Farahin Choudhury

Text classification: YouTube sports channels

IST736 Text Mining

Objective:

Find common tags and comments on videos from sports channels (a classification task)

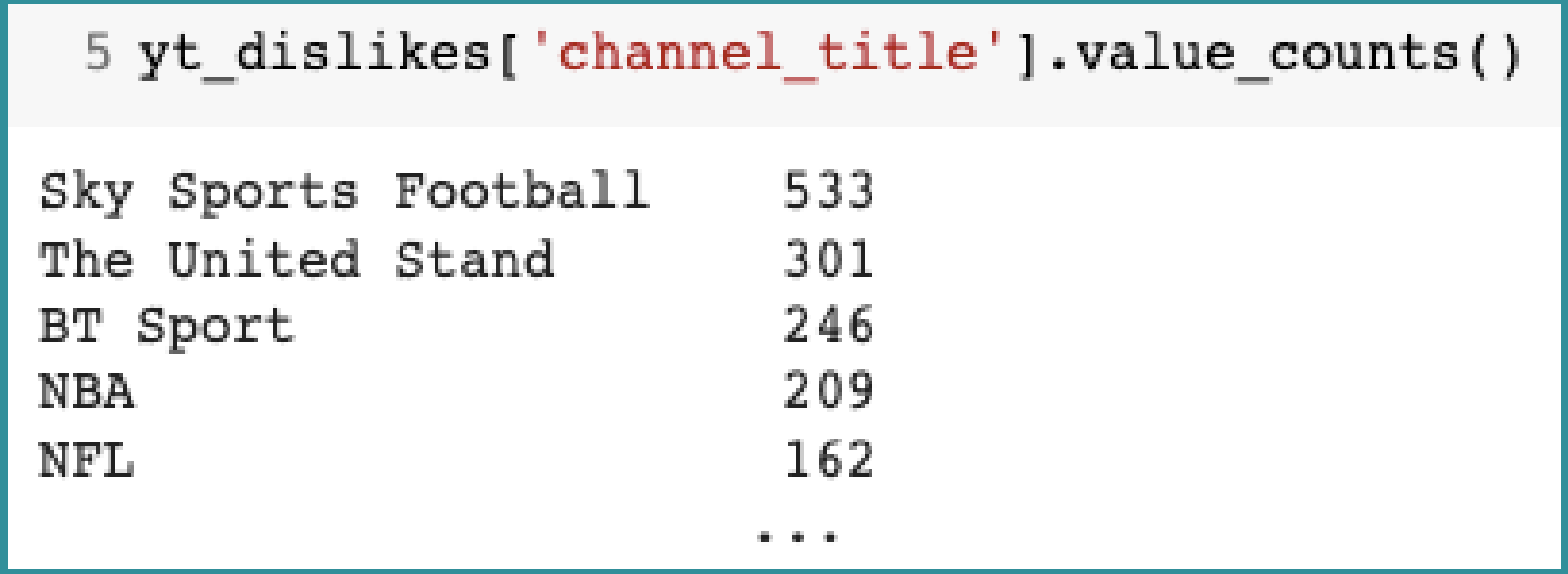
- **use supervised learning methods to uncover themes or sentiment (LDA, clustering)**
- **common terms across all videos (both in tags and comments)**
- **term frequency vectorization**

Why explore this area of analysis?

- Uncover sentiment in YouTube comments; YouTube commentary can reveal how users feel about a video
- Recurring tags on channels with high dislike count
- For a theme/topic like sports, so many types of conversations can happen related or unrelated to the sport itself

SPORTS DATA

HIGHEST NUMBER OF OBSERVATIONS ('CHANNEL_TITLE') ALL HAPPEN TO BE FOR SPORTS CHANNELS



```
yt_dislikes['channel_title'].value_counts()

Sky Sports Football    533
The United Stand       301
BT Sport               246
NBA                    209
NFL                    162
                      ...
```

data obtained from Kaggle

DATA PREP

CLEANING/MUNGING

Combine data, drop columns that do not have relevant data (video_id, channel_id),

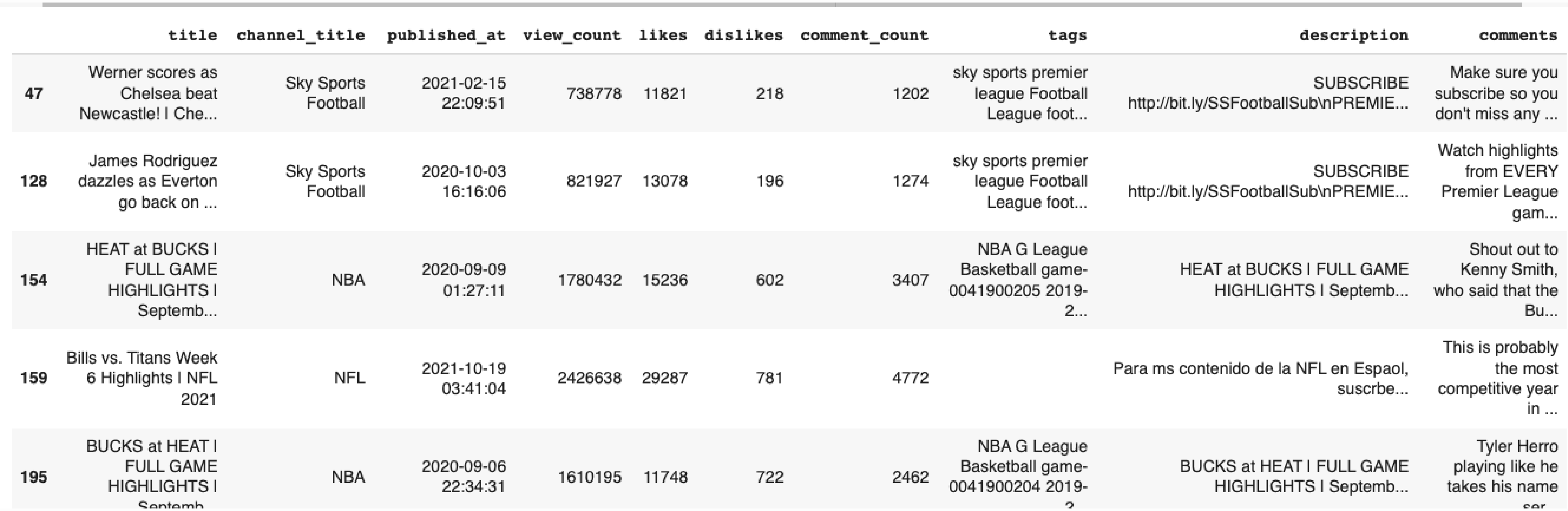
EXPLORATION

Understand the data and what is provided in the .csv file

PRESENTATION

Present the data in a concise manner, wordclouds of common terms/tags

data set example



	title	channel_title	published_at	view_count	likes	dislikes	comment_count	tags	description	comments
47	Werner scores as Chelsea beat Newcastle! | Che...	Sky Sports Football	2021-02-15 22:09:51	738778	11821	218	1202	sky sports premier league Football League foot...	SUBSCRIBE http://bit.ly/SSFootballSub\nPREMIE...	Make sure you subscribe so you don't miss any ...
128	James Rodriguez dazzles as Everton go back on ...	Sky Sports Football	2020-10-03 16:16:06	821927	13078	196	1274	sky sports premier league Football League foot...	SUBSCRIBE http://bit.ly/SSFootballSub\nPREMIE...	Watch highlights from EVERY Premier League gam...
154	HEAT at BUCKS | FULL GAME HIGHLIGHTS | Septemb...	NBA	2020-09-09 01:27:11	1780432	15236	602	3407	NBA G League Basketball game-0041900205 2019-2...	HEAT at BUCKS | FULL GAME HIGHLIGHTS | Septemb...	Shout out to Kenny Smith, who said that the Bu...
159	Bills vs. Titans Week 6 Highlights | NFL 2021	NFL	2021-10-19 03:41:04	2426638	29287	781	4772		Para ms contenido de la NFL en Espaol, suscrbe...	This is probably the most competitive year in ...
195	BUCKS at HEAT | FULL GAME HIGHLIGHTS | Septemb...	NBA	2020-09-06 22:34:31	1610195	11748	722	2462	NBA G League Basketball game-0041900204 2019-2...	BUCKS at HEAT | FULL GAME HIGHLIGHTS | Septemb...	Tyler Herro playing like he takes his name ser...

Python Modules

sci-kit learn

main library to analyze text

pandas

means of organizing data in a data frame (data came from .csv file)

nltk

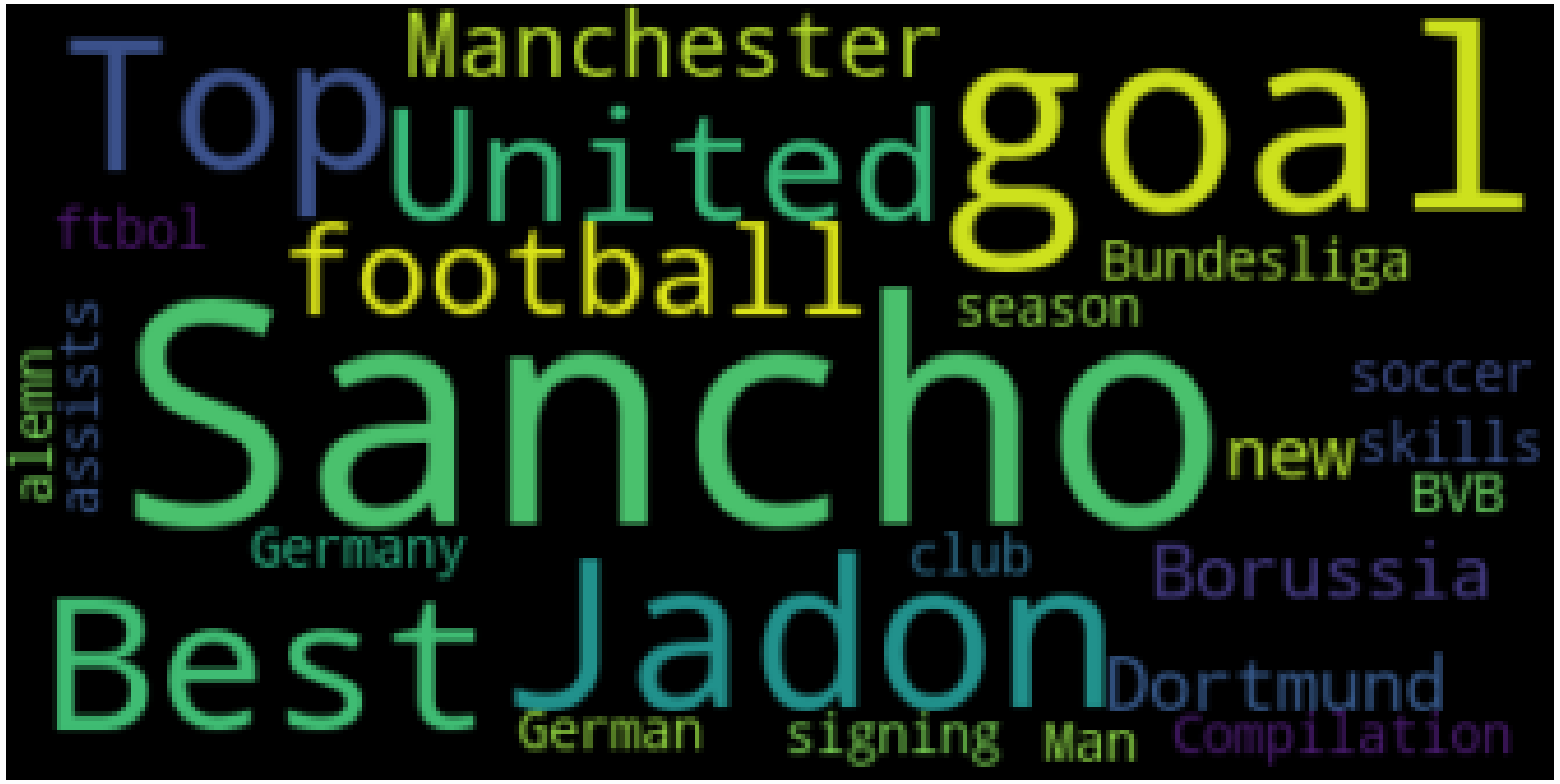
library to explore sentiment of text and to do clustering and categorization

gensim or LDA

topic modeling

wordcloud

TO VISUALIZE TERM FREQUENCY



SOME QUESTIONS TO EXPLORE

- WHICH TERMS ARE ASSOCIATED WITH VIDEOS WITH THE MOST DISLIKES?
- WITHIN THE TOPIC OF SPORTS, ARE THERE COMMON THEMES/KINDS OF COMMENTS? WHAT ARE THEY?
- HOW CAN FEATURE WEIGHTS AND RANKING HELP TO UNDERSTAND TRENDS?

Challenges

- Determining which algorithms and techniques to incorporate (what is ideal for text data not in a traditional corpus)

- The algorithms might need to be refined several times in order to obtain ideal results

- Comments in the 'comments' field are all lumped together so there is no separation to discern comments from one another

NEXT STEPS

Refine problem statement/goal

Determine the best algorithm(s)

Detect sentiment from comments

Thank you!