CS3103 Project

Applications of Web Crawler on Youtube

Alex Fu
Daryl Tan
Dominic Chong
Lau Yan Han

Objective

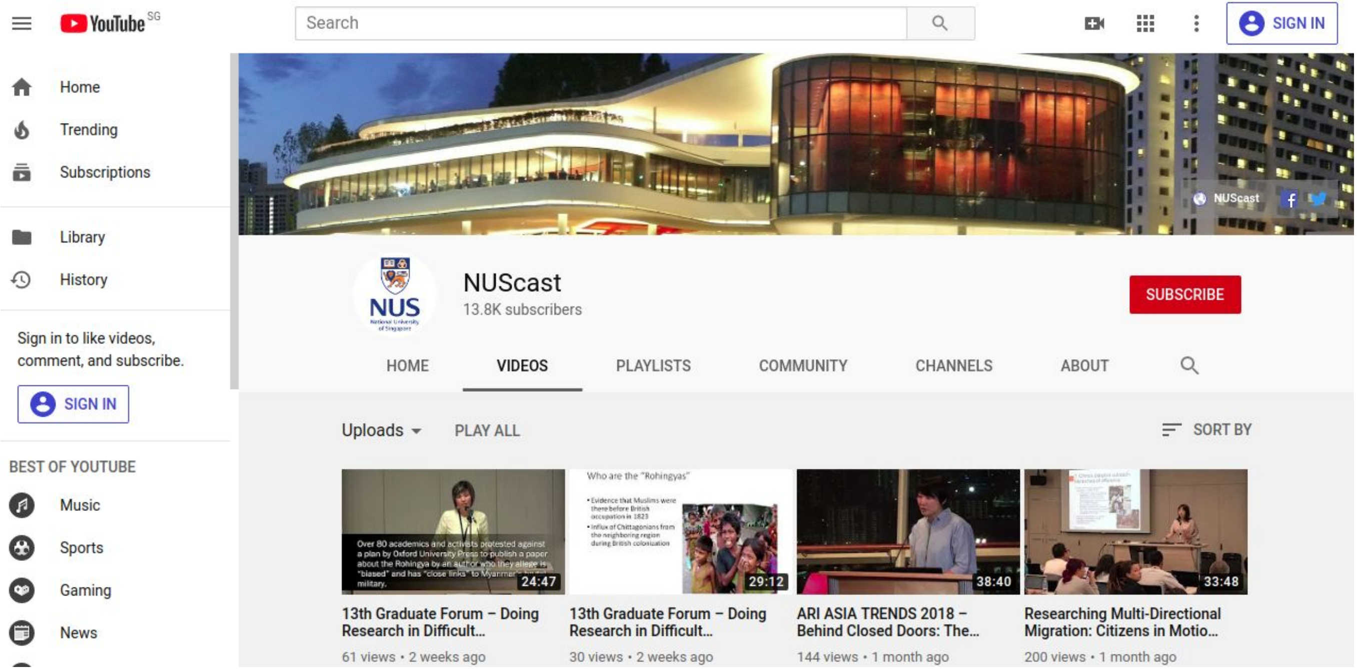
Project Proposal: **To find out the top users/channels that are related to a given online Youtube Channel**

This is done by:

- Parsing through the videos of said Youtube Channel
- Within each video, parse through the recommended videos, and find out the users who uploaded these videos
- Store these users in a text file, sorted accordingly to how frequently they appear on the “recommended videos” list

Preliminary Observations

Searching a sample Youtube Channel (https://www.youtube.com/user/NUScast/videos):


YouTube SG
Search
SIGN IN
Home
Trending
Subscriptions
Library
History
SIGN IN
BEST OF YOUTUBE
Music
Sports
Gaming
News
NUScast
NUS
National University of Singapore
SUBSCRIBE
HOME
VIDEOS
PLAYLISTS
COMMUNITY
CHANNELS
ABOUT
Uploads
PLAY ALL
SORT BY
Over 80 academics and activists protested against a plan by Oxford University Press to publish a paper about the Rohingya by an author who they allege is "biased" and has "close links" to Myanmar's brutal military.
24:47
Who are the "Rohingyas"
29:12
38:40
33:48

Sign in to like videos, comment, and subscribe.

NUScast

13.8K subscribers

Who are the "Rohingyas"

- Evidence that Muslims were there before British occupation in 1823
- Influx of Chittagonians from the neighboring region during British colonization

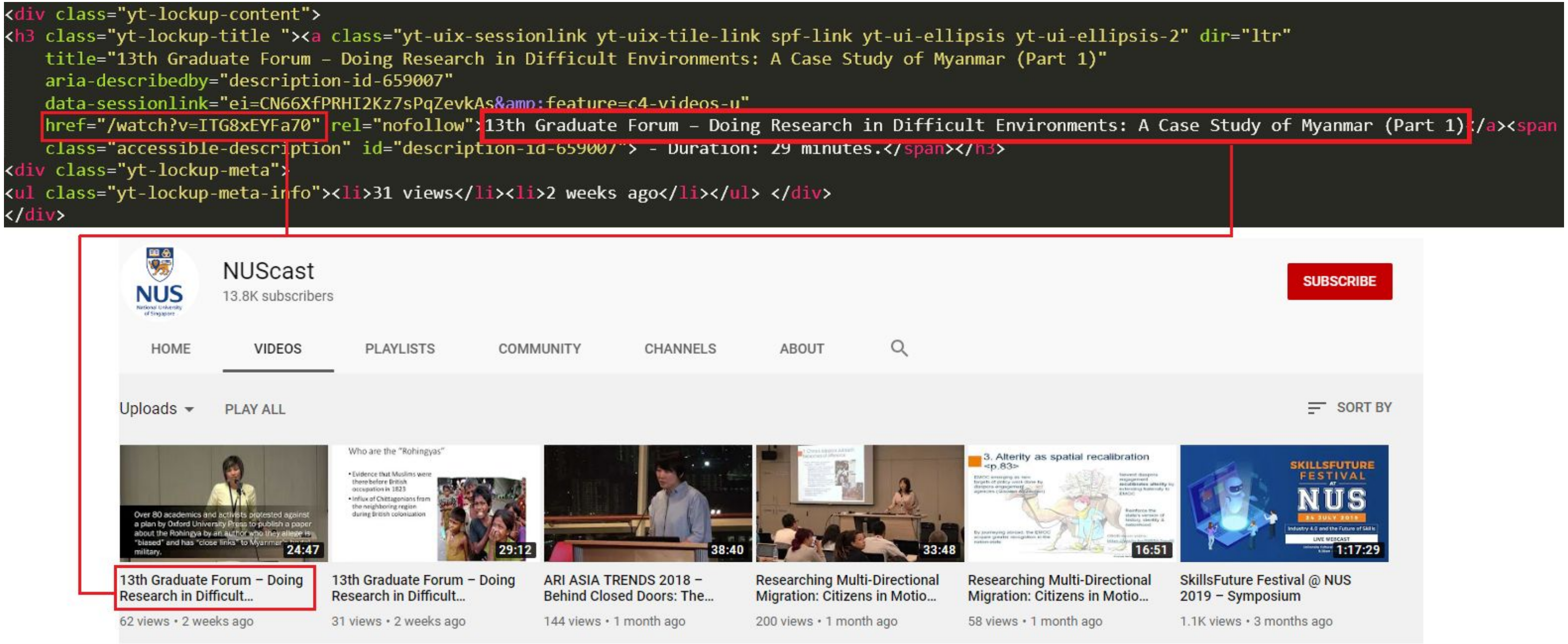
13th Graduate Forum – Doing Research in Difficult...
61 views • 2 weeks ago

13th Graduate Forum – Doing Research in Difficult...
30 views • 2 weeks ago

ARI ASIA TRENDS 2018 – Behind Closed Doors: The...
144 views • 1 month ago

Researching Multi-Directional Migration: Citizens in Motio...
200 views • 1 month ago

Preliminary Observations

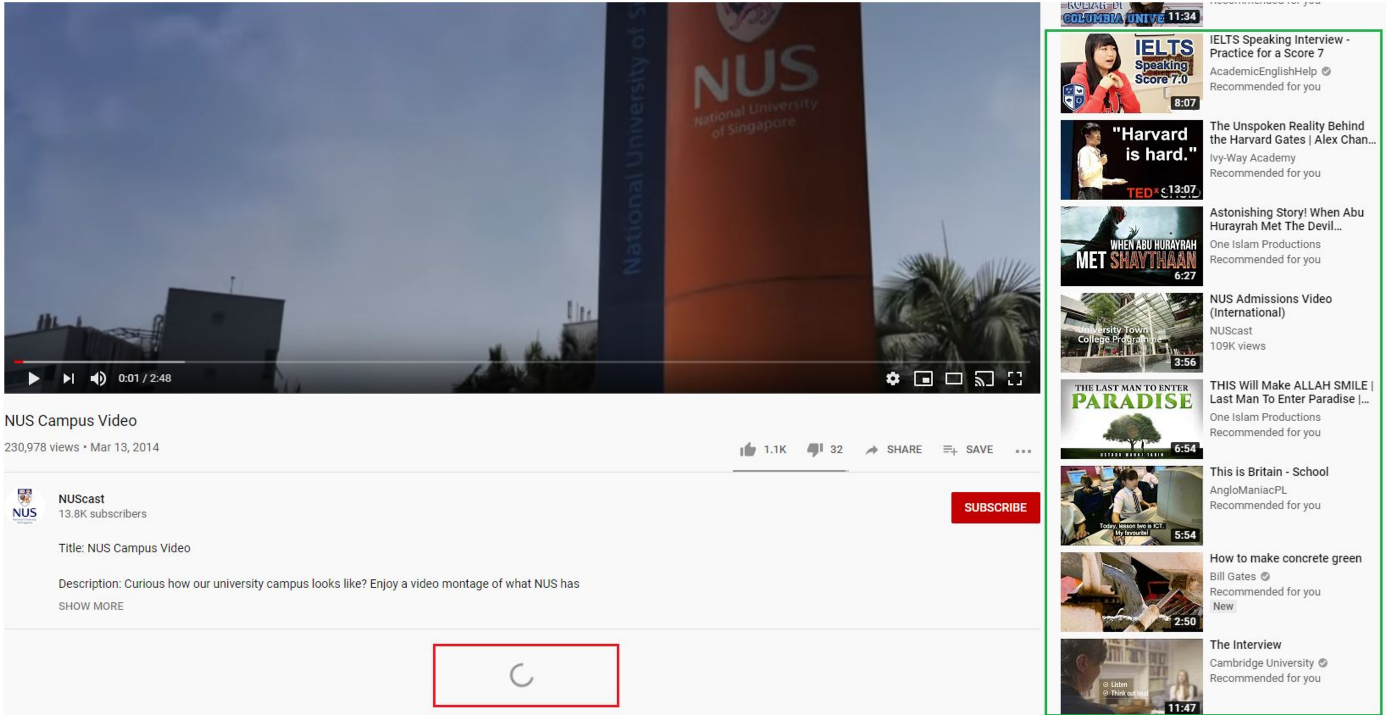


```
<div class="yt-lockup-content">
<h3 class="yt-lockup-title "><a class="yt-uix-sessionlink yt-uix-tile-link spf-link yt-ui-ellipsis yt-ui-ellipsis-2" dir="ltr"
    title="13th Graduate Forum – Doing Research in Difficult Environments: A Case Study of Myanmar (Part 1)"
    aria-describedby="description-id-659007"
    data-sessionlink="ei=CN66XfPRHI2Kz7sPqZevkAs&amp;feature=c4-videos-u"
    href="/watch?v=ITG8xEYFa70" rel="nofollow">13th Graduate Forum – Doing Research in Difficult Environments: A Case Study of Myanmar (Part 1)</a><span
    class="accessible-description" id="description-id-659007"> - Duration: 29 minutes.</span></h3>
<div class="yt-lockup-meta">
<ul class="yt-lockup-meta-info"><li>31 views</li><li>2 weeks ago</li></ul> </div>
</div>
```

NUScast
13.8K subscribers
SUBSCRIBE
HOME
VIDEOS
PLAYLISTS
COMMUNITY
CHANNELS
ABOUT
Uploads
PLAY ALL
SORT BY
24:47
29:12
38:40
33:48
16:51
1:17:29
13th Graduate Forum – Doing Research in Difficult...
62 views • 2 weeks ago
13th Graduate Forum – Doing Research in Difficult...
31 views • 2 weeks ago
ARI ASIA TRENDS 2018 – Behind Closed Doors: The...
144 views • 1 month ago
Researching Multi-Directional Migration: Citizens in Motio...
200 views • 1 month ago
Researching Multi-Directional Migration: Citizens in Motio...
58 views • 1 month ago
SkillsFuture Festival @ NUS 2019 – Symposium
1.1K views • 3 months ago

Upon acquiring the page source, we see that extracting the videos from a Youtube channel can be easily done using the "**/watch?**" keyword

Preliminary Observations



NUS Campus Video

230,978 views • Mar 13, 2014

NUScast
13.8K subscribers

Title: NUS Campus Video

Description: Curious how our university campus looks like? Enjoy a video montage of what NUS has

IELTS Speaking Interview - Practice for a Score 7
AcademicEnglishHelp
Recommended for you

The Unspoken Reality Behind the Harvard Gates | Alex Chan...
Ivy-Way Academy
Recommended for you

Astonishing Story! When Abu Hurayrah Met The Devil...
One Islam Productions
Recommended for you

NUS Admissions Video (International)
NUScast
109K views

THIS Will Make ALLAH SMILE | Last Man To Enter Paradise |...
One Islam Productions
Recommended for you

This is Britain - School
AngloManiacPL
Recommended for you

How to make concrete green
Bill Gates
Recommended for you
New

The Interview
Cambridge University
Recommended for you

Within each youtube video, we can parse through the list of **recommended videos** and extract the profile links of users who upload them (using the **'/user/"** keyword)

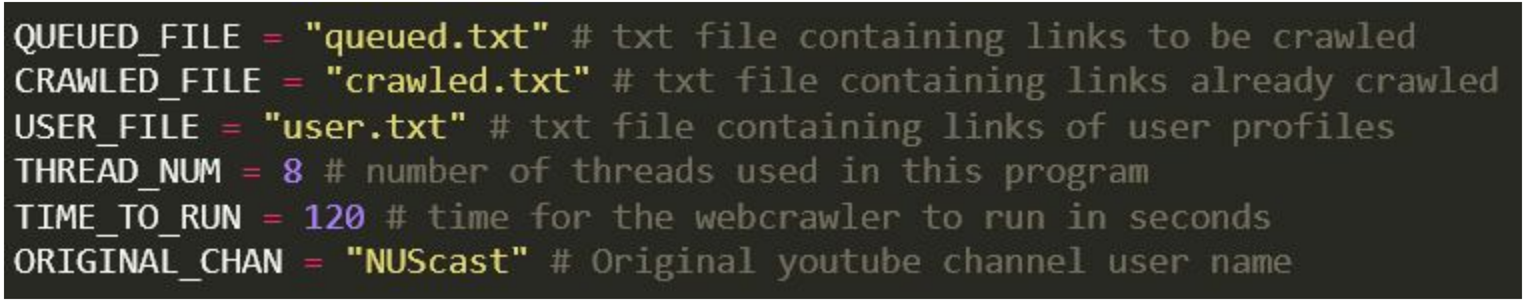
Planned Implementation

Planned Implementation:

- Employ the crawler to obtain the links of each post/video within a selected online channel
- Within each post, the crawler will proceed to extract the profile links of users who upload the list of recommended videos
- These user profile links, and the frequency of their appearance, will be stored in an appropriate data structure (e.g. hash table, binary tree) for frequency analysis

Demo

1. In webcrawler.py, we set the appropriate parameters:



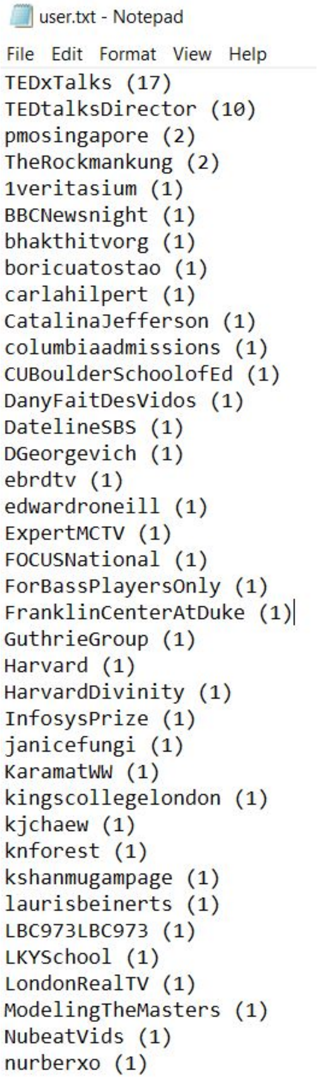
```
QUEUED_FILE = "queued.txt" # txt file containing links to be crawled
CRAWLED_FILE = "crawled.txt" # txt file containing links already crawled
USER_FILE = "user.txt" # txt file containing links of user profiles
THREAD_NUM = 8 # number of threads used in this program
TIME_TO_RUN = 120 # time for the webcrawler to run in seconds
ORIGINAL_CHAN = "NUScast" # Original youtube channel user name
```

2. Run webcrawler.py with the command: `python webcrawler.py`

3. Let the script run for TIME_TO_RUN seconds.

Demo

4. Once done, the script will output the names of all the user account names and the number of times its recommended videos show up according to Youtube’s algorithm, to the **`user.txt`** file.



```
TEDxTalks (17)
TEDtalksDirector (10)
pmosingapore (2)
TheRockmankung (2)
1veritasium (1)
BBCNewsnight (1)
bhakthitvorg (1)
boricuatostao (1)
carlahilpert (1)
CatalinaJefferson (1)
columbiaadmissions (1)
CUBoulderSchoolofEd (1)
DanyFaitDesVidos (1)
DatelineSBS (1)
DGeorgevich (1)
ebrdtv (1)
edwardroneill (1)
ExpertMCTV (1)
FOCUSNational (1)
ForBassPlayersOnly (1)
FranklinCenterAtDuke (1)
GuthrieGroup (1)
Harvard (1)
HarvardDivinity (1)
InfosysPrize (1)
janicefungi (1)
KaramatWW (1)
kingscollegelondon (1)
kjchaew (1)
knforest (1)
kshanmugampage (1)
laurisbeinerts (1)
LBC973LBC973 (1)
LKYSchool (1)
LondonRealTV (1)
ModelingTheMasters (1)
NubeatVids (1)
nurberxo (1)
```

Potential Applications

What can a Youtube channel do, given access to the list of related users (ordered in terms of frequency)?

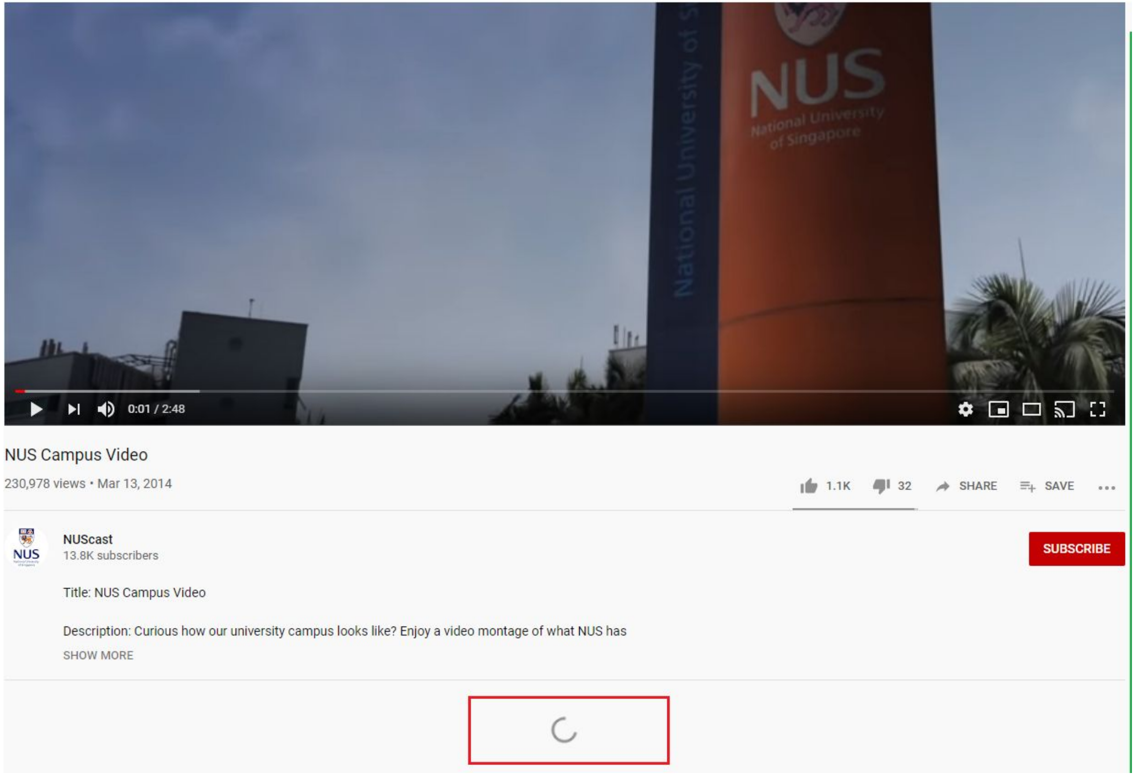
- Find out its top competitors
- Find out potential collaborators for future videos/expansion

Challenges Faced (1)

Initial Plan: We wanted to find out the **top fans of a Youtube Channel** by analyzing the comment frequency of the channel’s videos, and extract user profile links from these comments

At first glance, the implementation looks almost the same as our original plan...

Challenges Faced (1)



NUS Campus Video

230,978 views • Mar 13, 2014

NUScast

13.8K subscribers

Title: NUS Campus Video

Description: Curious how our university campus looks like? Enjoy a video montage of what NUS has

SHOW MORE

However, the links to the comments **don’t show up directly** in the parsed webpages!

Challenges Faced (1)

Perhaps it was possible to extract the comments through some Youtube provided API?

E.g. https://developers.google.com/youtube/v3/docs/comments/list

However, using the Youtube API requires authorization tokens, which are limited to Youtube developers!

Hence, the decision was made to switch the focus of the project to extracting of recommended videos — easier to obtain from a crawled webpage.

Challenges Faced (2)

Another challenge faced: How do we know whether our crawler is truly searching ALL the videos in a given channel?

- This is more relevant for large channels, where not all the videos show up unless the user scrolls down; only the top few videos are shown
- If the crawler is only searching the top few videos in the given channel, the resulting list of relevant users may not be complete.

Areas of Improvement

Scalability of the project

- Existing implementation: Specify a single Youtube Channel (“NUSCast” in our earlier example), run for a set time, retrieve list of related Youtube Channels and their frequency of appearance
- What if the channel is very large? How long do we need to run it to access ALL its videos?
- What if we want to run this crawler across several channels?