Language Detection

Maddie Bauer
DSC 680-T301 – Summer 2021
https://madelinebauer.github.io/LanguageDetection/

Which Domain?

Language detection is often an important step within a Natural Language Processing (NLP) problem. The goal is to try and predict the natural language of a word, sentence or other piece of text because it is necessary to know the language of text before any other actions, such as translating, sentiment analysis or autocompleting text, can take place. Text sources can come from many places, such as emails, news articles, comments, speech-to-text applications, and so many more, where a language detection model is needed in order to know what to do next with the available text. For example, in a Word document the spell-checking feature must first determine what language the text is being written in before it can provide you with the correct spelling of a word [1]. Another use cases could include applying the right subtitles or closed captioning to a presentation or video as well as being able to route emails to the right geographically located customer service department for a large company [1].

References

[1] https://towardsdatascience.com/how-i-trained-a-language-detection-ai-in-20-minutes-with-a-97-accuracy-fdeca0fb7724 - Language Detection AI

[2] https://towardsdatascience.com/deep-neural-network-language-identification-ae1c158f6a7d - Deep Neural Network Language Identification

[3] https://www.analyticsvidhya.com/blog/2021/03/language-detection-using-natural-language-processing/ - Language Detection Using NLP

[4] https://towardsdatascience.com/benchmarking-language-detection-for-nlp-8250ea8b67c - has different python packages/functions listed

[5] https://www.upgrad.com/blog/natural-language-processing-nlp-projects-ideas-topics-for-beginners/ - NLP projects to consider

[6] https://towardsdatascience.com/an-efficient-language-detection-model-using-naive-bayes-85d02b51cfbd -Language Detection using Naive Bayes

[7] https://www.analyticsvidhya.com/blog/2021/05/interesting-nlp-use-cases-every-data-science-enthusiast-should-know/ - Why NLP is important/useful today

[8] https://www.section.io/engineering-education/five-real-life-use-cases-of-natural-language-processing-nlp/ -use cases for NLP

[9] https://en.wikipedia.org/wiki/Language_identification - Language Detection definition & background

[10] https://algorithmia.com/blog/build-your-own-language-detection-microservice - Python Language Detection

Which Data?

I will be using the *sentences.csv* file from https://downloads.tatoeba.org/exports/ . This data consists of 9,714,915 pieces of text in 328 different unique languages. However, for my analysis I plan to simplify the dataset a little bit by using only English, Spanish, German, French and Italian. There are three columns in the dataset: Language, Target Language, and Text.

Research Questions? Benefits? Why analyze these data?

Many companies collect text data and need it to be analyzed. Being able to identify what language the text is in is essential, especially for international companies, before being able to analyze the data any further. Language detection is an essential first step in any NLP problem which is why I am choosing to analyze this dataset.

What Method?

I will be loading the dataset and filtering it so that I have only the rows with the languages listed above. This should make the dataset more usable on my personal computer, otherwise it'd be too large. I will then need to incorporate some feature engineering to get the data into a format that will work with a neural network. This might include using bag of words, n-grams, scaling and/or vectorizing depending on the research I do later on. I will also need to split the data into training and testing sets so that I can run and train a model. I plan to use the keras package to create neural network as my model. I am currently learning about neural networks in DSC 650 and hope to implement what I'm learning for this project. I plan to create a confusion matrix to evaluate the model.

Potential Issues?

It has been a while since I have worked with text data. I will definitely need to go back and find old examples from previous courses as well as spend time online researching the topic, different packages, and tools that are available to me to use. I personally think analyzing text data is extremely useful in today's world - because it is full on online posts, articles, comments, product reviews, texts, etc. - and being able to work with text data will be a skill I'll need moving forward.

Concluding Remarks

The goal of this analysis is to create and train a model to detect the difference between the English, Spanish, German, French and Italian languages and then when given some new text, it will be able to correctly predict the language it is in. NLP projects need to identify which language their input text is in order to move forward and complete any remaining tasks. Businesses can take advantage of language detection models to break the language barrier between countries as well as connect with more people all around the world.