Predicting English Variations using Doc2Vec

Jainabou Barry Danfa | MSI 2020 | Data Science Concentration | SI 630 (Natural Language Processing) Final Project

Introduction

English is the official language of 83 countries and spoken in 55 other countries. Understanding how different communities engage with the English language has various applications such as:

- Identifying the origin of specific texts
- Cybersecurity
- English as a second language resource development
- Tailored English messages for specific populations
- Enhanced English translation to fit specific contexts
- Social Media tailored ads to specific populations
- Creating culturally sensitive textbooks
- And many more!

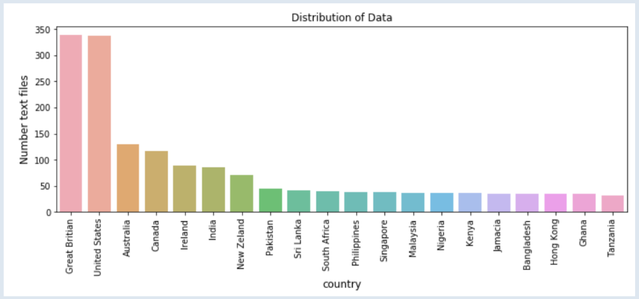
Doc2Vec is a unsupervised algorithm that generates vectors for any document. Doc2Vec was developed from the Word2Vec algorithm that represents words as a vector. We aim to use this algorithm to predict the English variation of various texts.

Project Update

This project is currently improving the accuracy and F1 score before deploying the finished prediction model. These are the current issues and proposed solutions:

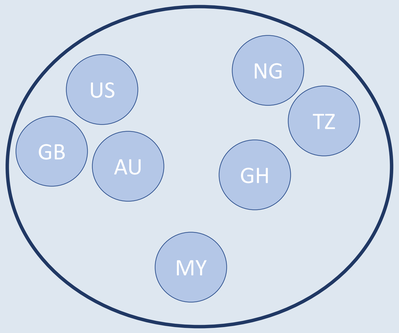
Issue	Proposed Solution
The input data has more text files from Great Britain and US compared to other countries. This is causing the model to train on more of those documents than the other countries.	By oversampling the minority countries in the dataset, we can have the model train on more of those samples by re-running them in the training dataset.
The model is currently predicting each document as originating from Great Britain. This issue may be from the large set of files from Great Britain or not enough differentiation in the documents	By gathering more training data or oversampling, we can have the model gather more differences from the documents based on country. Secondly, maybe Doc2Vec is not the best classification algorithm in determining significant differences among English types.

Data Cleaning and Processing



The data was obtained from the GloWbe (global Web-based English) corpus. The texts were mined from blog and general online texts. The countries represented in the dataset are highlighted in the chart above. Each document was tokenized by each word and the data was split 80/20 for training and testing.

Train Doc2Vec Model



Doc2Vec was implemented using genism and python. A distributed memory training algorithm was used to preserve the word order of each document in each class of documents.

Test Doc2Vec Model

Test	Score
Accuracy	7.66%
F1	1.09%

With the 20% of data held out for testing, we compared the model prediction of text country origin to the actual country origin. The F1 score accounts for false positives and negatives in the sample. If the scores are not acceptable, we will further train the model with different techniques to improve the scores.

Use Doc2Vec Model

Once the model has acceptable scores, we can predict the origin of texts by analyzing the document and determining which countries features are closest to the documents features.