An Investigation of Multiple Language Classifiers COMP 551 - Applied Machine Learning

Thomas Page
Department of Mechanical Engineering
McGill University
Montreal, Canada
thomas.page@mail.mcgill.ca
260771672

David Tamrazov
School of Computer Science
McGill University
Montreal, Canada
david.tamrazov@mail.mcgill.ca
260561439

Alexander Wong
School of Computer Science
McGill University
Montreal, Canada
alexander.wong4@mail.mcgill.ca
260602944

Notorious Language Classifiers
https://github.com/a22wong/notorious-language-classifier

October 23, 2017

I. Introduction

As a result of the first project, corpora in various languages are available to act as training sets for many language related machine learning tasks. The goal of this project is to create classifiers that can determine the language of a sample of text from corpora information developed in the first project. Based on a review of the methods discussed in COMP 551, we construct four classifiers: Naive Bayes, Linear Discriminant Analysis (LDA), Decision Trees, and Extremely Randomized Decision Trees.

The compiled training dataset contains 276516 labeled language samples from five different languages: Slovak, French, Spanish, German, and Polish. The test set contains 118507 unlabeled samples of zero to ten individual characters. This presented the challenge of classifying seemingly nonsensical data; a common task in practical machine learning.

Fig. 1

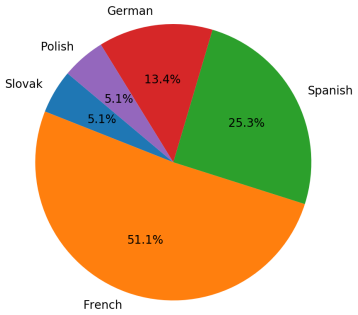

Fig. 1. Training Set Language Distribution.

A. Subsection Heading Here

Subsection text here.

1) Subsubsection Heading Here: Subsubsection text here.

II. Conclusion

The conclusion goes here.

Acknowledgment

The authors would like to thank...

References

[1] H. Kopka and P. W. Daly, *A Guide to LaTeX*, 3rd ed. Harlow, England: Addison-Wesley, 1999.