Preliminary Proposal

Alice Chang

Preliminary Datasets

1. Considering the substantial extent of undocumented and normalized instances of police brutality, can we build a model to predict the approximate number of police shootings in a state? Regression. Predictive. We will train our model on data gathered through the Washington Post's database of fatal shootings by U.S. police officers since January 2015 (https://www.kaggle.com/washingtonpost/police-shootings). We will consolidate the database with U.S. census data describing demographics of U.S. cities (https://www.kaggle.com/kwullum/fatal-police-shootings-in-the-us).

2. Can we build a model to predict whether and how a Reddit comment is "toxic"? Classification. Predicive. We will train our model on a dataset of May 2015 omments released by Reddit (https://www.kaggle.com/reddit/reddit-comments-may-2015/download). The final training dataset would likely involve our own qualitative coding and classification of comments in the existing data from May 2015 into "safe"/different categorizations of "toxic." Categories may include "insult," "threat," "obscene," and "identity-based hate."

3. Can we build a model to predict which Youtube video category a video falls into based on other features of the video (eg. comments, tags, views, likes etc.)? Classification. Predictive. We will train our model on data gathered through the Youtube API. The dataset provides a daily record of statistics and comments of top trending Youtube videos in the U.S., Great Britain, Germany, Canada, and France (https://www.kaggle.com/datasnaek/youtube-new). There are 29 categories on the Youtube video category list, including categories such as "Film and Animation," "Music," "Education," and "News and Politics."