You Tube Recommender

# YouTube Tag Prediction and Analysis:

## Classifying Video Features for a Refined Recommendation System

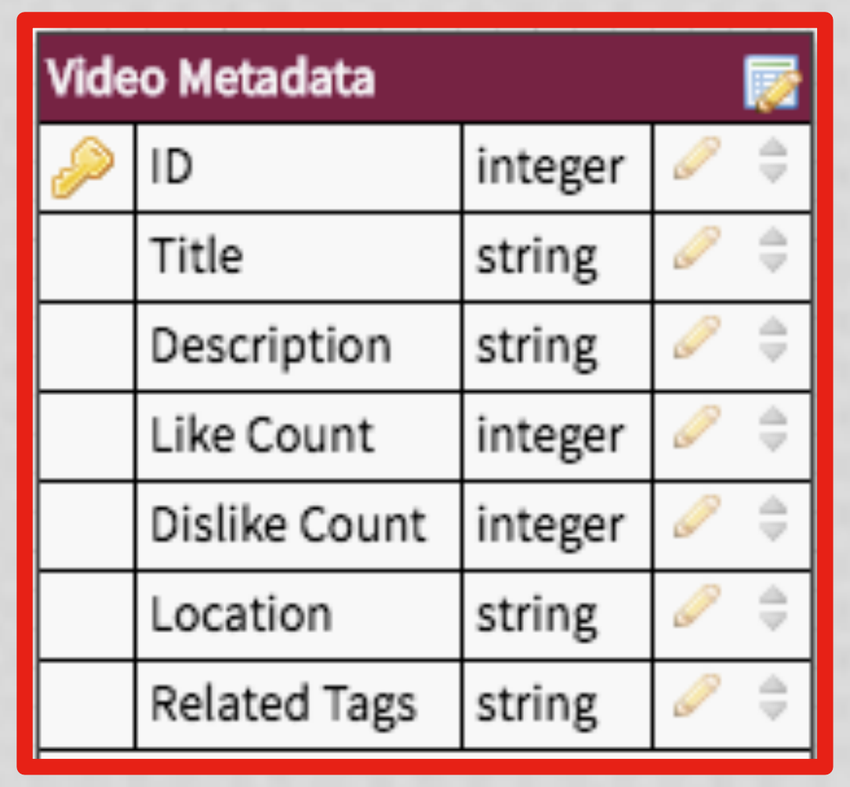
Abhishek Dutta, Preston Law, Natalie Roe, Aaron Wu, Wennie Zhang

## Vision

- Use user-defined tags as a basis to build a predictive model that assigns tags to YouTube videos while eliminating user error
- Apply tags as a basis for analysis of YouTube trends over time in various regions of world while considering sentimental variability (i.e. positive or negative popularity)

## Dataset

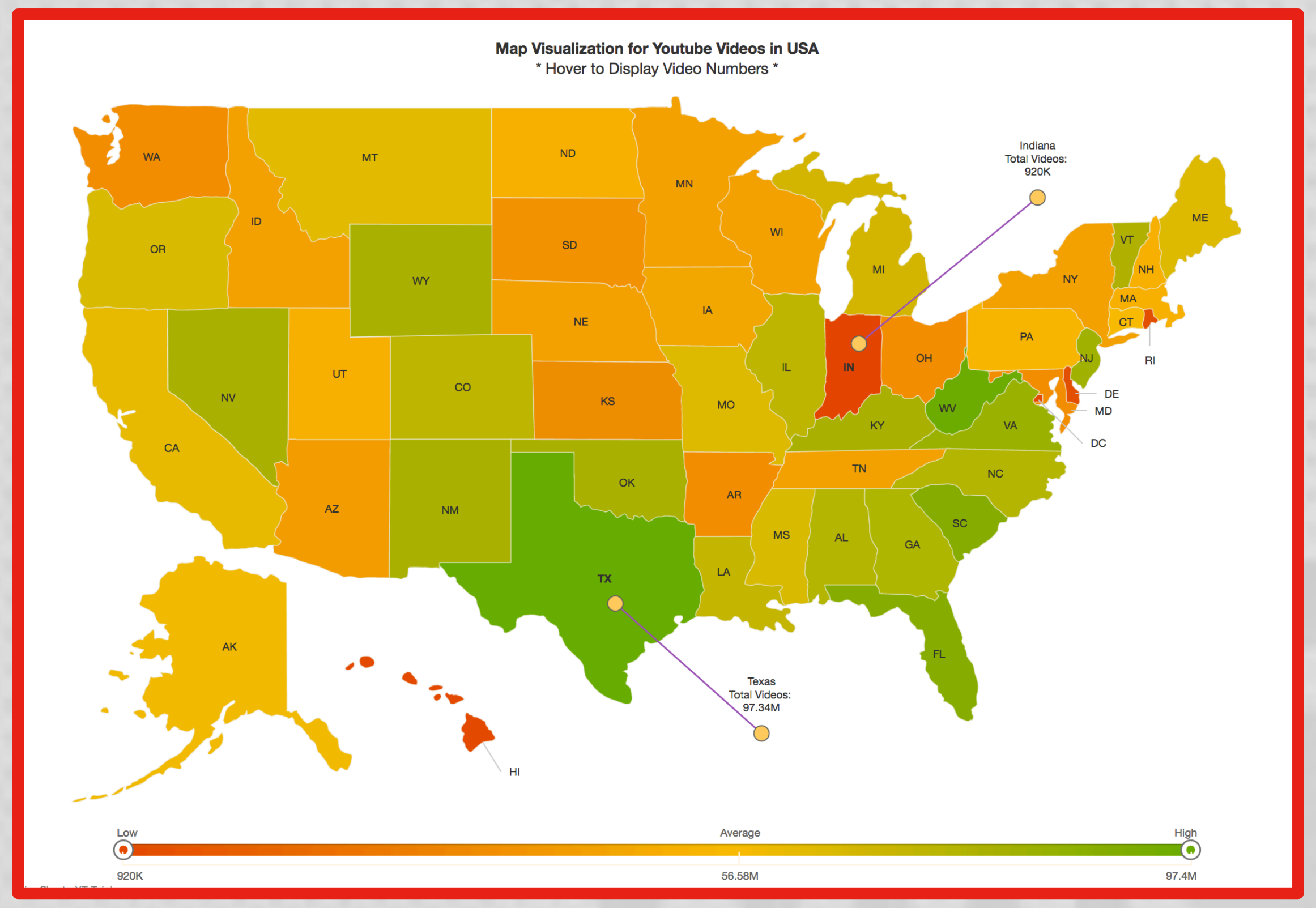
- Extracted data from YouTube data API
- Dataset consists of videos uploaded since March 1, 2017
- Training set: 90% of videos
- Test set: 10% of videos

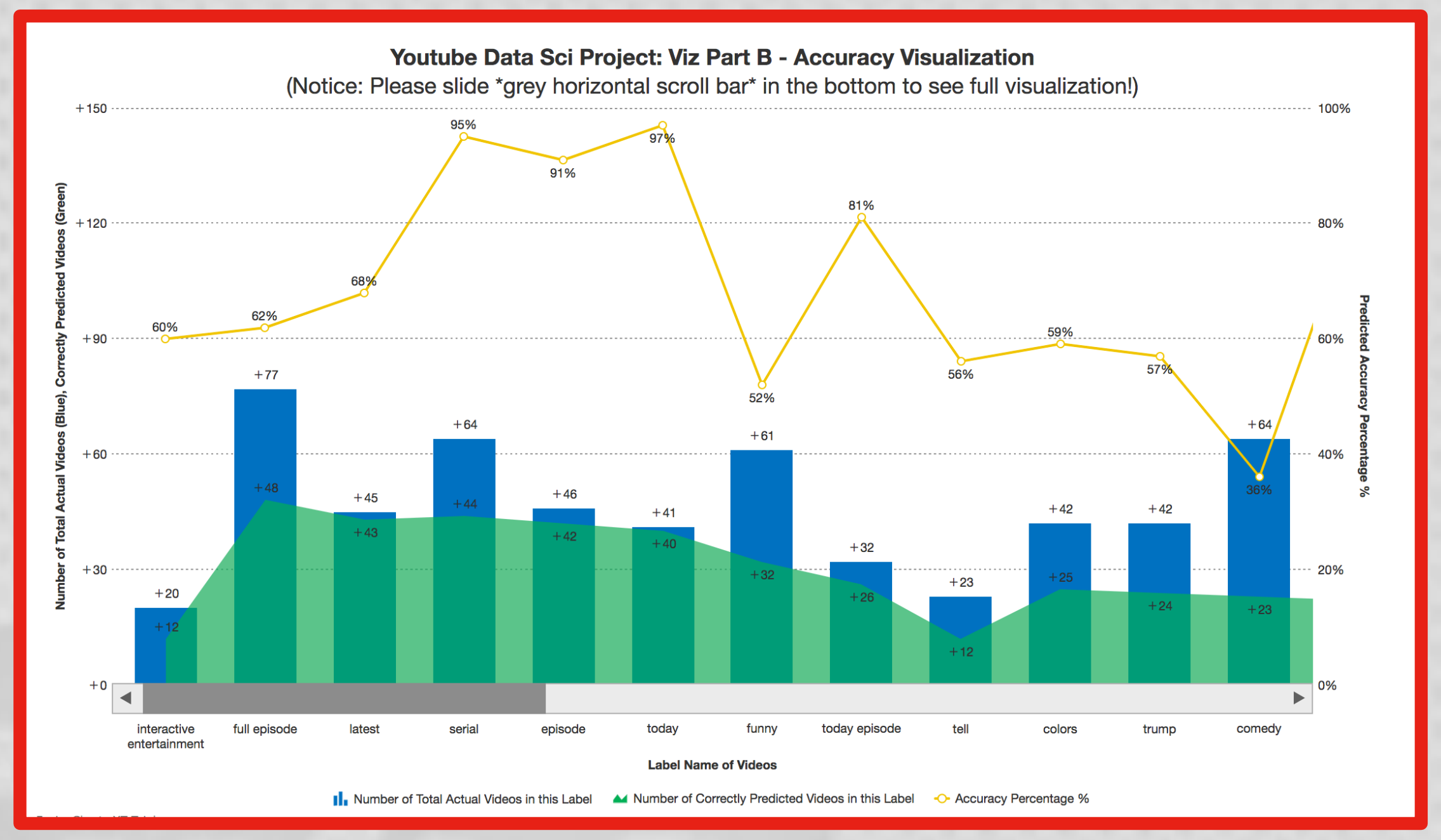


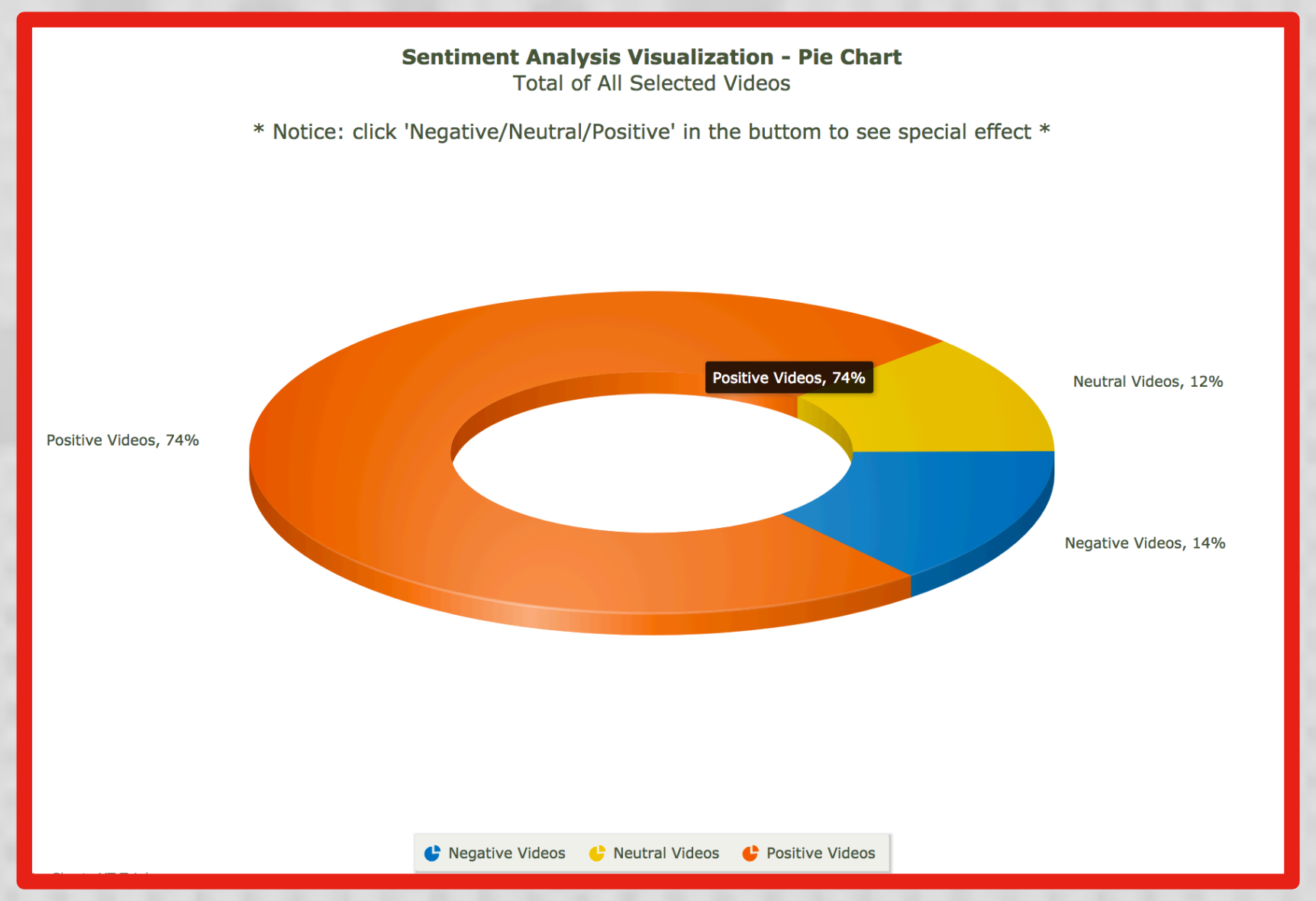
| Video Metadata | |
|---|---|
| ID | integer |
| Title | string |
| Description | string |
| Like Count | integer |
| Dislike Count | integer |
| Location | string |
| Related Tags | string |

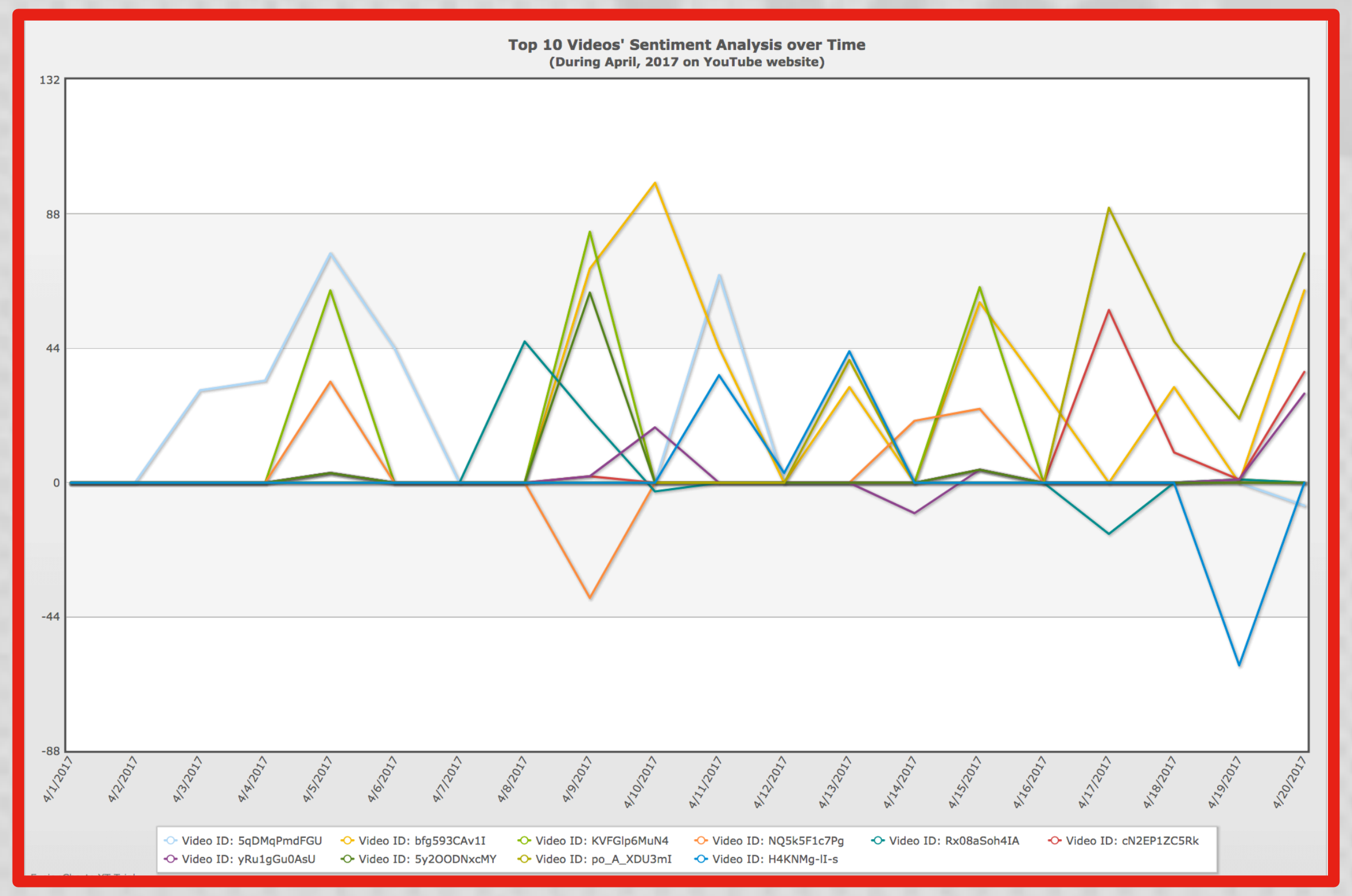
- Relevant metadata: ID, Title, Description, LikeCount, DislikeCount, Location, RelatedTags
- Extracted comment feed for top ten videos in each tag for sentiment analysis

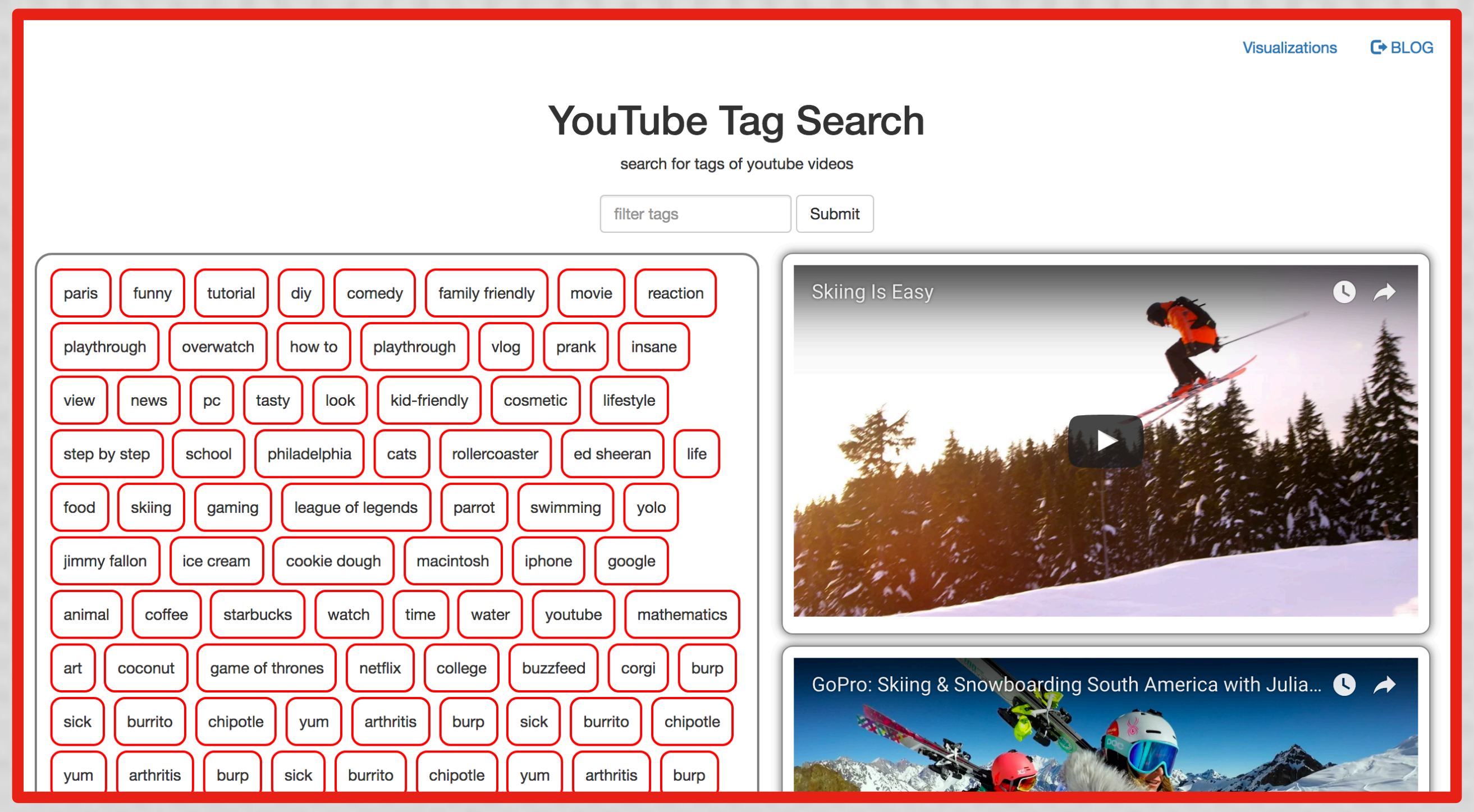
## Visualizations


Map Visualization for Youtube Videos in USA
* Hover to Display Video Numbers *
Indiana
Total Videos
Texas
Total Videos
Low
Average
High









## Results

- Model was able to predict labels accurately
- Over time generally videos with more positive sentiments trend
- User recommendation system accurately recommends top ten videos for a particular tag and videos maintain relevance to user query
- Viewers generally do not consider videos as neutral

## Methodology

- Stored cleaned data in MySQL database
- Condensed user-defined tags using synonym matching
- Used multi-label learning model which maps inputs to binary vectors rather than scalar outputs to predict multiple tags per video
- Tailored performance evaluation to suit multi-label classification by computing average recall, average precision, and average F1 score
- Performed sentimental analysis using Natural Language Processing (NLP) and Natural Language Tool Kit (NLTK)
- Computed top videos using likes-to-dislikes ratio

## Challenges

- Identifying a multi-label classification system
- User quota from YouTube API limits amount of extractable data per day
- Overcoming languages differences in videos
- Dealing with poorly-formatted user-defined labels such as “funny parrot” which should be split
- Limited quantity of data

3:05:45 3:09:03