EECS 445: PROJECT PROPOSAL

NAND DALAL
CHARLES LEWIS
JEFFREY LIN
ERIC TASEKI

1. Problem statement

StumbleUpon is currently hosting a competition on Kaggle. They have released the following problem statement: "StumbleUpon is a user-curated web content discovery engine that recommends relevant, high quality pages and media to its users, based on their interests. While some pages we recommend, such as news articles or seasonal recipes, are only relevant for a short period of time, others maintain a timeless quality and can be recommended to users long after they are discovered. In other words, pages can either be classified as "ephemeral" or "evergreen". The ratings we get from our community give us strong signals that a page may no longer be relevant - but what if we could make this distinction ahead of time? A high quality prediction of "ephemeral" or "evergreen" would greatly improve a recommendation system like ours."

2. Significance

As web applications such as StumbleUpon have proliferated due to more people having access to the internet, developers of these web apps have started to focus on how to retain their user base. Personalized and relevant content contribute heavily to user retention. This can be achieved in multiple ways. First, as StumbleUpon has experimented with, users can rate the content they are presented, which will gradually increase the proportion of relevant content they get as they use the application more. However, this does not make sense when StumbleUpon is trying to retain new users. A better method to provide relevant data is to automatically provide relevant data without having to learn about whether the content is relevant or not from the user base. However, the algorithms derived from this project can be applied to a more general problem of web page classification.

3. Related work/Novelty

StumbleUpon's competition on Kaggle relates to an important application of machine learning: web page classification. Much work has been done by researchers looking to classify web pages while crawling the web. For instance, when Googlebot crawls the web it indexes web pages according to genre of the web page in order to improve its search results. Similar algorithms have been implemented by web crawlers seeking to provide targeted content to users. One of the key problems in classifying web pages is the ability

to distinguish irrelevant from relevant words. In terms of machine learning, this applies to reducing the representative feature set or the dimensionality. We will have to implement similar algorithms and can refer to the ones used by web crawlers to understand how classification is achieved. However, the algorithm will need to be fine tuned to fit StumbleUpon's problem.

4. Proposed method/approach

The data set provided by StumbleUpon and Kaggle includes plain text web pages as well as binary evergreen labels indicating whether or not the web page would be considered an "evergreen" web page. We will start the project by performing a basic classification algorithm like Naive Bayes. This will allow us to categorize certain groups of words by whether their occurrence together on the same web page indicates the label for the web page. We can expand on this method by adding features to our data set including relative distance of words within the same web page. As we investigate the results of Naive Bayes, we will get a better idea of how to tweak our feature set. The majority of time spent on this project will be devoted to analyzing the result of applying various algorithms to the data set. As in many cases, the final algorithm may be a combination of multiple models.

5. Evaluation

Kaggle provides testing data in addition to training data for us to calculate the error in the algorithm. We will also implement cross validation, specifically k-fold cross validation, to estimate the true errors as the average error rate. This method of cross validation will allow us to asses the final model or combinations of models using the testing data. In order to make our submission to Kaggle, we will produce a csv file with the urlid and a label. According to Kaggle, "Submissions are judged on area under the ROC curve."

6. Data and Resources

- `http://www.kaggle.com/c/stumbleupon`
- `http://www.webology.org/2008/v5n1/a52.html`
- `http://www.cse.lehigh.edu/~xiq204/pubs/classification-survey/LU-CSE-07-010.pdf`