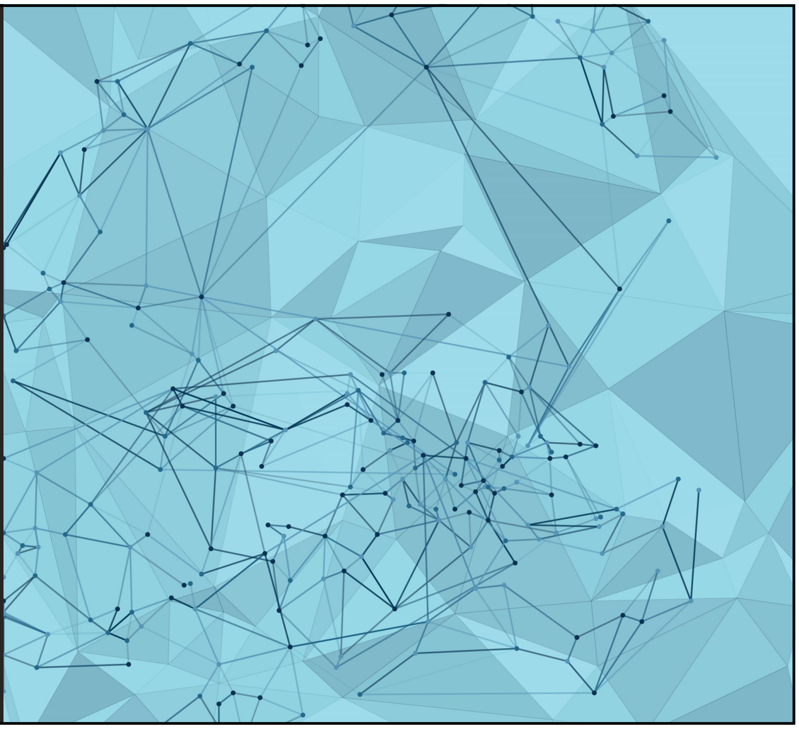
SPRING 2020 WAF DATA CHALLENGE

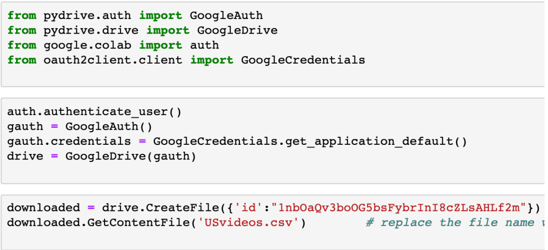
ABHISHEK PANDYA



1

CLEANING

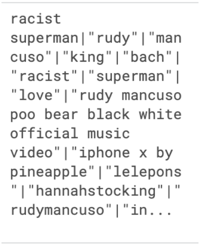
- Import Data
- Clean Tags



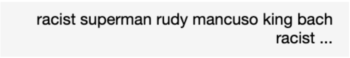
```python
from pydrive.auth import GoogleAuth
from pydrive.drive import GoogleDrive
from google.colab import auth
from oauth2client.client import GoogleCredentials
```

```python
auth.authenticate_user()
gauth = GoogleAuth()
gauth.credentials = GoogleCredentials.get_application_default()
drive = GoogleDrive(gauth)
```

```python
downloaded = drive.CreateFile({'id':"1nbOaQv3boOG5bsFybrInI8cZLsAHLf2m"})
downloaded.GetContentFile('USvideos.csv')        # replace the file name
```



```
racist
superman|"rudy"|"man
cuso"|"king"|"bach"|
"racist"|"superman"|
"love"|"rudy mancuso
poo bear black white
official music
video"|"iphone x by
pineapple"|"lelepons
"|"hannahstocking"|"
rudymancuso"|"in...
```



racist superman rudy mancuso king bach racist ...

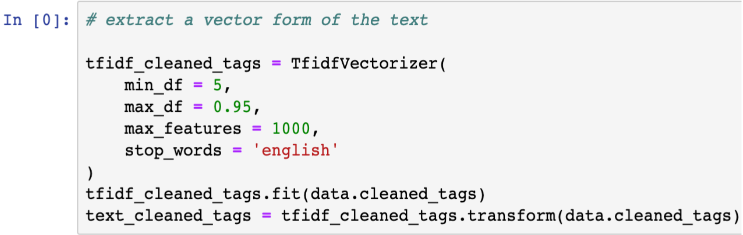
2

FEATURE EXTRACTION (TF-DIF)

What is TF-DIF?

- Compares the number of times a word appears in a doc with how often the word appears on the web
- Used to weigh factors in information retrieval and text mining
- Many search engines use this to rank relevancy of web pages and documents

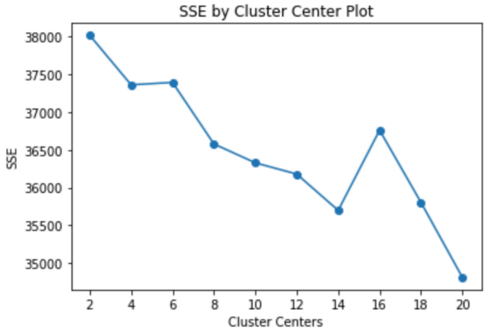
- I decided to use tags.
- I used TF-DIF to extract the most frequent/relevant tags.
- Used scikit-learn Tfidf Vectorizer

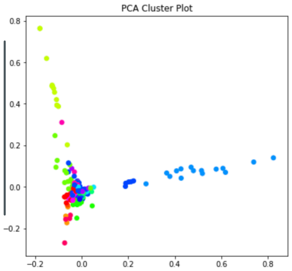


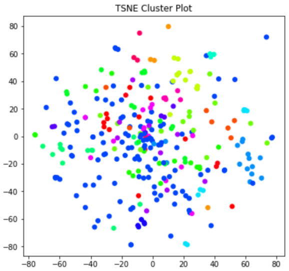
```
In [0]: # extract a vector form of the text

        tfidf_cleaned_tags = TfidfVectorizer(
            min_df = 5,
            max_df = 0.95,
            max_features = 1000,
            stop_words = 'english'
        )
        tfidf_cleaned_tags.fit(data.cleaned_tags)
        text_cleaned_tags = tfidf_cleaned_tags.transform(data.cleaned_tags)
```

3







ML MODEL - KMEANS

- 20 clusters had the least error
- PCA captures global structure of the data
- TSNE captures relations with neighbors

4

ANALYSIS – WHAT DID THE ML FIND?

- Science videos
- Celebrities/comedy
- Star wars
- Cooking/food
- Kittens, pets
- Failures: Cluster 0 (~50% of videos belong here, and have little relation)

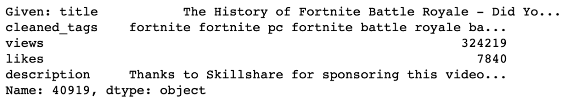
- Cluster 4: seth, fallon, rhett, link, nbc, night, sketch, snl, funny, comedy
- Cluster 10 kitchen, challenge, cook, make, chef, chocolate, cake, cooking, recipe, food
- Cluster 14: technology, random, space, physics, education, apple, tech, green, iphone, science
- Cluster 15: espn, sports, lebron, highlights, fortnite, games, dude, basketball, nba, game
- Cluster 0: christmas, love, react, live, video, new, animation, vlog, diy, funny

5

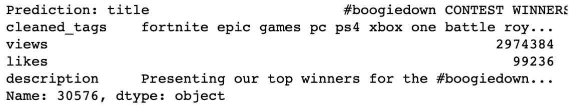
RECOMMENDATION ENGINE

Input Video → Finds cluster of video → Returns random video from cluster

Recognizes Videogames

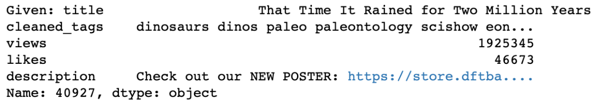


```
Given: title           The History of Fortnite Battle Royale - Did Yo...
cleaned_tags    fortnite fortnite pc fortnite battle royale ba...
views                                                      324219
likes                                                        7840
description     Thanks to Skillshare for sponsoring this video...
Name: 40919, dtype: object
```

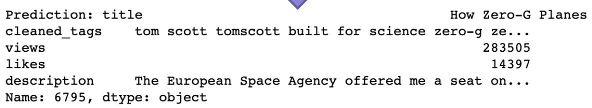


```
Prediction: title                          #boogiedown CONTEST WINNERS
cleaned_tags    fortnite epic games pc ps4 xbox one battle roy...
views                                                    2974384
likes                                                      99236
description     Presenting our top winners for the #boogiedown...
Name: 30576, dtype: object
```

Recognizes Science Videos



```
Given: title                        That Time It Rained for Two Million Years
cleaned_tags    dinosaurs dinos paleo paleontology scishow eon...
views                                                    1925345
likes                                                      46673
description     Check out our NEW POSTER: https://store.dftba....
Name: 40927, dtype: object
```



```
Prediction: title                                    How Zero-G Planes
cleaned_tags    tom scott tomscott built for science zero-g ze...
views                                                     283505
likes                                                      14397
description     The European Space Agency offered me a seat on...
Name: 6795, dtype: object
```

6