Project Proposal Final Draft 1st Quarter 2009-2010

Edwin Zhang

October 29, 2009

1 Title of the project

Learning to Classify documents

2 Purpose and scope of the research project

The goal of this project is to use a Machine Learning technique to automatically build a document classification model from a set of training documents and the model is able to determine what an article is talking about based on the words that are in the article. I dont plan to include every possible subject in my program, which would be very difficult and time-consuming, but I do plan to include a good number of categories, such as sports.

3 Background and review of current literature/research in this area

Ive read the NLTK website and my dad has explained to me on the algorithm for the project. Also, I have read a chapter from a Machine Learning book. Also, my second Lit Review paper discussed text classification a little and talked about the algorithm and what text classification was.

4 Procedure and Methodology

The Machine Learning algorithm used in my project is Nave Bayes Classifier. Nave Bayes Classifier computes the conditional probability p(T—D) for a given document D for every topic T and assigns the document D to the topic with the largest conditional probability. Nave Bayes Classifier converts the calculation of the conditional probability into a formula that is easy to calculate using the Bayes rule. For my project, I will first write a program to select a subset of relevant features (terms) that can best distinguish one topic from another, the probability a selected term occurs in a document from a given topic is computed using training documents. The conditional probability p(T—D) will be computed based on the probabilities of all se-

lected terms that occur in the document. I will use Python/Java for my project.

5 Expected Results & and Value

This program will be very helpful to people because it will allow them to identify the topic of a document without reading the document. This can be used to determine, for example, if an e-mail is spam or not or whether a document is important to your task.