greatlearning
Learning for Life

How to Build Language Detection Model using NLP

Sravan Malla

Outline

- Basics of NLP
- Text Pre-processing & Vectorization
- Steps to build Language detection model using NLP
- Questions & Answers

What is NLP?

NLP (Natural Language Processing) is analysis or generation of natural language text using computers.

For Example:

- Language Detection
- Machine Translation
- Next Word Prediction
- Automated query answering
- Speech Parsing

What is NLP Based on?

NLP is primarily based on

- Probability and Statistics
- Machine Learning/Deep Learning
- Linguistics
- Common sense

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

Types of use-cases in NLP

- Text Classification
- Named Entity Recognition
- Text Parsing
- Text Synthesis
- Reasoning

Text Pre-Processing

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

Vectorization

- Bag-of-Words (Count Vectorizer)

  Bag of Words converts text into set of vectors containing the count of word occurrences in the document.

- TF-IDF

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

- Word2Vec

  Word2vec creates vectors that are numerical representations of word features, features such as the context of individual words. The purpose and usefulness of Word2vec is to group the vectors of similar words together in vector space. That is, it detects similarities mathematically.

Build Language Detection Model

Importing Libraries:

```
import string
import re
import codecs
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn import feature_extraction
from sklearn import linear_model
from sklearn import pipeline
from sklearn.model_selection import train_test_split
from sklearn import metrics
```

Build Language Detection Model

Loading English Raw Data:

```
eng_df = pd.read_csv("train\\english.txt","utf-8",header=None, names=["English"])
eng_df.head()
```

	English
0	2\t0.00 0.00% How the mighty have fallen.
1	3\t0.00 0.00% Major companies coming out with ...
2	4\t01 JANUARY 2015, MAGAZINE Why are there so ...
3	5\t0:44 Now watching Up next 2016 Word Associa...
4	6\t0:46 autoplay autoplay Copy this code to yo...

Build Language Detection Model

Loading German Raw Data:

```
ger_df = pd.read_csv("train\\german.txt","utf-8",header=None, names=["German"])
ger_df.head()
```

	German
0	1\t04.01.15 Wissenschaft Welche Rolle das Lich...
1	2\t04.04.2014 â€“ 09:54Touristik "Das Magazin ...
2	3\t04. November 2015 18:29 Russland greift nur...
3	4\t05.06.2015 â€“ 10:16 Fernsehen MÃ¼nchen (ot...
4	5\t05. Oktober 2015 11:28 Wenn KÃ¶rper & Geist...

Build Language Detection Model

Loading French Raw Data:

```
fre_df = pd.read_csv("train\\french.txt","utf-8",header=None, names=["French"])
fre_df.head()
```

	French
0	1\tLe prÃ©sident de l'OM, Jean-Claude Dassier,...
1	2\tIl a signÃ© jeudi Ã l'issue du programme l...
2	3\tClub du 4e chapeau, la Chorale aura sans do...
3	4\tL'Espagnol Pau Gasol, crÃ©ditÃ© de 22 point...
4	5\tManuel Osborne-Paradis ne croit pas quÂ'il ...

Build Language Detection Model

Loading Spanish Raw Data:

```
spa_df = pd.read_csv("train\\spanish.txt","utf-8",header=None, names=["Spanish"])
spa_df.head()
```

	Spanish
0	1\tDenuncia IEM probable fraude con actas elec...
1	2\tA pesar de la organizaciÃ³n del movimiento,...
2	3\tEs decir, el BM entrega prÃ©stamos (evident...
3	4\tSin embargo, el juego no tenÃa construida ...
4	5\tBuscados por las autoridades, trabajamos en...

Build Language Detection Model

Loading Chinese Raw Data:

```
chi_df = pd.read_csv("train\\chinese.txt",header=None, names=["Chinese"])
chi_df.head()
```

	Chinese
0	2 0.00 0.00%強者如何墮落。
1	3 0.00 0.00%主要公司推出他們最新的季度數據包括Dave＆Buster's，Men...
2	2015年1月4日，雜誌為什麼有這麼多Magna Cartas?
3	5 0:44現在觀看2016年下一屆Word協會與Brad Woodhouse共和黨總統選舉...
4	6 0:46 autoplay autoplay將此代碼複製到您的網站或博客搜索者在阿拉斯加...

Build Language Detection Model

Data Pre-Processing

```
for char in string.punctuation:
    print(char, end=" ")
translate_table = dict((ord(char), None) for char in string.punctuation)
```

```
! " # $ % & ' ( ) * + , - . / : ; < = > ? @ [ \ ] ^ _ ` { | } ~
```

```
for i,line in eng_df.iterrows():
    line = line['English']
    if len(line) != 0:
        line = line.lower()
        line = re.sub(r"\d+", "", line)
        line = line.translate(translate_table
        data_eng.append(line)
        lang_eng.append("English")
```

```
for i,line in ger_df.iterrows():
    line = line['German']
    if len(line) != 0:
        line = line.lower()
        line = re.sub(r"\d+", "", line)
        line = line.translate(translate_table)
        data_ger.append(line)
        lang_ger.append("German")
```

Build Language Detection Model

Data Pre-Processing

```
for char in string.punctuation:
    print(char, end=" ")
translate_table = dict((ord(char), None) for char in string.punctuation)
```

```
! " # $ % & ' ( ) * + , - . / : ; < = > ? @ [ \ ] ^ _ ` { | } ~
```

```
for i,line in fre_df.iterrows():
    line = line['French']
    if len(line) != 0:
        line = line.lower()
        line = re.sub(r"\d+", "", line)
        line = line.translate(translate_table)
        data_fre.append(line)
        lang_fre.append("French")
```

```
for i,line in spa_df.iterrows():
    line = line['Spanish']
    if len(line) != 0:
        line = line.lower()
        line = re.sub(r"\d+", "", line)
        line = line.translate(translate_table)
        data_spa.append(line)
        lang_spa.append("Spanish")
```

Build Language Detection Model

Data Pre-Processing

```
for char in string.punctuation:
    print(char, end=" ")
translate_table = dict((ord(char), None) for char in string.punctuation)
```

```
! " # $ % & ' ( ) * + , - . / : ; < = > ? @ [ \ ] ^ _ ` { | } ~
```

```
for i,line in chi_df.iterrows():
    line = line['Chinese']
    if len(line) != 0:
        line = line.lower()
        line = re.sub(r"\d+", "", line)
        line = re.sub(r"[a-zA-Z]+", "", line)
        line = line.translate(translate_table)
        data_chi.append(line)
        lang_chi.append("Chinese (Traditional)")
```

Build Language Detection Model

```
df = pd.DataFrame({"Text":data_eng+data_ger+data_fre+data_spa+data_chi,
                   "language":lang_eng+lang_ger+lang_fre+lang_spa+lang_chi})
print(df.shape)
```

```
(149787, 2)
```

Data Before and After Pre-Processing

English
2\t0.00 0.00% How the mighty have fallen.
3\t0.00 0.00% Major companies coming out with ...

German
1\t04.01.15 Wissenschaft Welche Rolle das Lich...
2\t04.04.2014 â€“ 09:54Touristik "Das Magazin ...

Text	language
\t how the mighty have fallen	English
\t major companies coming out with their late...	English

Text	language
\t wissenschaft welche rolle das licht im wiss...	German
\t â€“ touristik das magazin fã¼r die freiheit...	German

Build Language Detection Model

Data Before and After Pre-Processing

French
1\tLe prÃ©sident de l'OM, Jean-Claude Dassier,...
2\tIl a signÃ© jeudi Ã l'issue du programme l...

Text	language
\tle prã©sident de lom jeanclaude dassier y co...	French
\til a signã© jeudi ã lissue du programme lib...	French

Spanish
1\tDenuncia IEM probable fraude con actas elec...
2\tA pesar de la organizaciÃ³n del movimiento,...

Text	language
\tdenuncia iem probable fraude con actas elect...	Spanish
\ta pesar de la organizaciã³n del movimiento s...	Spanish

Chinese
2 0.00 0.00%強者如何墮落。
2015年1月4日，雜誌為什麼有這麼多Magna Cartas?

Text	language
%強者如何墮落。	Chinese (Traditional)
年月日，雜誌為什麼有這麼多？	Chinese (Traditional)

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited

Build Language Detection Model

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

```
X, y = df.iloc[:,0],df.iloc[:,1]
X_train, X_test, y_train, y_test = train_test_split(X,
                                                    y,
                                                    test_size=0.2,
                                                    random_state=0)

print(X_train.shape)
print(X_test.shape)
print(y_train.shape)
print(y_test.shape)
```

```
(119829,)
(29958,)
(119829,)
(29958,)
```

Build Language Detection Model

Vectorizer and Model fitting Pipeline

```
vectorizer = feature_extraction.text.TfidfVectorizer(ngram_range=(1,3),analyzer='char')

pipe_lr_r13 = pipeline.Pipeline([
    ('vectorizer', vectorizer),
    ('clf', linear_model.LogisticRegression())
])
```

Model Fitting

```
pipe_lr_r13.fit(X_train, y_train)
```

Build Language Detection Model

Model Prediction

```
y_predicted = pipe_lr_r13.predict(X_test)
```

Model Evaluation

```
acc = (metrics.accuracy_score(y_test, y_predicted))*100
print(acc,'%')
```

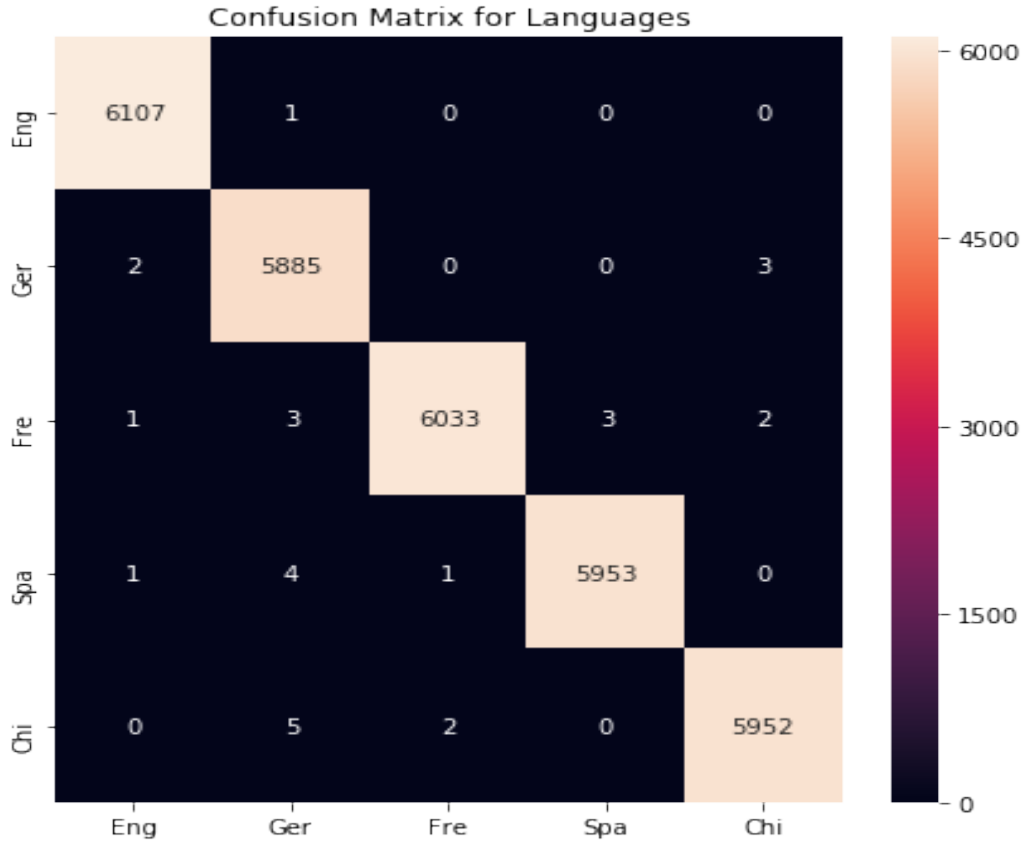
```
99.9065358168102 %
```

```
matrix = metrics.confusion_matrix(y_test,y_predicted)
print('Confusion matrix : \n',matrix)
```

```
Confusion matrix :
 [[6107    1    0    0    0]
 [   2 5885    0    0    3]
 [   1    3 6033    3    2]
 [   1    4    1 5953    0]
 [   0    5    2    0 5952]]
```

Build Language Detection Model

Confusion Matrix



Build Language Detection Model

Model Saving

```
import pickle
# Persist model so that it can be used by different consumers
lrFile = open('LRModel.pckl', 'wb')
pickle.dump(pipe_lr_r13, lrFile)
lrFile.close()
```

Model Loading

```
global lrLangDetectModel
lrLangDetectFile = open('LRModel.pckl', 'rb')
lrLangDetectModel = pickle.load(lrLangDetectFile)
lrLangDetectFile.close()
```

Build Language Detection Model

Method Definition to call Trained Model and Make Predictions

```
def lang_detect(text):
    import numpy as np
    import string
    import re
    import pickle
    translate_table = dict((ord(char), None) for char in string.punctuation)

    global lrLangDetectModel
    lrLangDetectFile = open('LRModel.pckl', 'rb')
    lrLangDetectModel = pickle.load(lrLangDetectFile)
    lrLangDetectFile.close()

    text = " ".join(text.split())
    text = text.lower()
    text = re.sub(r"\d+", "", text)
    text = text.translate(translate_table)
    pred = lrLangDetectModel.predict([text])
    prob = lrLangDetectModel.predict_proba([text])
    return pred[0]
```

Build Language Detection Model

Predictions

```
lang_detect("Hello I just built my own language detection model")
```

```
'English'
```

```
lang_detect("Hallo, ich habe gerade mein eigenes Spracherkennungsmodell erstellt")
```

```
'German'
```

```
lang_detect("Bonjour, je viens de créer mon propre modèle de détection de langue")
```

```
'French'
```

```
lang_detect("Hola, acabo de construir mi propio modelo de detección de idioma")
```

```
'Spanish'
```

```
lang_detect("您好，我剛剛建立了自己的語言檢測模型")
```

```
'Chinese (Traditional)'
```

Any Questions

greatlearning
Learning for Life

Thank You