Morgan Baccus and Gabriel Byram
CptS 315
Project Proposal

Toxic Comment Classification

1. Data Mining Task:

Our data mining task is to create an algorithm that can categorize Wikipedia comments based on whether they are toxic, and which category of toxicity they fall under. The categories will be clean (not toxic), toxic, extremely toxic, obscene, threat, insult, and identity hate. The motivation behind choosing this problem is to create an algorithm for social media platforms that is able to hide toxic comments once they are labeled. Hopefully by doing this, we will be able to reduce negativity and cyber-bullying on social media.

2. Dataset:

We obtained our data from Kaggle's Toxic Comment Classification Challenge. Kaggle provided separate .csv files for the training and test datasets. These datasets are provided under the data section of the challenge on Kaggle:
https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/data

3. Methodology:

We will first parse the data into a format we can use. We will use NLTK to tokenize each word from each comment, and remove punctuation and words without meaning on their own, such conjunctions and articles. We will also break each word in a comment down to its stem using NLTK.

We then will train our program to identify the different categories of toxicity based on the testing data we have parsed using a machine learning library, such as TensorFlow. We will use Numpy and Pandas to store our data into data structures we can more easily manipulate.

Once our program has been trained, we will test it by allowing it to categorize each comment from the testing data, and then checking to see how accurate its categorization was by comparing its categorization to the actual categorization of the comment. After we obtain our results, we will use Matplotlib to represent our data using graphics, such as pie charts and bar graphs.

4. Final product:

For the outcome of this project, we want to provide a template for filtering out negativity and toxicity in online communities. If toxicity in online communities can be reduced, the decrease in mental health associated with harassment and online bullying may be reduced as well. This could lead to overall improved mental health in communities with high online involvement. We will measure the success of our project based on whether we can create an algorithm that can accurately measure levels of toxicity in a comment. We can measure the accuracy of our algorithm by creating a graph to compare the scores that our algorithm assigned the comments from our test data to the scores that the comments from the test data actually received. This project will help us explore how to better define toxicity in online communities, and learn how to create a more accepting online environment.