Language Detection Web Application Using NLP

Contributors:
Md. Masum Billah
Sakib Alam Snigdha

Supervisor:
Enamul Hassan

Outline

- Why NLP
- Text Pre-processing
- Vectorization
- Processing Datasets
- Training Model
- Model Evaluation
- Frontend Application
- Results & Findings
- End Note

Why NLP?

- Language is one of the defining characteristics of our species
- NLP helps to resolve ambiguity in language and adds useful numeric structure to the data
- A large corpus of knowledge can be organized and easily accessed using NLP

Text Preprocessing

- Tokenization
- Stop words Removal
- Lower case conversion
- Removing numeric/digits
- Removing Punctuations/Special Characters
- Removing characters (for foreign languages)
- Normalization

Text Preprocessing

- Remove punctuations
  ! " # $ % & ' ( ) * + , - . / : ; < = > ? @ [\ ] ^ _ ` { | } ~
- Remove line breaks, empty spaces
  Hello ↲World -> Hello World
- Transform to lowercase
  Hello World -> hello world

Vectorization

TF-IDF

TF-IDF creates vectors from text which contains information on the more important words and the less important words as well.

Algorithm:

$$tf\ idf\ (t, d, D) = tf\ (t, d)\ .\ Idf(t,D)$$

Where,

tf = term frequency

idf = inverse document frequency

t = word

d = document

D = set of documents

Processing Datasets

English Data

Total number of data: 1,80,657

	English
0	Musharraf's Last Act?
1	Desperate to hold onto power, Pervez Musharraf...
2	His goal?
3	To stifle the independent judiciary and free m...
4	Artfully, though shamelessly, he has tried to ...
5	Nothing could be further from the truth.
6	If Pakistan's history is any indicator, his de...
7	General Musharraf appeared on the national sce...
8	Many Pakistanis, disillusioned with Pakistan's...
9	The September 11, 2001, terrorist attacks on A...

Processing Datasets

Spanish Data

Total number of data: 1,59,260

	Spanish
0	Â¿El Ãºltimo acto de Musharraf?
1	En su desesperaciÃ³n por mantenerse en el pode...
2	Â¿Su meta?
3	Asfixiar la independencia del poder judicial y...
4	Con habilidad, pero de manera descarada, ha tr...
5	Nada podrÃ a estar mÃ¡s alejado de la verdad.
6	A juzgar por la historia de PakistÃ¡n, su deci...
7	El general Musharraf apareciÃ³ en el escenario...
8	Muchos pakistanÃ es, decepcionados de la clase...
9	Los ataques terroristas del 11 de septiembre d...

Processing Datasets

French Data

Total number of data: 1,47,251

	French
0	Le dernier numÃ©ro de Moucharraf ?
1	DÃ©sespÃ©rÃ© de conserver le pouvoir, Pervez M...
2	Dans quel but ?
3	Pour Ã©touffer un systÃ¨me judiciaire indÃ©pen...
4	Il a tentÃ© de faire passer cette action -- in...
5	On ne saurait Ãªtre plus loin de la vÃ©ritÃ©.
6	Si l'on se fie Ã l'histoire du Pakistan, la d...
7	Le GÃ©nÃ©ral Moucharraf est apparu sur la scÃ¨...
8	Bon nombre de Pakistanais -- qui avaient perdu...
9	Le 11 septembre 2001, les attaques terroristes...

Processing Datasets

German Data

Total number of data: 1,61,805

	German
0	Musharrafs letzter Akt?
1	In dem verzweifelten Versuch, an der Macht fes...
2	Sein Ziel?
3	Die unabhÃ¤ngige Justiz und die freien Medien ...
4	Listig, aber schamlos hat er versucht, seine M...
5	Dies ist absolut unwahr.
6	Falls Pakistans Geschichte ein Indikator ist, ...
7	General Musharraf betrat am 12. Oktober 1999 d...
8	Viele von der politischen Klasse ihres Landes ...
9	Am 11. September 2001 brachten die Terroransch...

Training Model

Splitting Data into Train and Test sets

80:20 (Train : Test)

```
X Train : (518481,)
X Test : (129621,)
y Train : (518481,)
y Test : (129621,)
```

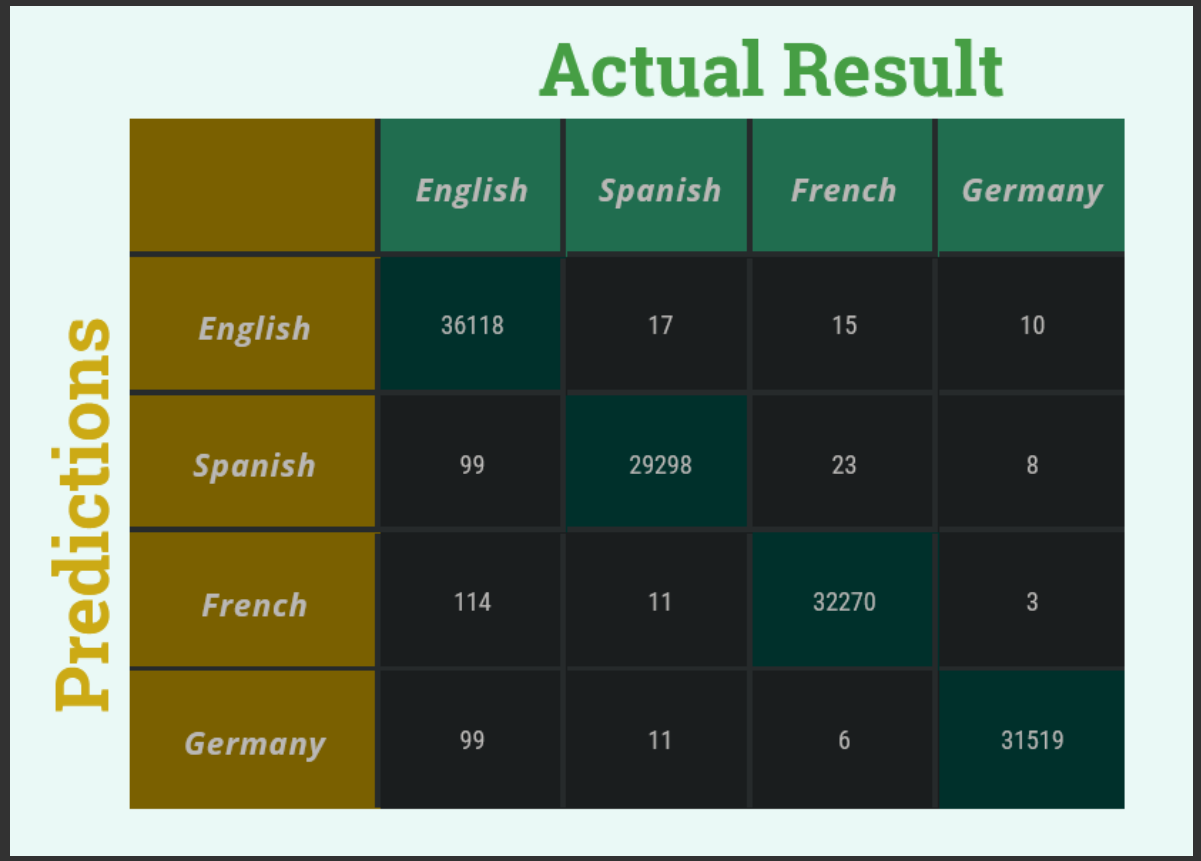
Model Evaluation

Accuracy & Confusion Matrix

```
Accuracy:  99.67906434914096 %
Confusion Matrix:
[[36118    17    15    10]
 [   99 29298    23     8]
 [  114    11 32270     3]
 [   99    11     6 31519]]
```

Model Evaluation

Confusion Matrix



Model Evaluation

Accuracy & Confusion Matrix with Separate Dataset

On Trained Model

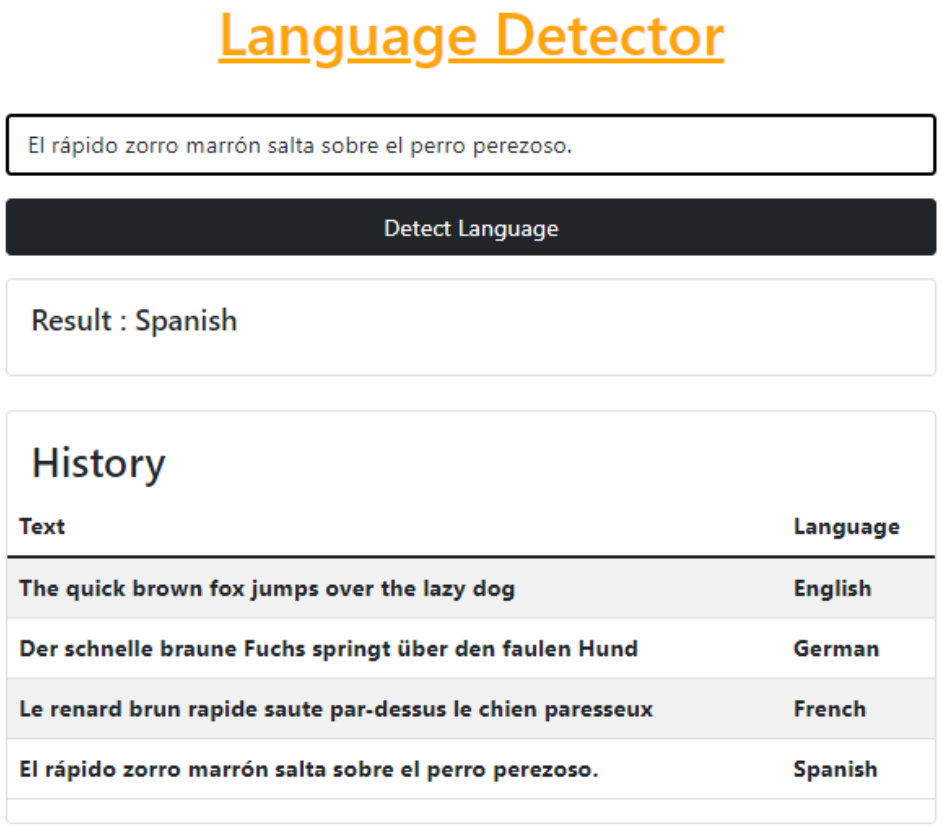
```
Accuracy:  98.47731510254818 %
Confusion Matrix:
 [[1380    4    0    1]
 [  16  988    7    3]
 [   0    0    0    0]
 [  11    5    2  801]]
```

Model Evaluation

Prediction

- It is raining today.
  English
- Il pleut aujourd'hui.
  French
- Es regnet heute.
  German
- Hoy esta lloviendo.
  Spanish

Frontend Application



Result & Findings

- 99.67% accuracy proved that the model is successful on determining language of English, German, French and Spanish.
- As per the accuracy on the testing dataset the model is 1.2% overfitting.
- The web application is very responsive.
- Further development of this project will enhance the performance and adding more languages in the future.

Thank you