CS 440: Project 2
A Document Classifier
Due April 26th

Gabe Dunham, Jeff Milner, Ryan Quackenbush

Abstract

A brief summarization the paper in a single paragraph. This documents aims to discuss the strategies our group took in implementing a basic document classifier between three document types: Deed of Reconveyance, Deed of Trust, and Liens. We will use three different strategies and to parse and determine a type for each document in the test files given.

Introduction

Given we have three different document types: Deeds of Reconveyance, Deed of Trust, and Liens, we are to parse through all the test data and be able to determine what type of document they are. We are given three strategies: intelligrep, naive bayes, and perceptrons to evaluate and test the documents on through training, if necessary, and testing. The results will then be used to determine the percentage of the test data which is classified correct. The data provided in for each classification will be used to train the naive bayes and perceptrons. A TEST data directory will contain all necessary files to run our tests and compare against a results file to display the percentage of documents classified correctly.

Initial Strategies

IntelliGrep Strategy

The initial strategy was to classify each document on whether a certain word / phrase appeared most often in the document. The words used were "Deeds of Reconveyance", "Deeds of Trust", and "Lien" for Deeds of Reconveyance, Deeds of Trust, and Liens respectively. This strategy on our test data came out with values around 15 percent accuracy overall. Initially this was due to attempting to run the algorithm on the training data. Switching our data around to work on the test data improved our accuracy to about 60 percent.

Naive Bayes

The base strategy for Naive Bayes worked out reasonably well to start. The bag of words was initially not boolean, therefore causing issues with the output not being correct

Perceptron

Improved Strategies

IntelliGrep Strategy

Naive Bayes

Perceptron

Conclusion

Based on your results, argue which of your classiers you expect to perform the best on the hidden data set and what elements of that classier you view as