CSCI 544: Applied Natural Language Processing
Course Project Proposals Jonathan May and Nanyun Peng

1. What is the name of your task?

Age Classification of Youtube Captions.

2. Why this is an interesting data set? What would be learned from building a good classifier of this data ?

This project will attempt to annotate and classify Youtube videos taking into account the content of the video and its composition. While youtube flags content inappropriate for young audiences by requiring viewers to sign in, a lot of youtube content is generally unaudited if the uploader of the video does not flag it so. Also there is no distinction between which content is appropriate for what age groups. We will classify content based the film rating system: G, PG, PG-13 and R. We will also apply a binary classification for classifying clickbait videos.

A public corpus of youtube captions which analyses a large volume of data does not exist. There is a huge opportunity to annotate massive amounts of data and to gather insights with respect to the kind of content (read clickbait) and the kind of audience it appeals to.

3. How will the data be collected and labeled ?

We have been able to successfully collect data 100 youtube videos from Casey Neistat's channel. We have used a third party service to upload the url and get the captions of a video. We also currently also exploring extracting youtube captions directly using its publicly available API.

The data will be labelled with respect to the film rating guidelines; violence, abusive language, substance abuse, etc. We have come up with the list of words and kind of language (after referring the guidelines) that we will be using to classify. We will be labelling data manually, and will also be using helper scripts.