Automatic Tag Recommendation In YouTube

Afshin Moin (afshinm), Abhishek Bharani (abharani), and Peng Seng Kuok(pkuok)

Content sharing has been the key ingredient of the biggest success stories in the last decade. Social networks like Facebook and LinkedIn, video sharing services like YouTube, Ecommerce websites like Amazon and online travel agencies like TripAdviser and Expedia are all examples of online platforms that provide a convenient way for their users to exchange content, goods or services.

It is obvious that the success of such online platforms depends on how effectively they link the end users to relevant content. This latter cannot be achieved without proper content classification and user feedback analysis. This motivated us to apply machine learning techniques to automate the process of user feedback analysis and content classification.

We have identified a number of challenges we are planning on to deal with in this project. We will apply our solutions to the YouTube sample database. Nevertheless, same techniques can be applied to any other database with similar structure like Yelp and Netflix. The databases can be downloaded from https://www.kaggle.com/datasnaek/youtube. They include daily data from the 200 most trending YouTube videos in US and UK. The databases contain video title, channel title, category, tags, number of views, likes and dislikes and user reviews.

Sentiment analysis based on the reviews is a potential area of interest. We will extract features from reviews and use classification techniques such as logistic regression and SVM to predict whether a review bears positive or negative meaning. We will use likes and dislikes of the corresponding users as data labels.

Tags are an important type of data that can be used by search algorithms to enhance the quality of results. We will predict the category of a video given its tags. Using a generative method, the distribution function of tags given video category is computed. This distribution function may be used for automatic tag recommendations. Same techniques can be used to predict the distribution of tags based on review features. Likewise, text analysis techniques like TF-IDF lend themselves well to estimate the relevance of a tag to a video based on its reviews.

We will train the models on a number of releases of the YouTube database. We then test them on a few other releases. Measures like precision, recall and F-score will be used to compare the performance of different techniques with each other. Splitting a database into training and test set and doing cross validation is another way of evaluating our models.