MAIS 202 Final Project- Deliverable 2

Tahsin Kabir, Omar Lahlou, Abderrahman Laoufi

Our project will consist of a language detection program that will detect the language of a short text. The scope of the project is limited to the 235 languages found on Wikipedia.

We will use the WiLi-2018 data set, as it contains 235000 paragraphs in 235 languages from Wikipedia and is already cleaned and separated into training and test sets. The size of the data set, as well as its convenience, were the main reasons for choosing it.

**We have not changed anything from the last deliverable, in terms of the dataset.**

The data is already processed, as it contains cleaned paragraphs for each language. It is additionally separated into training and data sets, allowing us to begin working directly on the model.

**Our model remains the same as well.** A Multinomial Naïve Bayes model is best suited for our multiclass classification problem.

We are currently trying to figure out the vectorize function for our problem, we are getting numbers other than binary values. We would require more time to fix the code and then we can send in the preliminary results.

This is also why we are not able to make a comment on the efficiency or accuracy of the code yet. We will be able to make a judgement once we have figured out the model.

Moving forward, our goal is to get the vectorize function working for our testing sample and to come up with ways to make it more time efficient, and hopefully run confusion matrix and accuracy loss.