DS-GA 1003 Project Proposal: Extreme Classification

Group Members

Andrew Yeh *ay1626*, Eelis Virtanen *ev933*, Mrinal Jain *mj2377* **(submitter)**, Alec Hon *abh466*

Chosen Project

We have chosen the **Extreme Classification** project. Legal domain documents have been tokenized and marked with multiple labels. The resulting dataset has 5000 features with approximately 4000 classes. Each feature has been encoded into a sparse format, and our goal is to create a model to predict the probability of our inputs belonging to each of the classes.

Proposed Approach

To start out, we will do some initial feature engineering on the raw data. The primary goal of our feature engineering is to find better representations of the data either by making things more easily separable or by increasing the computational efficiency of training with a more compact representation.

Next, we will try a few simple baseline models applicable to such high-dimensional data. This is just to get an idea of what models work given the unusual class distribution--there are only 25.7 examples per label--and see which ones have some predictive power to iterate on.

Last, based on exploratory results, we will decide on 1-2 models for hyper-parameter tuning to achieve a score for every one of the approximately 4000 labels. We may re-iterate over the feature engineering process if necessary.

To ensure model robustness, we will use a training and test split on the available data and use some cross-validation techniques on the training set to tune the hyper-parameters.

Suggested Experiments

Some initial feature engineering ideas are applying some dimensionality reduction techniques (like PCA, t-SNE) to find features that linearly combine many tokens that vary in the same way and potentially co-occur. This reduction in the dimensionality of the feature space can help to deal with computational problems that can occur from these issues.

Some baseline models we may try include:

- Logistic Regression
- K-nearest neighbors
- Naive Bayes
- Support Vector Machines
- Bagging best models

We will evaluate which one of these models performs well using LRAP (label ranking average precision score) before deciding to follow through with the best performers.