Language Detection Model

Alok Bhawankar PD09
Vivek Ray PD17
Anmol Singh PD26

Language Detection

1. Language detection is an important aspect of many Business Process such as Query Resolution, Question Answering System and so on.
2. Figuring out a document’s source language is an essential first step for many cross-language tools.
3. Our trained language models in an ensemble to determine which language, or languages, an unknown input document is written in.

Libraries used:

- Numpy
- Pandas
- Seaborn
- Matplotlib
- Pickle
- Sklearn

Text Vectorization:

Text Vectorization is the process of converting text into numerical representation. Here is some popular methods to accomplish text vectorization:

- Bag-of-Words : Converts text into set of vectors containing the count of word occurences.
- TF-IDF: Creates vectors from text which contains information on the more important words and the less important ones as well.
- Word2Vec: Creates vectors that are numerical representations of word features.

We'll be using TF-IDF vectorization in our model.

Data Preprocessing

-> For English, French and German:

- Remove Punctuations in dataset.
- Convert to lower case text
- Remove digits and numerics.

French	Text	language
1\tLe prÃ©sident de l'OM, Jean-Claude Dassier,...	\tle prã©sident de lom jeanclaude dassier y co...	French
2\tIl a signÃ© jeudi Ã l'issue du programme l...	\til a signã© jeudi ã lissue du programme lib...	French

Spanish	Text	language
1\tDenuncia IEM probable fraude con actas elec...	\tdenuncia iem probable fraude con actas elect...	Spanish
2\tA pesar de la organizaciÃ³n del movimiento,...	\ta pesar de la organizaciã³n del movimiento s...	Spanish

Train and Test Data:

- Splitting Data into Train and Test sets (80:20)

Vectorizer and Model fitting Pipeline:

- For N-Grams: ngram_range(1,3)
- Analyzer = ‘char’
- Model: linear_model.LogisticRegression()