Youtube trending prediction

SW24/2015 Jelena Maravić, SW60/2015 Sara Petrović

1. Motivation

As the second most visited video sharing platform on the web, YouTube shows all the most popular and currently trending topics from around the world. It is host to the latest news, current events, and viral content. Its popularity was our main motivation but also challenging dataset.

2. Research questions

Our main goal was to find out how long video can live on the trending list but also how much views it can have.

After searching for the dataset we had a lot of troubles. It had too much-unneeded features and it was too large so our task was to fix and clean up the data.

Song's data from its first showing on the trending list was remembered. Also, we had to count the maximum number song's consecutive days on Youtube trending list.

All non-numeric data was translated into numeric for its easier manipulation.

3. Related work

We have read almost all the discussions people have about our dataset and the prediction they had done. We did not come up to prediction we decided to do for the video trending but we have found out that people tried to predict the views of the video. They said that their main problem was the conversion of data.

4. Methodology

For the methodology, we decided to split the problem into several parts. The first part was for processing data - counting the trending days, remembering all important and removing unnecessary data like ids, video errors, titles and more.

The second part was for the training data and prediction. Before we got our model, we used the grid search to find out which parameters were the best for the concrete classifier.

For the last part, we used the F1 score micro to find out which classifier gave us the best value.

5. Discussion

For the testing procedure, we tried to find out which parameters give better result for each classifier. This is done by grid search.

After that we used bagging to choose which classifier is the best. All classifiers gave similar results, but SVC gave us the best result.

6. References

<https://www.kaggle.com/datasnaek/youtube-new>