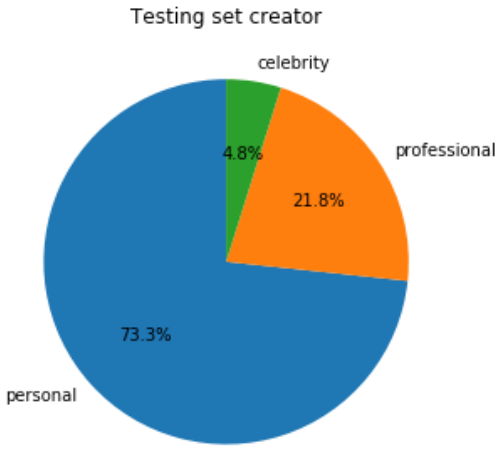
counted within their respective groups. After obtaining predictive tagging using the classifier, we explore the relationship between each tag category and the amount of likes or views they receive and accumulate statistics on most frequently used keywords.

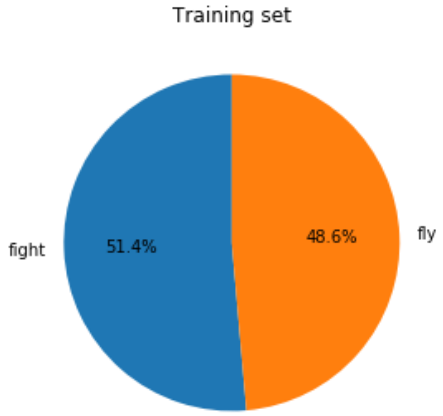
Lastly, we utilize a natural language processing package, *nltk*, to generate Cosine similarity vectors between texts and use the tokenized data to train our classifier.

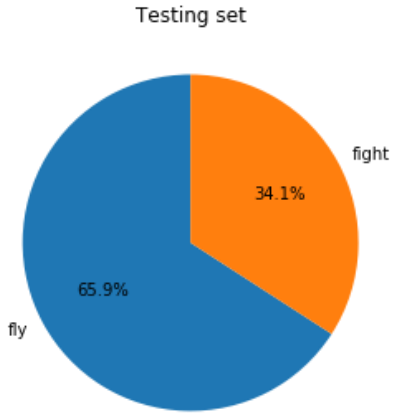
Result



In the above figures are tags with most views and most video count. As predicted in our data aggregation method, “COVID-19” and “Stayhome withme” and their variations are among tags with most video count. However, the result regarding the viewership shows some discrepancy between that of video count.



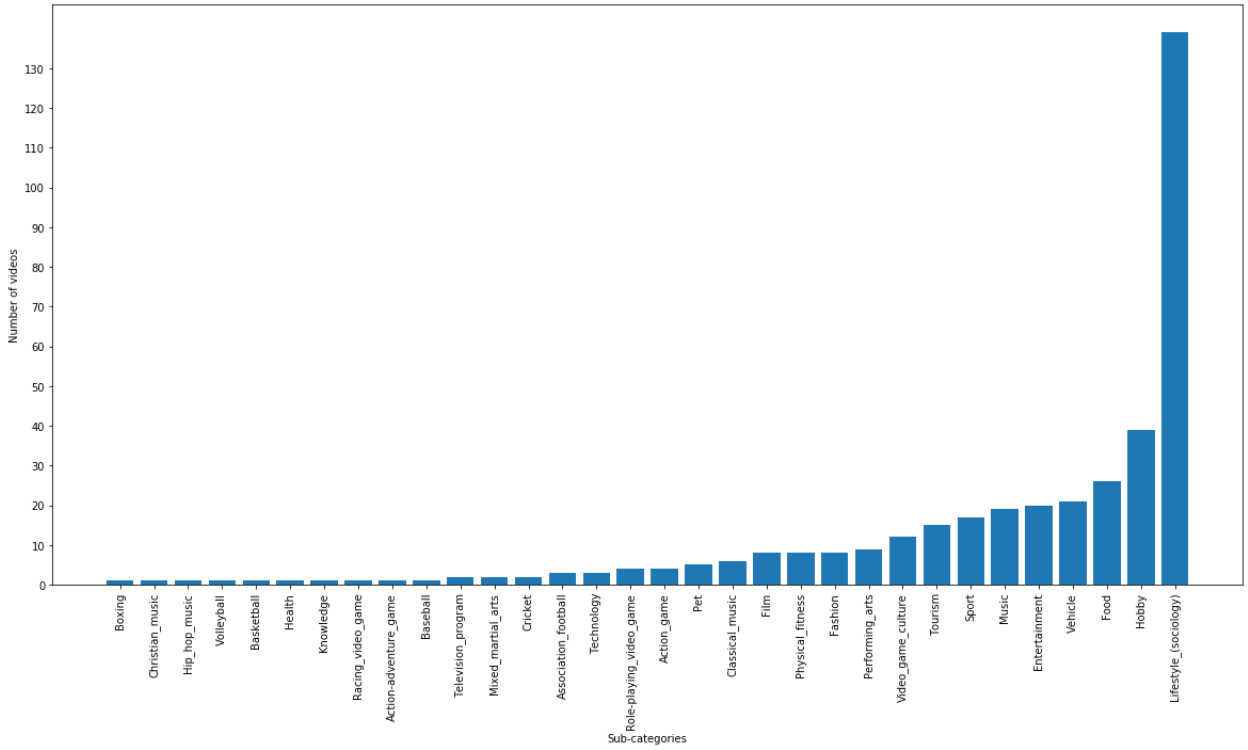




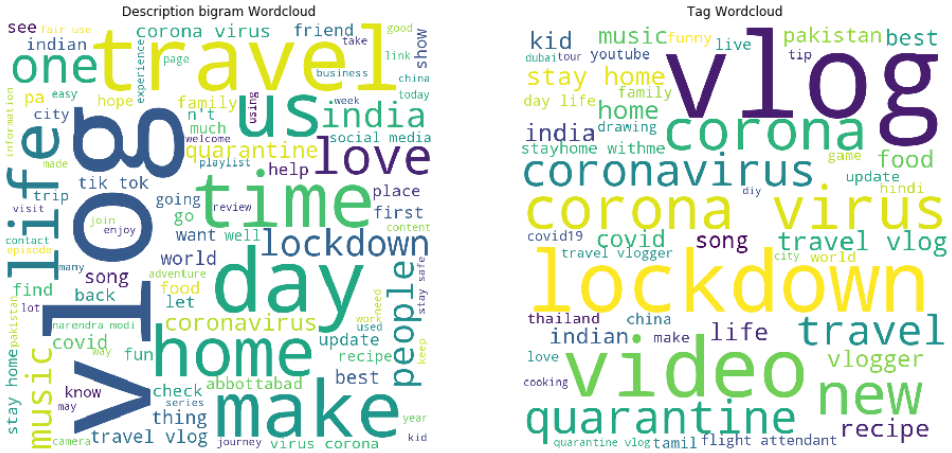
For our hypothesis testing, we test out a hypothesis that there is statistically significant difference between the two groups “fight” and “fly.” The result shows that while our initial prediction that there would be more “fly” videos than “fight,” there is not enough data to suggest that there is considerable gap between the amount of likes and views between them. We also suggest different approach in our methodology based on current results. Running ANOVA Type II on predicted subcategory i.e. fly vs fight, and number of views and likes, we got respective p-values of 0.967 and 0.478. This means the null hypothesis is not rejected and there is not enough difference in means between the two predicted subcategory and their views or likes.

	Average view	Average like
Fly	74844.1	1278.57
Fight	73645.2	1020.81

We reckon the disparity in viewership between different celebrity status and will change our methodology for subsequent research projects, for example, by performing hypothesis testing within each celebrity group or grouping the audience and/or creators through different means.



The most popular subcategories for videos that are predicted “fight” are visualized in the figure. We conclude that “Lifestyle,” “Hobby,” and “Food” are the most preferred subcategory for content creators during this time, with videos related to “Lifestyle” having a noteworthy number of videos. This is what we predicted before running the analysis, because we believe most videos created during this time tend to have more in the subject of recipes or crafts-making.



The figure shows most frequently used tags and most frequent tokens in video's description.

Discussion

As the pandemic is still an ongoing distress, early data analysis would spearhead interesting research and spark discussion within the community. Results show that most creators intend to help their viewers by performing tasks to stay healthy or showing everyday activities that people can do at home due to the isolation either through fight or fly videos. We believe this project will help us understand better about people's response and coping mechanism during an incredibly unusual time, specifically how content creators would incorporate the changing elements of the world surrounding them into their videos. We also see a great number of videos coming from amateurs, indicating that the sudden social distancing has created more of an opportunity for everyone, not just professionals or celebrities, to contribute to the community by simple cooking or crafts making instructional videos. It seems that creators and viewers uptake about the situation by routine-related contents. Thus, to get informative contents to as many viewership as possible, it should follow the trend of public interest to approach those in need.

A point of interest for the UN is to analyze country-specific policies that might have affected the content on YouTube created during this time. We can see how different creators

respond to their local policy by the same analysis principle. This could be an intriguing path worth investing time into, especially if we are interested in the political landscape shaped by YouTube's viewership. For example, recent protests are also broadcasted and aired by amateurish creators, which could be a rich source of content to analyze tags about policies.

Appendix A

	video_id	video_title	channel_id	channel_title	description
0	--ueGGPBc2U	011 Шахсий бухгалтерия Хисоботлар Харажатла...	UCnYBmZ9aSaV8JbgwsVfE1rg	LEADERS GROUP	Шахсий бухгалтерия. 1С 8.3 да узимиз учун хисо...
1	-1kTxtkwiS0	Mission: A fit and healthy Bangladesh	UC9DzYLMuT9-Knt9h9r14QxA	Zahurul Shuvo	My aim and mission to see a fit and healthy Ba...
2	-2Ya3O0WaQU	4th Week Salinas Ecuador Jan 2018 msbjpeart	UCEjpSE4pOhftupgVjQw1UGA	BRENDA J. PEART	#ConnectingtheDotsAcrosstheDiaspor \n\n\r\nht...
3	-3BBx74PEE0	BERFİN AKTAY - NEWROZ #StayHome #WithMe #Evd...	UC4L1zk8leEoJTte3XGHoVXA	Berfin Aktay	NEWROZ PÎROZ BE\nStay home and sing with me.....
4	-3GpaGQQnqY	#stayhome #withme Corona5g ? is there a scient...	UCO3sl5WQ1eDTIhykB5y-p7Q	**FREEDOM TV**	
...	...	...	...	...	...
3501	zxFJQuluq1w	Lockdown Log 17	UChIo_isfgXCDzPn2RBnPHfw	James Alan Anslow	Brief daily vlogs from an English villag duri...
3502	zydeidC3Kds	INTERNATIONAL Quarantine: Corona Virus Cancell...	UCUvKgU6SY_BhitpIVl1tsWw	Sarah Magdy	From #Egypt to #Germany to #Switzerland to #In...
3503	zz5khF9ix64	#stayhome #withme Reading time with Donna	UCUX5WMe_0FxRzQFnDkxnZNw	Donna Vee Comedy	Entertaining kids by reading so you ca have a...

Above is a DataFrame example of our collected data.

Appendix B

```
# Draw WordCloud of video description and tags

bigram=Phrases(df_testing.tokens, min_count=5)

bigrams=[bigram[tl] for tl in df_testing['tokens']]
df_testing['bigram'] = bigrams


def draw_WordCloud(df):
    bigram_list=[]
    for sent in df['tokens']:
        bigram_list.extend(sent)

    tag_list =[]
    for k in df['tag']:
        if k!= None:
            tag_list.extend(k)

    bigram_wordcloud=WordCloud(width=800, height=800,
                      background_color ="white",
                      min_font_size=20).generate(" ".join(bigram_list))

    tag_wordcloud=WordCloud(width=800, height=800,
                      background_color ="white",
                      min_font_size=20).generate(" ".join(tag_list))
```

Above is a code snippet to generate WordClouds corresponding to most frequent words used in the description and most used hashtags.

```
def filter_df(adf):
    for index, row in adf.iterrows():
        isDropped = False
        if row["default_lang"][:2]!="en" and row["default_lang"]!="NA":
            adf.drop(index, inplace=True)
            isDropped = True
            continue
        for key in filter_keyword:
            if key.lower() in row["channel_title"].lower():
                adf.drop(index, inplace=True)
                isDropped = True
                break
        if not isDropped:
            video_title = row["video_title"]
            try:
                video_title.encode('ascii')
            except UnicodeEncodeError:
                continue
            lang = detect(video_title)
            adf.loc[index]["default_lang"] = lang
    adf = adf.loc[adf["default_lang"] == "en"]
    adf.reset_index(inplace=True, drop=True)
    return adf
```

Our method “filter_df” is used to filter out videos that contain foreign characters or are from news channels.

Appendix C

Third-party packages include langdetect, gensim, WordCloud, and nltk.

References

#Envision2030 Goal 3: Good Health and Well-being Enable. (n.d.). Retrieved from https://www.un.org/development/desa/disabilities/envision2030-goal3.html

Alexander, J. (2020, March 28). "With me" videos on YouTube are seeing huge spikes in viewership as people stay home. Retrieved from https://www.theverge.com/2020/3/27/21197642/youtube-with-me-style-videos-views-coronavirus-cook-workout-study-home-beauty

Cherry, K. (2019, August 18). The Fight-or-Fly Response Prepares Your Body to Take Action. Retrieved from https://www.verywellmind.com/what-is-the-fight-or-fly-response-2795194

Romero, D. (2020, April 02). YouTube thrives as a window for those isolated by coronavirus. Retrieved from https://www.nbcnews.com/tech/social-media/youtube-thrives-window-those-isolated-coronavirus-n1173651