textual group

1. Collect video id : top 10 videos by view count from 2020-03-01 to 2020-05-31.

<https://github.com/LynetteGao/image_as_data/blob/main/video%20metadata%20collection/videos_collection.py>

Note: Change line 29 to your own api\_key

####### Obtain your YouTube API Key #############

You need to have a gmail

Then, get onto: <https://console.developers.google.com/>

Create a New Project and you can name it as YouTube COVID19

At the Dashboard page, click + ENABLE APIS AND SERVICES -> In the Search Box, type YouTube -> choose YouTube Data API v3 and after you click into it, please click the “ENABLE” button.

Go to Credentials and generate an API Key for this project. This is the API Key you will use to collect YouTube Videos.

Scholars have already written nice git to help us use the YouTube API, please check and follow instruction on this git: <https://github.com/SMAPPNYU/youtube-data-api> In this repo, the scholars from NYU developed the package youtube-data-api, which allows us to collect YouTube video data. Please read and you can see that my students had used this package during their data collection.

####### Obtain your YouTube API Key #############

* Expected similar output: <https://uwprod-my.sharepoint.com/:x:/r/personal/kchen67_wisc_edu/_layouts/15/Doc.aspx?sourcedoc=%7B05C877C1-478F-4285-9856-E04717A7BFE0%7D&file=Coronavirus%20Vaccination_May24.csv&action=default&mobileredirect=true>

At this point, we collected more ids than original conspiracy datasets. Since lots of them will be abandoned after preprocessing.

(Note: aim for around N = 299 (# conspiracy) at the end)

M1: Please share the output csv Lynette know when you generate before proceeding to the next step.

textual group

1. Subset English videos

* Input: csv from step 1 with all metadata

<https://github.com/LynetteGao/image_as_data/blob/main/video%20metadata%20collection/remove_non_english_video.R>

* Output: csv with all metadata

1. Collect transcript: textual group

* Input: csv from step 2 with all metadata

https://github.com/LynetteGao/image\_as\_data/blob/main/video%20metadata%20collection/transcript-collection.ipynb

* Output: a json file

1. processing the output transcript in json file + cleaning: textual group

<https://github.com/LynetteGao/image_as_data/blob/main/video%20metadata%20collection/process_caption.R>

* Input: json file from step 2
* Output: a csv containing all metadata + transcript info columns

M2: Please let Lynette know when you generate this output csv, Lynette need to double check the output before proceeding to the next step.

(discuss:: do we want to further subset the current video datasets into “ videos with transcript”, depending on the amount of videos we have now)?

1. Running dictionary on transcript: textual group

<https://github.com/LynetteGao/image_as_data/blob/main/video%20metadata%20collection/calculating%20transcript%20statistics.R>

* Input: csv from step 4
* Output: a “big” csv containing all info needed for textual analysis

Visual group

1. Collect mp4

* Input: csv from step 4 with all video\_id

<https://github.com/LynetteGao/image_as_data/blob/main/visual%20feature%20extraction/mp4_collection.ipynb>

* Output: mp4s + csv with a new ‘video availability’ column

(please share this new csv on the chat)

1. Extracting frames from each mp4 collected on step 6

<https://github.com/LynetteGao/image_as_data/blob/main/visual%20feature%20extraction/ExtractFrame.ipynb>

* Input: a folder with all mp4s
* Output: image folders

1. Calculating visual statistics

<https://github.com/LynetteGao/image_as_data/blob/main/visual%20feature%20extraction/ColorFeatureExtraction.ipynb>

* Input: a folder with all images
* Output: a large csv with all visual features for each video