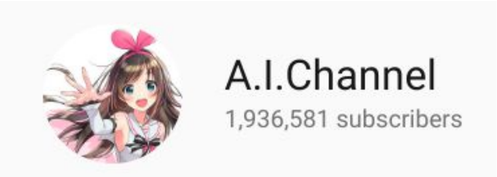
YouTube Subtitles Analysis

Chris Dong

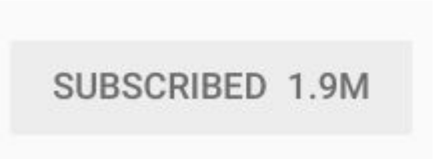
Data

- Subtitles for 342 + 102 videos
- Channels arbitrarily chosen, can be anything

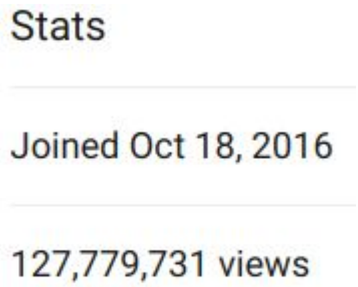


A.I.Channel

1,936,581 subscribers



SUBSCRIBED 1.9M

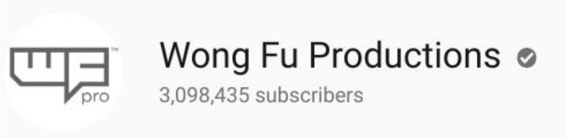


Stats

Joined Oct 18, 2016

127,779,731 views

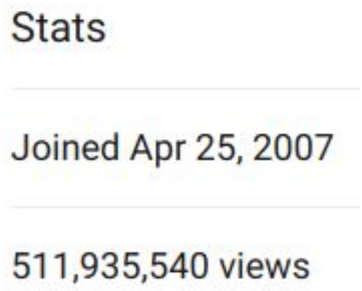
- “vlog” style - talks about random topics
- “Virtual YouTuber” pretending to be an AI



Wong Fu Productions

3,098,435 subscribers

SUBSCRIBED 3M



Stats

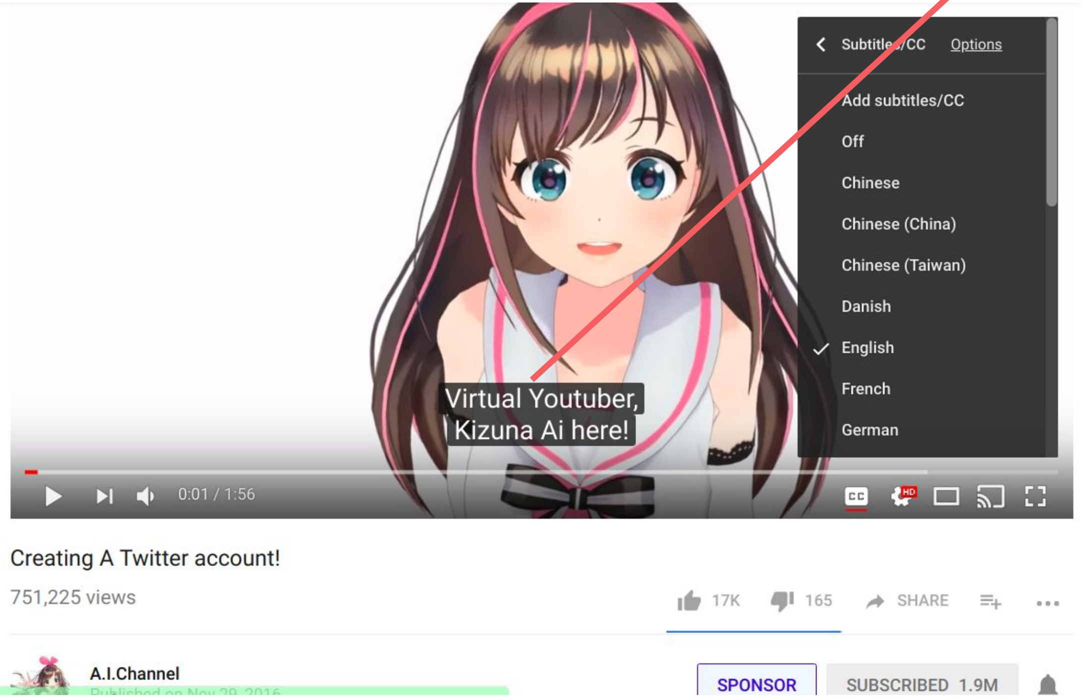
Joined Apr 25, 2007

511,935,540 views

- Makes mostly short films

Data

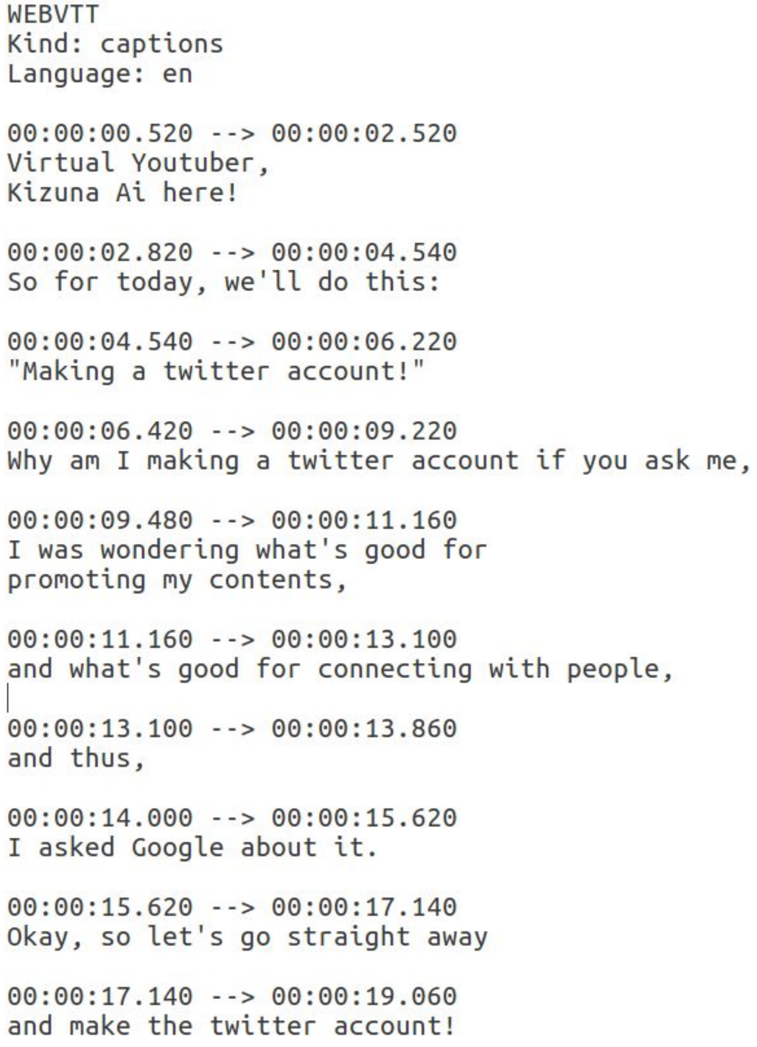
- Scraped from YouTube captions



Virtual Youtuber,
Kizuna Ai here!

Creating A Twitter account!

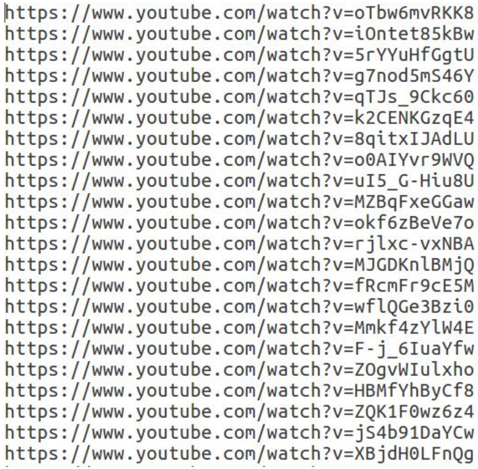
751,225 views



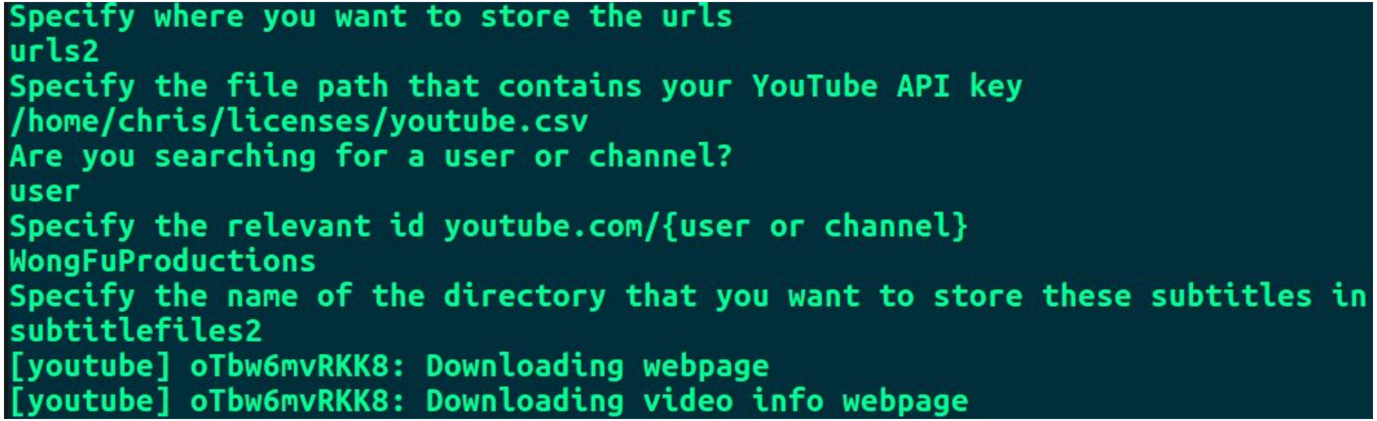
```
WEBVTT
Kind: captions
Language: en

00:00:00.520 --> 00:00:02.520
Virtual Youtuber,
Kizuna Ai here!

00:00:02.820 --> 00:00:04.540
So for today, we'll do this:

00:00:04.540 --> 00:00:06.220
"Making a twitter account!"

00:00:06.420 --> 00:00:09.220
Why am I making a twitter account if you ask me,

00:00:09.480 --> 00:00:11.160
I was wondering what's good for
promoting my contents,

00:00:11.160 --> 00:00:13.100
and what's good for connecting with people,

00:00:13.100 --> 00:00:13.860
and thus,

00:00:14.000 --> 00:00:15.620
I asked Google about it.

00:00:15.620 --> 00:00:17.140
Okay, so let's go straight away

00:00:17.140 --> 00:00:19.060
and make the twitter account!
```

Data

- Given a user’s channel name, retrieve a list of URLS
- For each url, download subtitles (videos that don’t have subtitles will be skipped
- Create bash script to automate downloading

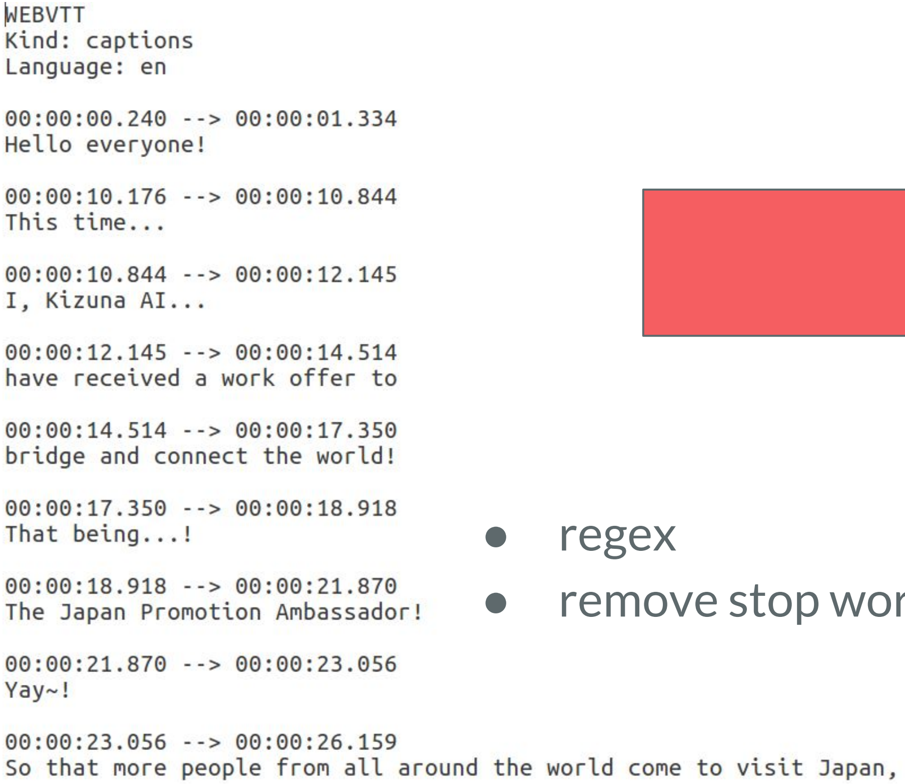


```
https://www.youtube.com/watch?v=oTbw6mvRKK8
https://www.youtube.com/watch?v=iOntet85kBw
https://www.youtube.com/watch?v=5rYYuHfGgtU
https://www.youtube.com/watch?v=g7nod5mS46Y
https://www.youtube.com/watch?v=qTJs_9Ckc60
https://www.youtube.com/watch?v=k2CENKGzqE4
https://www.youtube.com/watch?v=8qitxIJAdLU
https://www.youtube.com/watch?v=o0AIYvr9WVQ
https://www.youtube.com/watch?v=uI5_G-Hiu8U
https://www.youtube.com/watch?v=MZBqFxeGGaw
https://www.youtube.com/watch?v=okf6zBeVe7o
https://www.youtube.com/watch?v=rjlxc-vxNBA
https://www.youtube.com/watch?v=MJGDKnlBMjQ
https://www.youtube.com/watch?v=fRcmFr9cE5M
https://www.youtube.com/watch?v=wflQGe3Bzi0
https://www.youtube.com/watch?v=Mmkf4zYlW4E
https://www.youtube.com/watch?v=F-j_6IuaYfw
https://www.youtube.com/watch?v=ZOgvWIulxho
https://www.youtube.com/watch?v=HBMfYhByCf8
https://www.youtube.com/watch?v=ZQK1F0wz6z4
https://www.youtube.com/watch?v=jS4b91DaYCw
https://www.youtube.com/watch?v=XBjdH0LFnQg
```



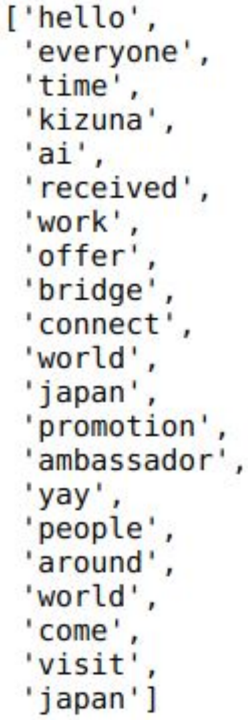
```
Specify where you want to store the urls
urls2
Specify the file path that contains your YouTube API key
/home/chris/licenses/youtube.csv
Are you searching for a user or channel?
user
Specify the relevant id youtube.com/{user or channel}
WongFuProductions
Specify the name of the directory that you want to store these subtitles in
subtitlefiles2
[youtube] oTbw6mvRKK8: Downloading webpage
[youtube] oTbw6mvRKK8: Downloading video info webpage
```

Preprocess



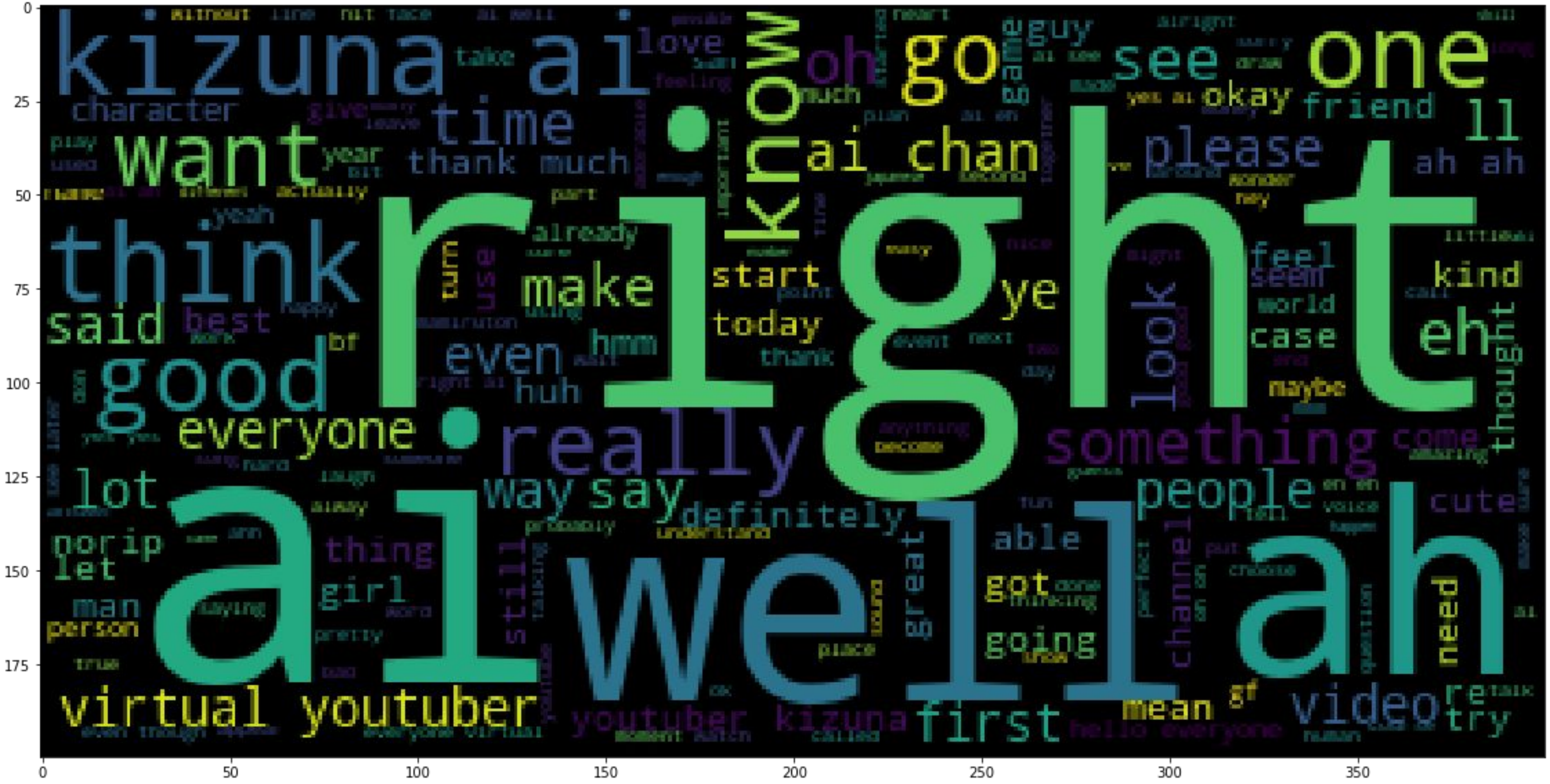
```
WEBVTT
Kind: captions
Language: en

00:00:00.240 --> 00:00:01.334
Hello everyone!

00:00:10.176 --> 00:00:10.844
This time...

00:00:10.844 --> 00:00:12.145
I, Kizuna AI...

00:00:12.145 --> 00:00:14.514
have received a work offer to

00:00:14.514 --> 00:00:17.350
bridge and connect the world!

00:00:17.350 --> 00:00:18.918
That being...!

00:00:18.918 --> 00:00:21.870
The Japan Promotion Ambassador!

00:00:21.870 --> 00:00:23.056
Yay~!

00:00:23.056 --> 00:00:26.159
So that more people from all around the world come to visit Japan,
```

- regex
- remove stop words, punctuation, etc



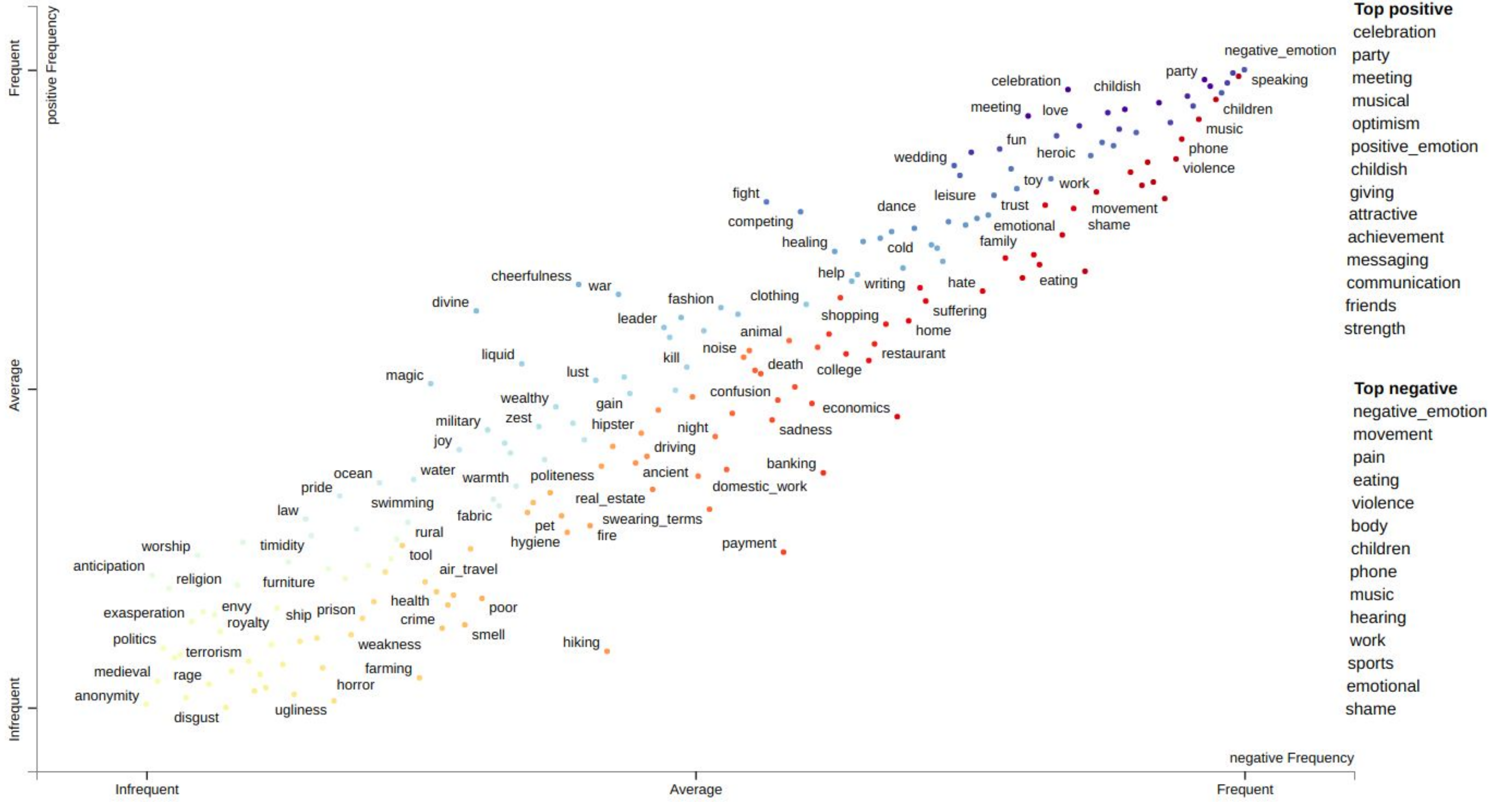
```
['hello',
 'everyone',
 'time',
 'kizuna',
 'ai',
 'received',
 'work',
 'offer',
 'bridge',
 'connect',
 'world',
 'japan',
 'promotion',
 'ambassador',
 'yay',
 'people',
 'around',
 'world',
 'come',
 'visit',
 'japan']
```

Summary


right
ai
well
ah
kizuna ai
think
want
good
know
go
one
time
ai chan
oh
see
please
really
something
everyone
virtual youtuber
youtuber kizuna
first
video
people
look
make
said
lot
eh
thought
feel
kind
okay
friend
ll
ah ah
guy
game
love
character
thank much
year
yeah
already
start
today
ye
even
huh
hmm
best
bf
way say
definitely
able
cute
come
case
world
seem
great
channel
still
norip
thing
let
girl
man
person
got
going
mean
need
try
hello everyone
maybe
take
give
0
25
50
75
100
125
150
175
0
50
100
150
200
250
300
350

Summary

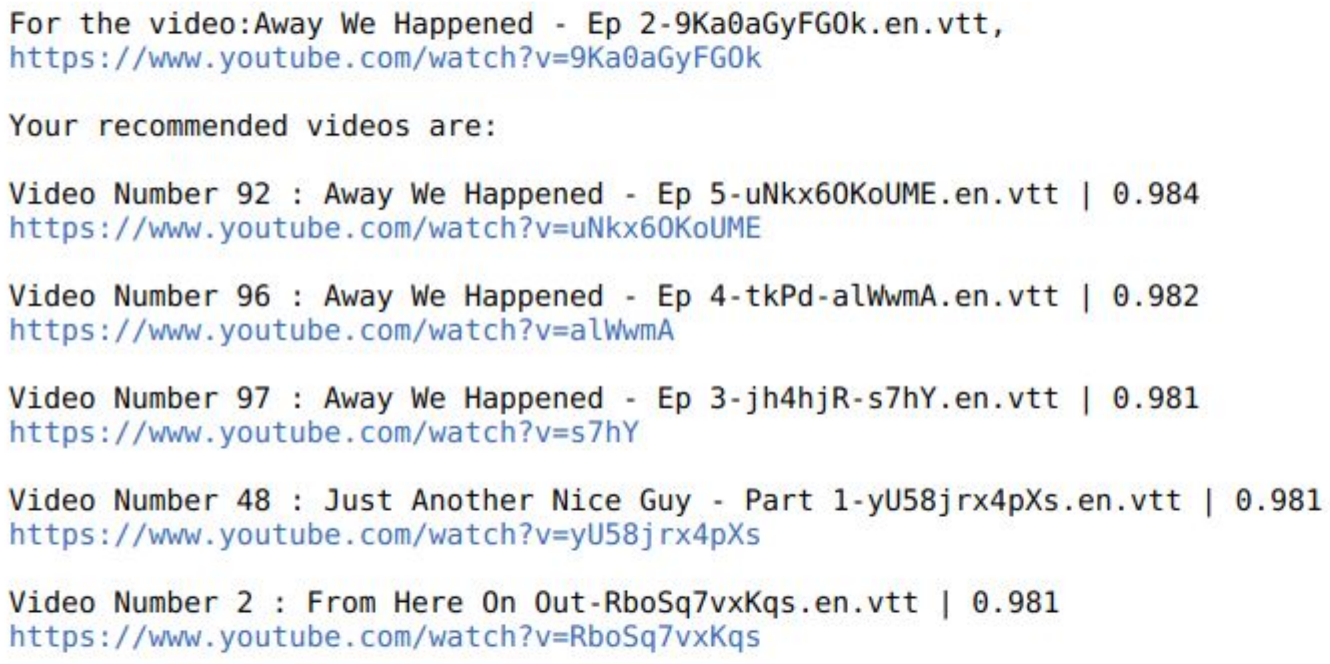
- Interactive (HTML)
- scattertext



Model Fitting

Recommend video based on subtitles

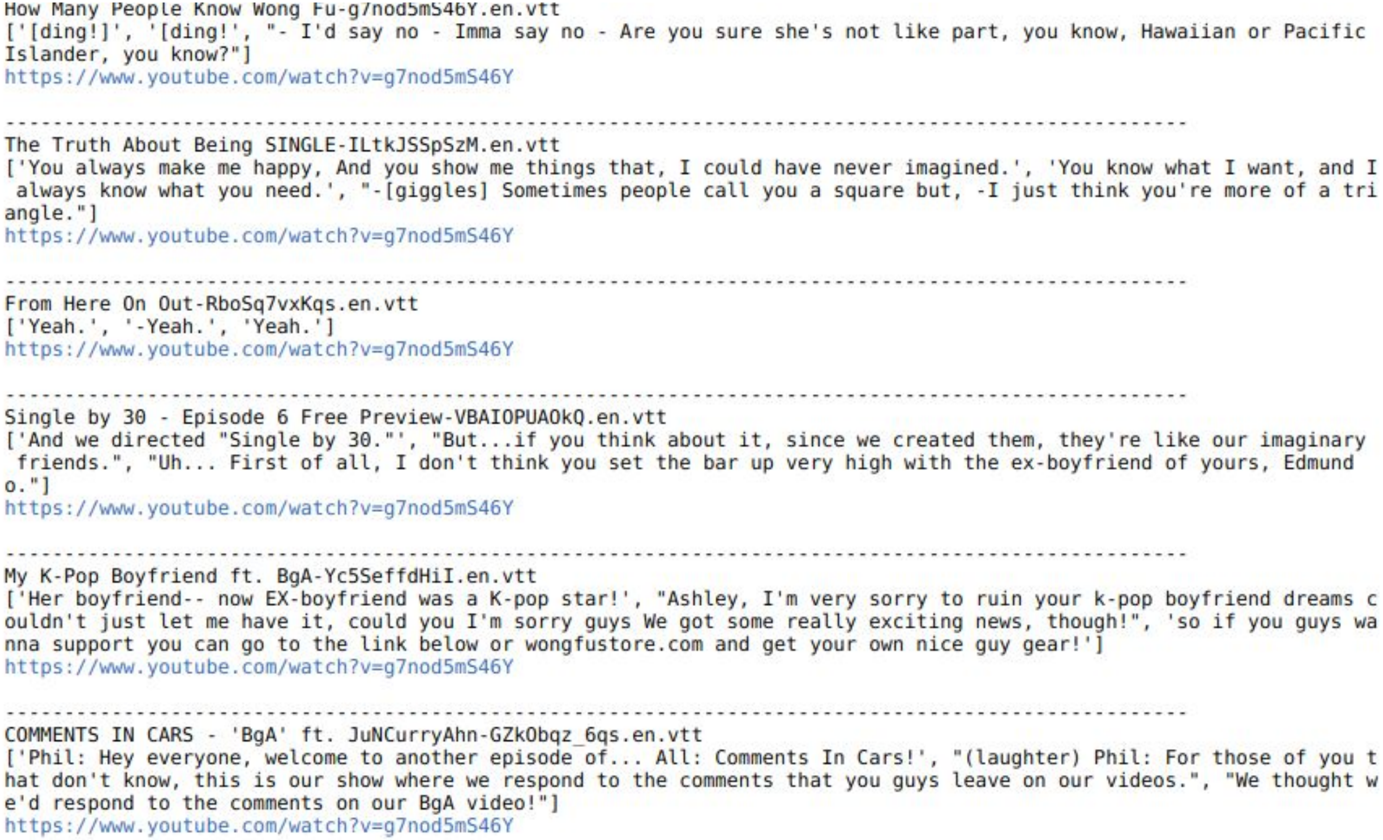
- GloVe vectors (cosine similarity)



```
For the video:Away We Happened - Ep 2-9Ka0aGyFGOk.en.vtt,
https://www.youtube.com/watch?v=9Ka0aGyFGOk

Your recommended videos are:

Video Number 92 : Away We Happened - Ep 5-uNkx6OKoUME.en.vtt | 0.984
https://www.youtube.com/watch?v=uNkx6OKoUME

Video Number 96 : Away We Happened - Ep 4-tkPd-alWwmA.en.vtt | 0.982
https://www.youtube.com/watch?v=alWwmA

Video Number 97 : Away We Happened - Ep 3-jh4hjR-s7hY.en.vtt | 0.981
https://www.youtube.com/watch?v=s7hY

Video Number 48 : Just Another Nice Guy - Part 1-yU58jrx4pXs.en.vtt | 0.981
https://www.youtube.com/watch?v=yU58jrx4pXs

Video Number 2 : From Here On Out-RboSq7vxKqs.en.vtt | 0.981
https://www.youtube.com/watch?v=RboSq7vxKqs
```

Model Fitting

Summarizer



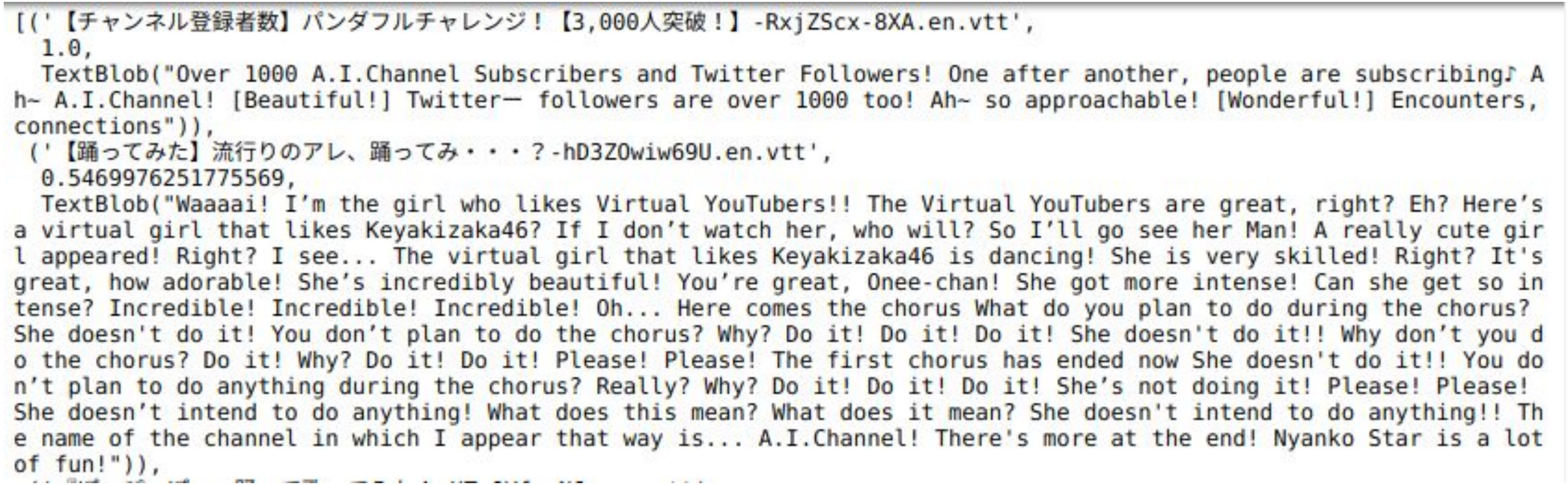
```
How Many People Know Wong Fu-g7nod5mS46Y.en.vtt
['[ding!]', '[ding!', "- I'd say no - Imma say no - Are you sure she's not like part, you know, Hawaiian or Pacific
Islander, you know?"]
https://www.youtube.com/watch?v=g7nod5mS46Y

----------------------------------------------------------------------------------------------------
The Truth About Being SINGLE-ILtkJSSpSzM.en.vtt
['You always make me happy, And you show me things that, I could have never imagined.', 'You know what I want, and I
 always know what you need.', "-[giggles] Sometimes people call you a square but, -I just think you're more of a tri
angle."]
https://www.youtube.com/watch?v=g7nod5mS46Y

----------------------------------------------------------------------------------------------------
From Here On Out-RboSq7vxKqs.en.vtt
['Yeah.', '-Yeah.', 'Yeah.']
https://www.youtube.com/watch?v=g7nod5mS46Y

----------------------------------------------------------------------------------------------------
Single by 30 - Episode 6 Free Preview-VBAIOPUAOkQ.en.vtt
['And we directed "Single by 30."', "But...if you think about it, since we created them, they're like our imaginary
 friends.", "Uh... First of all, I don't think you set the bar up very high with the ex-boyfriend of yours, Edmund
o."]
https://www.youtube.com/watch?v=g7nod5mS46Y

----------------------------------------------------------------------------------------------------
My K-Pop Boyfriend ft. BgA-Yc5SeffdHiI.en.vtt
['Her boyfriend-- now EX-boyfriend was a K-pop star!', "Ashley, I'm very sorry to ruin your k-pop boyfriend dreams c
ouldn't just let me have it, could you I'm sorry guys We got some really exciting news, though!", 'so if you guys wa
nna support you can go to the link below or wongfustore.com and get your own nice guy gear!']
https://www.youtube.com/watch?v=g7nod5mS46Y

----------------------------------------------------------------------------------------------------
COMMENTS IN CARS - 'BgA' ft. JuNCurryAhn-GZkObqz_6qs.en.vtt
['Phil: Hey everyone, welcome to another episode of... All: Comments In Cars!', "(laughter) Phil: For those of you t
hat don't know, this is our show where we respond to the comments that you guys leave on our videos.", "We thought w
e'd respond to the comments on our BgA video!"]
https://www.youtube.com/watch?v=g7nod5mS46Y
```

Model Fitting

Sentiment Analysis

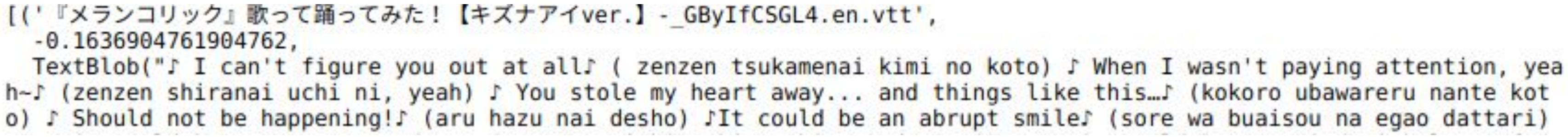
+

+



```
[('【チャンネル登録者数】パンダフルチャレンジ！【3,000人突破！】-RxjZScx-8XA.en.vtt',
  1.0,
  TextBlob("Over 1000 A.I.Channel Subscribers and Twitter Followers! One after another, people are subscribing♪ A
h~ A.I.Channel! [Beautiful!] Twitter— followers are over 1000 too! Ah~ so approachable! [Wonderful!] Encounters,
connections")),
 ('【踊ってみた】流行りのアレ、踊ってみ・・・？-hD3ZOwiw69U.en.vtt',
  0.5469976251775569,
  TextBlob("Waaaai! I’m the girl who likes Virtual YouTubers!! The Virtual YouTubers are great, right? Eh? Here’s
a virtual girl that likes Keyakizaka46? If I don’t watch her, who will? So I’ll go see her Man! A really cute gir
l appeared! Right? I see... The virtual girl that likes Keyakizaka46 is dancing! She is very skilled! Right? It's
great, how adorable! She’s incredibly beautiful! You’re great, Onee-chan! She got more intense! Can she get so in
tense? Incredible! Incredible! Incredible! Oh... Here comes the chorus What do you plan to do during the chorus?
She doesn't do it! You don’t plan to do the chorus? Why? Do it! Do it! Do it! She doesn't do it!! Why don’t you d
o the chorus? Do it! Why? Do it! Do it! Please! Please! The first chorus has ended now She doesn't do it!! You do
n’t plan to do anything during the chorus? Really? Why? Do it! Do it! Do it! She’s not doing it! Please! Please!
She doesn’t intend to do anything! What does this mean? What does it mean? She doesn't intend to do anything!! Th
e name of the channel in which I appear that way is... A.I.Channel! There's more at the end! Nyanko Star is a lot
of fun!")),
```

−

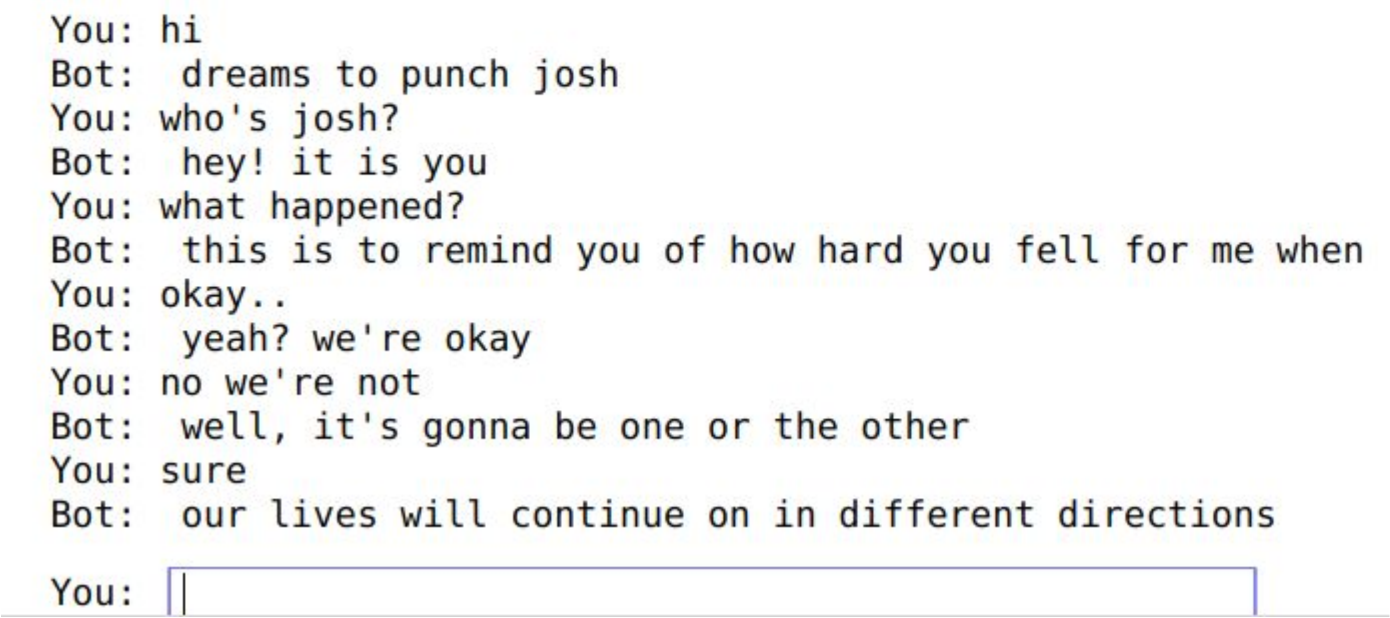


```
[('『メランコリック』歌って踊ってみた！【キズナアイver.】-_GByIfCSGL4.en.vtt',
  -0.1636904761904762,
  TextBlob("♪ I can't figure you out at all♪ ( zenzen tsukamenai kimi no koto) ♪ When I wasn't paying attention, yea
h~♪ (zenzen shiranai uchi ni, yeah) ♪ You stole my heart away... and things like this…♪ (kokoro ubawareru nante kot
o) ♪ Should not be happening!♪ (aru hazu nai desho) ♪It could be an abrupt smile♪ (sore wa buaisou na egao dattari)
```

Model Fitting

Chatbot

- Bot doesn’t make sense
- Training data not that great (“garbage in, garbage out”)



```
You: hi
Bot:  dreams to punch josh
You: who's josh?
Bot:  hey! it is you
You: what happened?
Bot:  this is to remind you of how hard you fell for me when
You: okay..
Bot:  yeah? we're okay
You: no we're not
Bot:  well, it's gonna be one or the other
You: sure
Bot:  our lives will continue on in different directions

You:
```

Documentation

https://github.com/dongchris/subtitles-chat-bot

Lessons Learned

- Working with YouTube API can be frustrating
- Subtitle text data is messy
- Having a *good* chatbot is a difficult problem