Project Deliverable 2

Ze Dian Xiao

February 24, 2019

Problem Statement In my first project deliverable, I had thought I could possibly create a machine learning model for pdf to LaTeX. However, under Isaac's guidance, I chose another project, which consists of insincere comment classification. The data consists of Quora questions, with binary labels 0 and 1.

Data Preprocessing Dealing with text data is often very complicated in terms of data preprocessing, especially when the dataset contains 1.3 million entries. To preprocess my data, I first cleaned it by removing all elements that are non alphabetic or spaces. Then, I removed possible contractions. And finally, I lemmatized and removed stop words as well as punctuation all of this in order to get cleaner data, which has less noise. Unfortunately, I greatly underestimated how much time this preprocessing part will take me. Hence, I was unable to actually fit my model to it in time, maybe by the time you check my git it will be there.

Machine Learning Model This part will come soon!