

TED UNIVERSITY

Faculty of Engineering

Department of Computer Engineering

**Weekly Report**

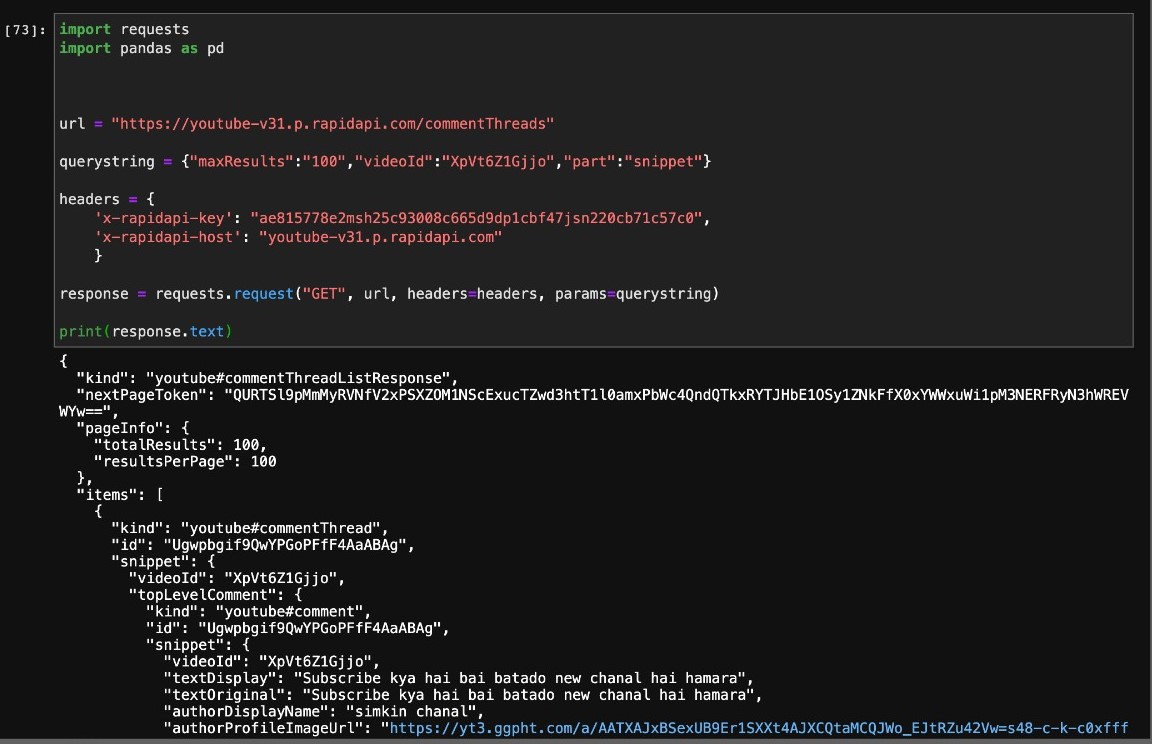
by

Yağız ÇİMEN – Efe TONTU – Mesut Nadir SEYFELİOĞLU

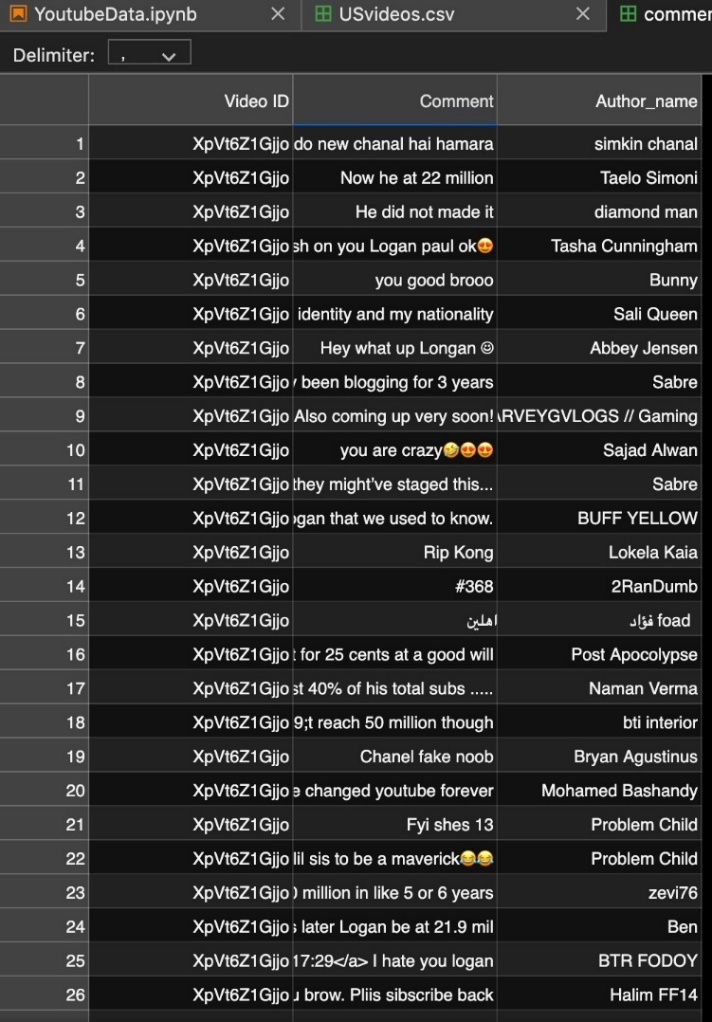
**Supervisor:** Assoc. Prof. Dr. Tansel DÖKEROĞLU

1. **Gather data from YouTube API**

This week we got data from YouTube API using **rapidapi** on python. We used rapidapi because it is open source and free. Data we got contains 100 comments, their authors on corresponding video. We send request by using **video\_id** (XpVt6Z1Gjjo) from **USvideos.csv** dataset for one video comments. It took like a second to show the result. Then we convert the json file, which is result of call, to csv, Codes are [here](https://github.com/mnadirs/e-big).

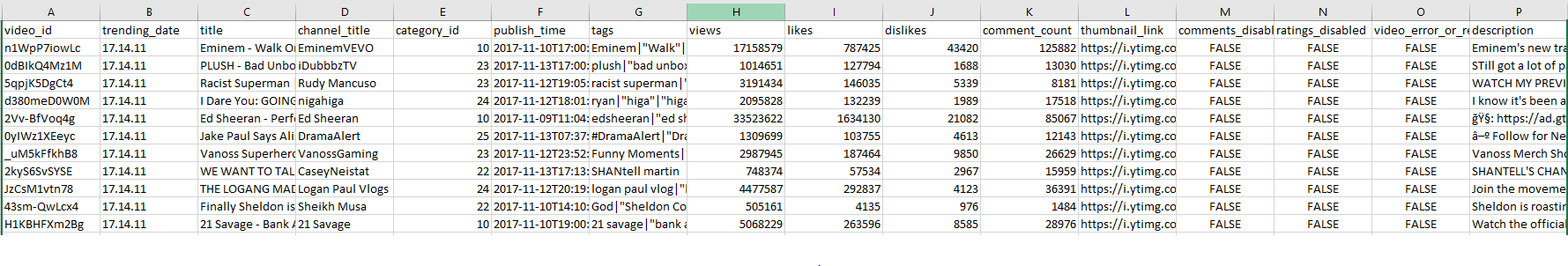


Result (100 comment and author):



1. **Preprocess datasets**

The process we just did was only for 1 video which returns 100 comments according to video\_id. Our aim is, do the same thing for all the videos on our datasets. By doing that we will get lots of useful columns like video\_id, trending\_date, title, channel\_title, publish\_time, tags, views, likes, dislikes, comment\_count, comments\_disabled, ratings\_disabled, video\_error\_or\_removed, description **+ Comment, Author\_name (Above)** for each row in dataset.

****

In total we have 10 datasets which contains 40,000 data each. If we apply gathering data process to all, at the end we are going to have 40,000\*10\*100 which is **40,000,000 data** in as 1 dataset that contains everything. **For the upcoming week our plan is to create this huge dataset by using YouTube API on rapidapi**. Then we will upload the dataset to Hadoop Distributed File System, and it will be ready for analyses phase. Codes are [here](https://github.com/mnadirs/e-big).

1. **Analyze phase**

We are going to implement the machine learning to category id column in data set. Category id column tells us the community of videos. As we have the comment and their author we can simply connect the communities and users because whether the comment is good or bad it means that if a person commented on a video means he or she is related to that community at one point. By analyzing the category id, comments and tags we will apply classification algorithms here. Then by analyzing tags and comment columns we will try to specify the community subgroup and make predictions for users like which class and which subcategory they belong to. This is an example image for our categories json file we are going to use for main user classification. Codes are [here](https://github.com/mnadirs/e-big).

.