Language Detection using stopwords

Aniket Singh Solanki-16BCP001

Karan Lohia-16BCP019

# INTRODUCTION

Finding a documents source language is an important step for numerous cross-language tools. Natural Language models are usually specific to a discrete language. If one is not sure of the receiving document’s language, it becomes very hard to provide anyone with a good experience.

This kind of program is particularly helpful in online chat boxes of companies which interact with the customer and give a wholesome customer care experience. The software can identify the language and then the further conversations would take place in the same language.

Language Identification is an important objective in the text mining process. Successful study of the extracted text with natural language processing or machine learning training demands a good language identification algorithm. Any natural language software be it sentiment analysis or toxic comment analysis cannot take place without identifying the language first.

**LITERATURE SURVEY**

Concerning, Language Detection, there is an existing Google API available which directly gives us the result. Along with this n-gram approach has also been widely used. We have used the stop words approach where we compare the stop words with the given data set. Every language consists certain key words which could be used to determine the language.

**DATASET DESCRIPTION**

This dataset is meant for use by researchers aiming to use machine learning techniques to build automatic language detection algorithms.

It Contains 140 000 sentences (14 000 per language class) for training and 50 000 sentences (5000 per language class) for testing.

Each Variant contains the following files

### **train\_set.txt**

Contains sentences for training and integer labels representing the language classes

### **test\_set.txt**

Contains sentences for testing and integer labels representing the language classes

##### Both the train and test files are organized as sentence - label pairs with the tab characer "\t" separating them.

### **chars.json**

A single json file containing two arrays: "char\_to\_idx" mapping characters to Integers and "idx\_to\_char" mapping Integers to characters

### **languagemap.json**

A json file mapping Integer labels to the languages they represent

### Included Languages

* English
* French
* Russian
* Chinese Mandarin
* Hebrew
* Portugese
* Polish
* Dutch
* Japanese
* Italian

**Methods and Ensemble**

In this section we study baseline methods for the above mentioned common challenges. Further, we propose our ensemble learning architecture. Its goal is to minimize errors by detecting optimal methods for a given comment.

We have used stop words contained in the nlkt library to analyze the dataset. We match the stopwords with the words occurring in the dataset and find out which language’s stopwords have been matched the most. The language whose matching of stopwords to the words in the dataset is the highest, that language is correctly classified.

POTENTIAL BUG- Languages whose vocabulary is the same cannot be classified using the stop words approach.

**RESULTS**

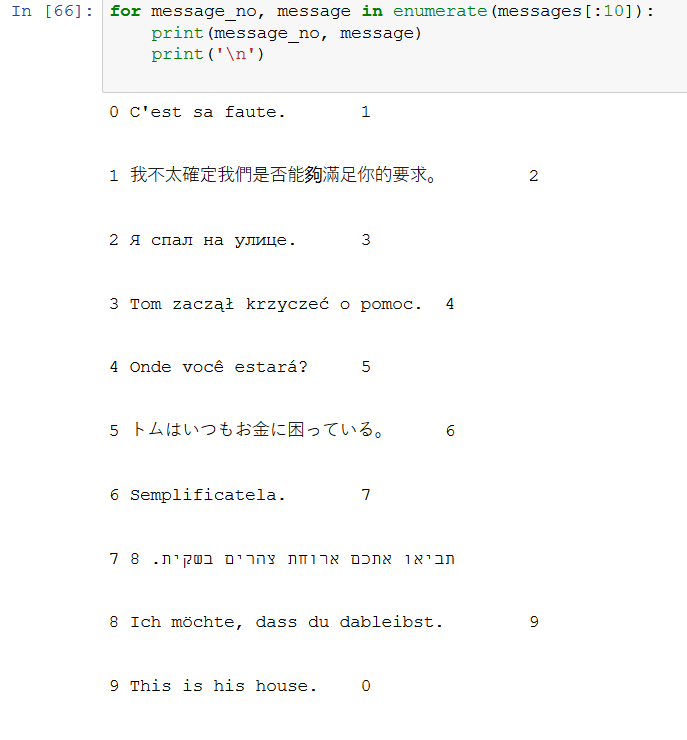
Stop words for the following languages are available in the nlkt library in python.

****

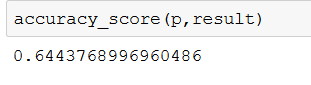
In the dataset we add two column names to the existing data. This is message and label. This would help us to classify the messages according to their labels.

****

Now the given dataset is classified according to the stop words in the nlkt library as mentioned above before.



The accuracy comes out to be 64.437%. The reason for this low accuracy is that German and Dutch languages have the same vocabulary, hence they cannot be classified using the available stop words.



**CONCLUSION**

In this work we presented analysis using Stop words for language detection. We find that a large source of error is due to some European languages having the same vocabulary while been differentiated on the terms of grammar and pronunciation. To successfully classify these languages we need to understand the underlying grammar that differentiates them.

**REFERENCES**

* Dataset acquired from johnlafenwa ling10 dataset on github.