Greg Szumel and Erik Kessler
CSCI 375: Final Project Proposal

Overview

For this project, we plan to build a classifier that, given two YouTube videos with the same content, will predict which video will get more views based on its title and description. Video creators try to make their videos sound enticing by putting a flashy title that gets a particular user to feel interested or curious enough to click on the title. Views are not just a representation of popularity, though, they also represent how much a creator will be paid for their content. So, classifying titles by how many views those videos could accumulate would benefit the YouTube community as a whole: content creators could get more views and more money, and YouTube would be able to increase the total number of videos people watch. Ideally, we would also be performing an analysis of how "clickbait" is impacting YouTube.

Existing Work

Youmna Borghol, Sebastien Ardon, Niklas Carlsson, Derek Eager, and Anirban Mahanti. 2012. The untold story of the clones: Content-agnostic factors that impact YouTube video popularity. *In Proceedings of KDD*.

> Looked at how features like social networks and view count histories impact view counts. We look to build on this by including titles and descriptions. We will use their dataset of duplicate videos.

Himabindu Lakkaraju, Julian McAuley, and Jure Leskovec. 2013. What's in a name? Understanding the interplay between titles, content, and communities in social media. *In Proceedings of ICWSM*.

> Identifies linguistic features of titles that impact popularity.

Chenhao Tan, Lillian Lee, and Bo Pang. 2014. The effect of wording on message propagation: Topic- and author-controlled natural experiments on Twitter. In *Proceedings of ACL*.

> Exploring a similar idea but used tweets and retweets instead of titles and view counts.

Formulation

We plan to formulate this a classification problem. Thus, we will focus on developing features and use some off-the-shelf classifiers such as Naive Bayes. We want to try to control for creator subscriber count and ideally, content of the video as well, through a carefully crafted dataset.

Implementation

Dataset: http://www.ida.liu.se/~nikca89/papers/kdd12.html (YouTube videos grouped into "clone sets")

Additional Data: May be gathered manually through seeking recent 'clone' videos.

Programming Language: Python

Initial Libraries: NLTK, Stanford NLP Libraries, NumPy

Timeline

Now-5/3	Dataset collection & project setup
5/4-5/14	Features/Classifier Implementation
5/15	Test and evaluate
5/16	Outline final report
5/17-5/19	Write final report
5/19	Final report & code submission