

TED UNIVERSITY

Faculty of Engineering

Department of Computer Engineering

**Weekly Report**

by

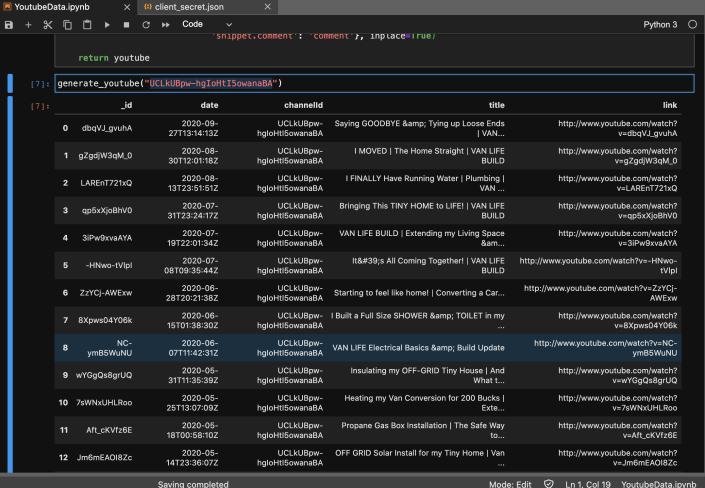
Yağız ÇİMEN – Efe TONTU – Mesut Nadir SEYFELİOĞLU

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

1. **Data Collecting Processing**

We can import youtube datsets by using google apı in python language. Our aim is to search the video comments by traversing id column on our set. By doing that we have all the video information and video comments (with user information). Then we will mix these 2 sets together.

At the end we will have a quite big dataset which contains youtube video information and their related comments by user (with user info).



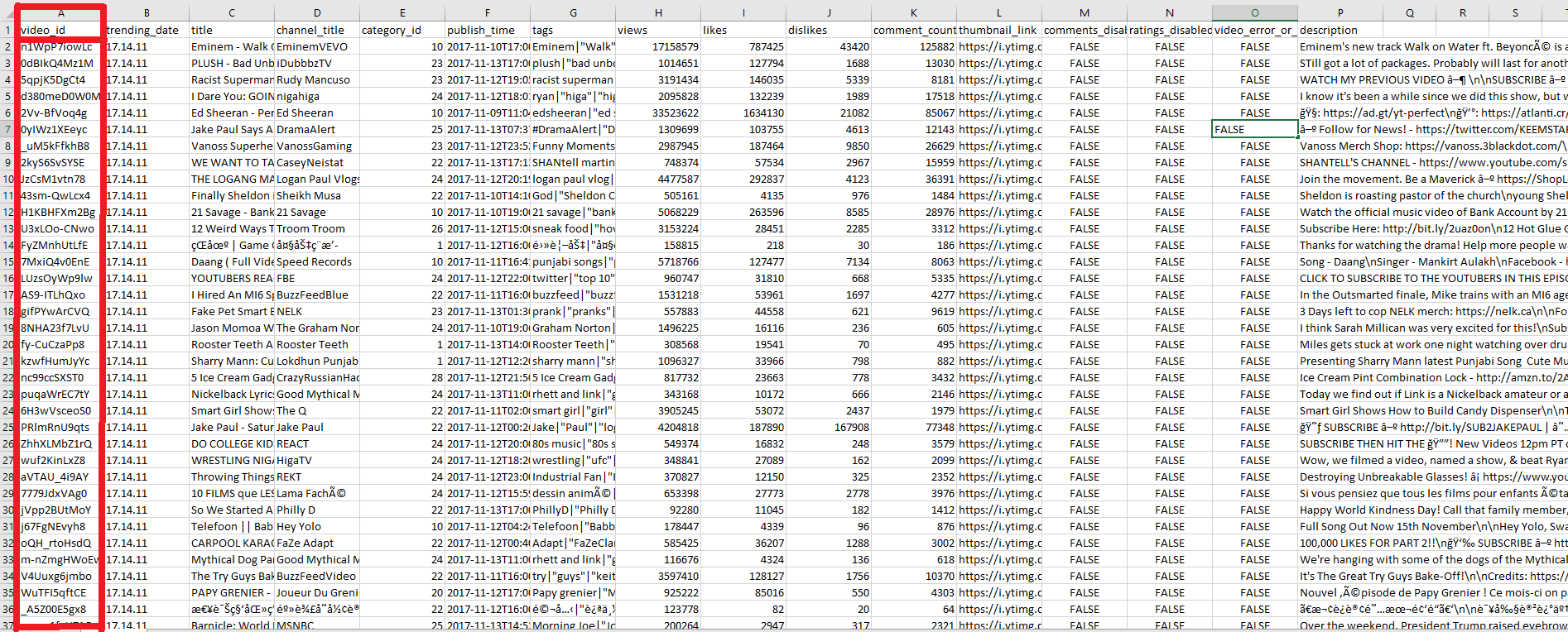
1. **Python PyCaret Library**



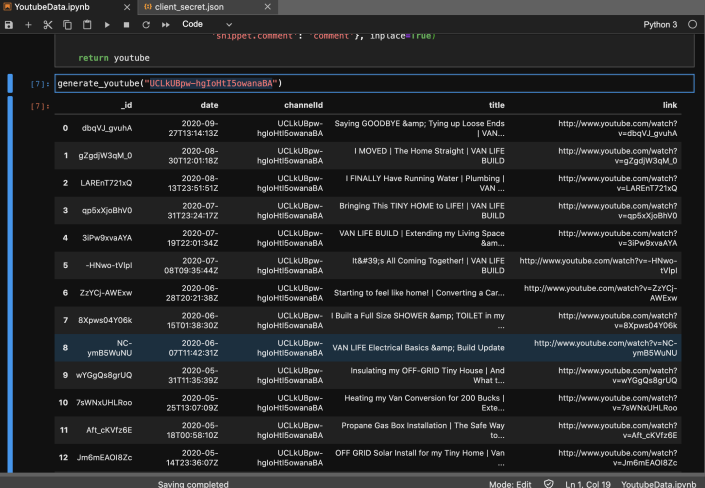
Link to lib: <https://pycaret.org/>. We found a library that analyses the given dataset by implementing every machine learning algorithm on dataset. Then reports the results by showing graph and performance reports. Our aim is to give the mixed dataset (Done in first part) to library and get the results for all ml algorithms.

**AFTER FEEDBACK**

These graphs are from main website of pycaret they are NOT our analyse reports, we just want to show the example images of results. First, we need to gather more data from youtube api which contains user id and comments. Then we take rows of videoID column one by one and add data to our main dataset which is

**Dataset 1: We already have**

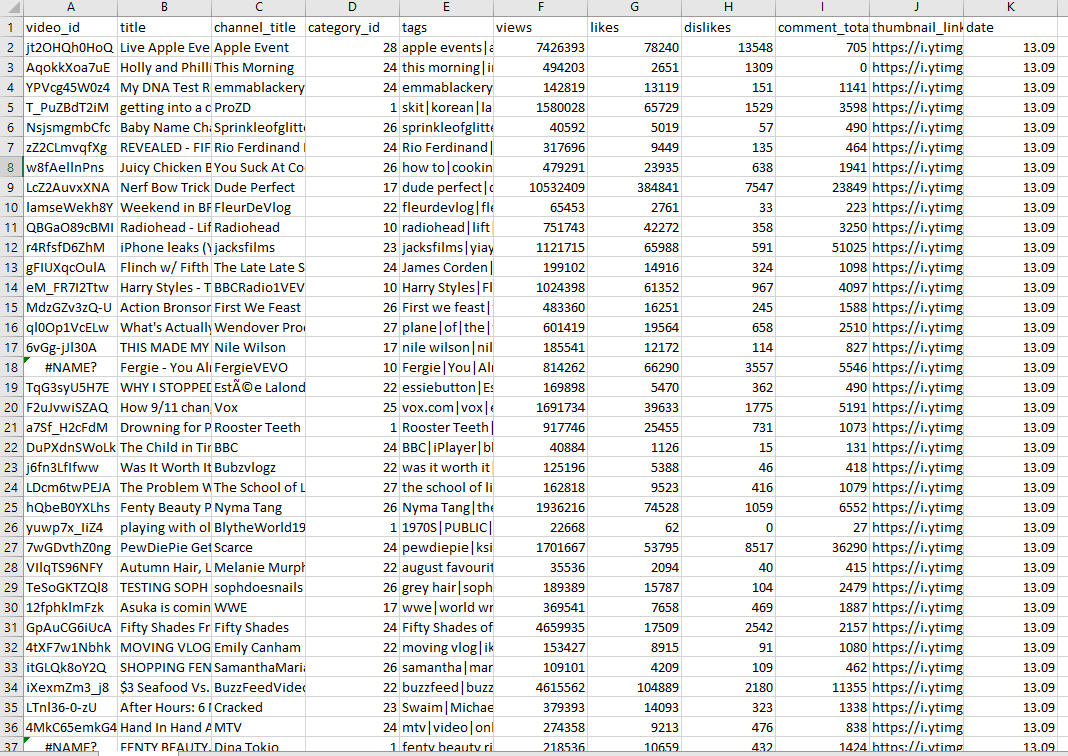
**Dataset 2: We want to add**

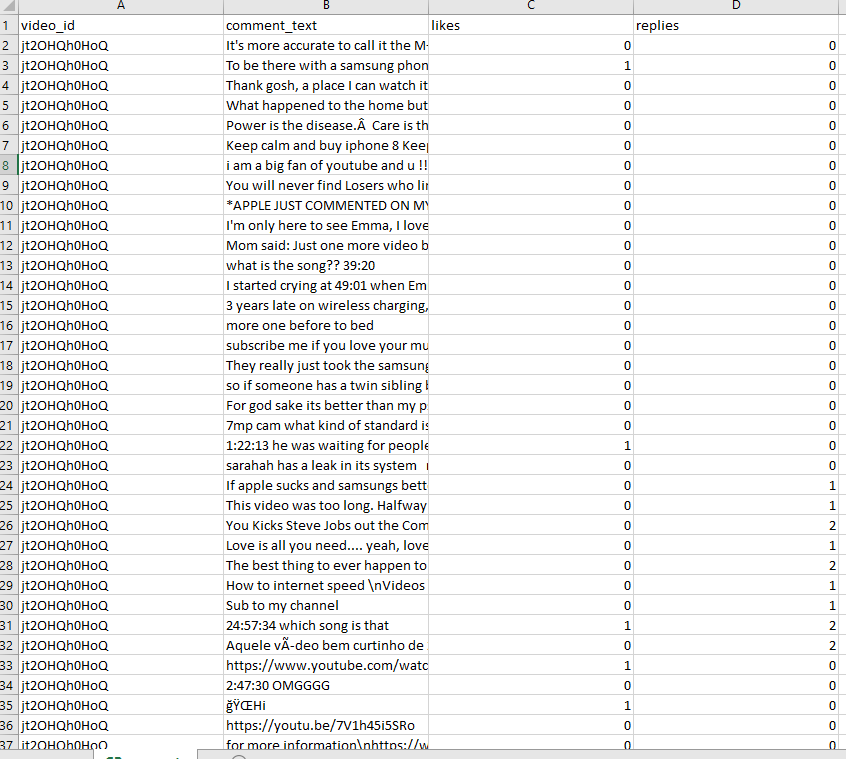


In second photo we wrote a python code which can find video comments. Search type can change to anything video id, channel title… Our aim is to get A comment dataset also contains user id. Then we will merge these two into 1 graph by looking videoID probably not decided clearly yet.

On the other hand, we have another data which contains youtube video features and 100 user comment for each video (Without user id). Which is

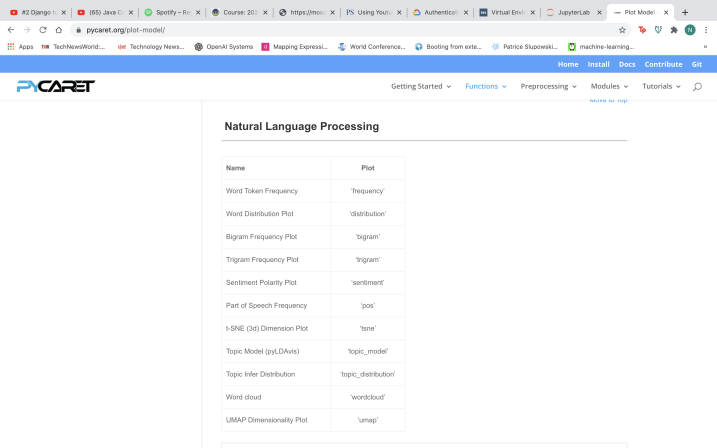
**Dataset 3: We already have**

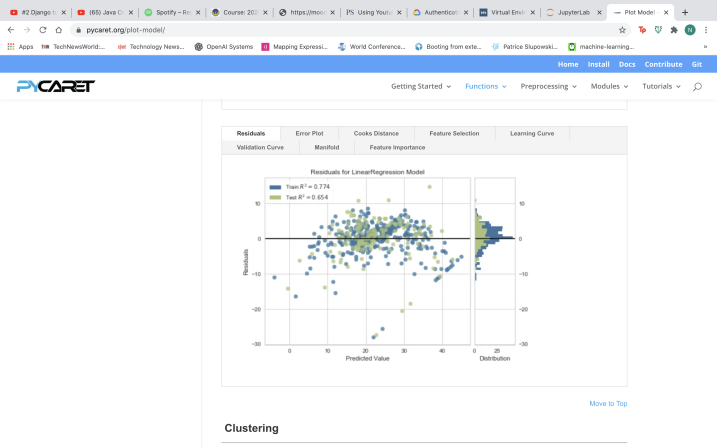


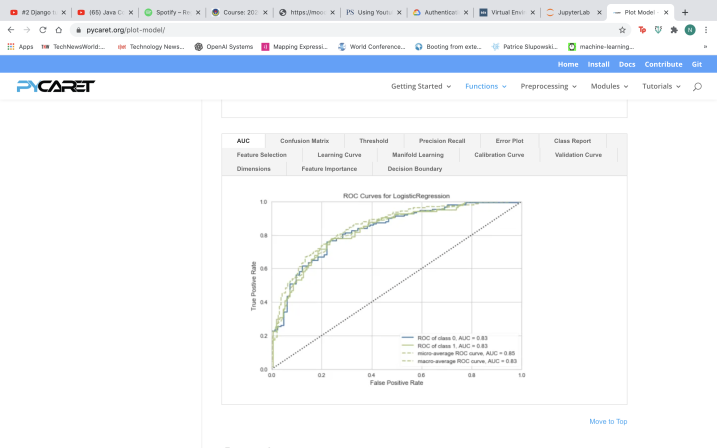
**Dataset 4: We already have**

We will process the datasets to have just 1 big data set which contain everything with comments, user info, video info etc…









1. **Implementation to Spark MLlib**

After analyzing the result for pycaret we will select the algorithm with best performance and implement it to spark mllib in java. Probably logistic regression is going to fit our data.

1. **Performance Comparison**

We aim to work in both languages (java, python) and compare the result for both in scale of efficiency. It can help us to pick the best things in both technologies.

1. **Different platforms for Product**

Our product is going to work something like basic user interface. We will want user to pick a platform from combo box, and enter the keyword (Nasa, coffee…). Then analyses and reporting will be done in background. We plan to do this analysis to different platforms in real world like google, Linkedin, twitch, Instagram…

1. **Tools**

**For android OS:** Android Studio

**For IOS OS:** Angular and swift

**For python analyse:** jupyter notebook, jupyterlab (lib: pycaret, googleapiclient, google\_auth\_oathlib)

**For Java anlayse**: Intellij IDE (lib: Apache Spark)

**Database:** Hadoop HDFS

**Website:** Angular, Javascript, DJango