# Language Detection

by

Vishal Singh 11722015
Arun Borale 11722001
hritik Madke 11722010

Guide:

Lect. Mamta Patil

Department of Computer Engineering

New Horizon Institute of Technology and Management

University of Mumbai

(2021-2022)

NEW HORIZON INSTITUTE OF TECHNOLOGY

AND MANAGEMENT

# CERTIFICATE

This is to certify that the Mini Project entitled "**Language Detection"** is a Bonafede work of **"Vishal Singh" (**11722015 **), "Arun Borale" (**11722001**) and "Hritik Madke" (**11722010 **)** submitted to the University of Mumbai in partial fulfillment of the requirement for the award of the degree of "Bachelor of Engineering" in "Computer Engineering".

(Lect. Mamta patil)
Guide

(Dr. Sanjay N. Sharma)
Head of Department

(Dr. Prashant D. Deshmukh)
Principal

NEW HORIZON INSTITUTE OF TECHNOLOGY
AND MANAGEMENT

# Project Report Approval for B.E.

This project report entitled “**Language Detection**” by ***Vishal Singh,*** **Arun Borale** ***and hritik madke*** is approved for the ***Mini Project in Computer Engineering***, ***2021-22***.

Examiner Name Signature

1. ____________________ ____________________

2. ____________________ ____________________

Date:

Place:

# Abstract

Language identification is the task of automatically detecting the language(s) present in a document based on the content of the document. In this work, we address the problem of detecting documents that contain text from more than one language (multilingual documents). We introduce a method that is able to detect that a document is multilingual, identify the languages present, and estimate their relative proportions. We demonstrate the effectiveness of our method over synthetic data, as well as real-world multilingual document collected from the web.

# Index

# Chapter 1

# Introduction

## 1.1 Objective

Take Output From user perform preprocess method and analyze perform machine learning algorithm to detect that given language is given in which language.

## 1.2 Scope

User can easily detect the disease he/she is facing, just by  Giving by different input to the system . thus this system can provide you with the answer that the given input is in which langauge.

# Chapter 2

# Present Investigation

## 2.1 Problem Definition

Language identification is the process of determining which language given content is in. The basic goal of the language identification (LI) system is to accurately identify the language

## 2.2 Feasibility Analysis

This system is completely design in python and dataset are imported from CSV files. Hence user has to install python and a prerequisite application to use the system. Even updating of dataset can be done easily as they are stored in CSV files.

# Chapter 3

# Implementation Details

## 3.1 Implementation plan

- We collected datasets from various resources.
- We have Two column Text and language.
- There are 2000 Text row  in our datasets.
- First we apply datapreprocessing on data like stemming,removing punctuation ,removing stopwords.
- After Preprocessing we split our Dataset And then we will apply Feature Engineering.
- We will fit Our data in machine learning model and train it .
- Finally after training we will use pipeline method to detect given input is in which language.

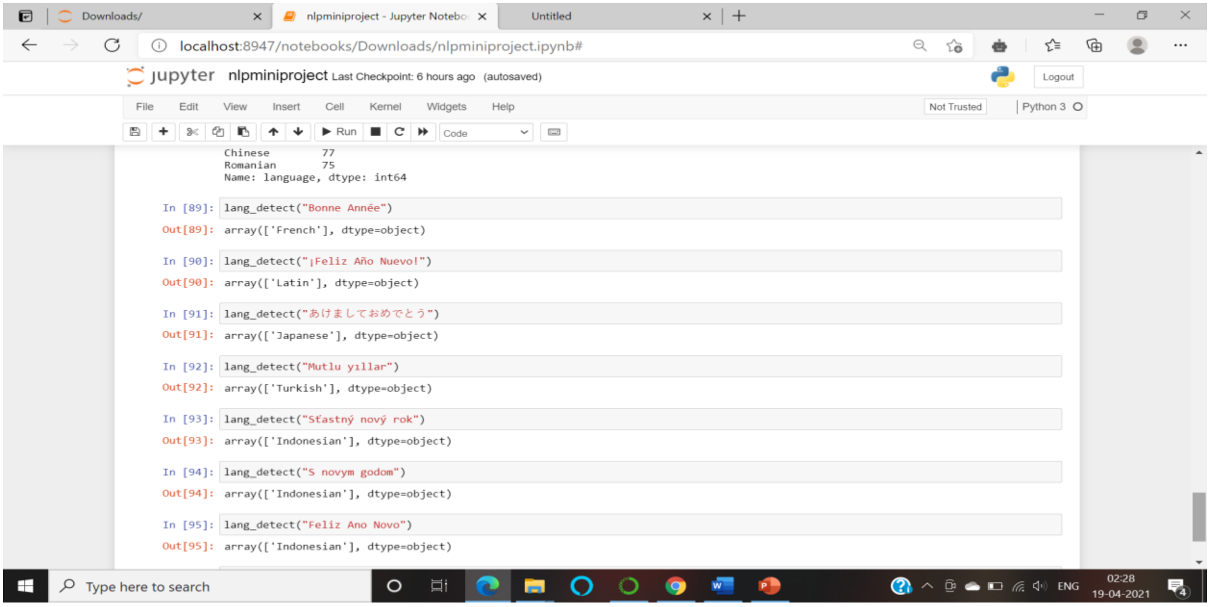
## 3.2 Code

```
#!/usr/bin/env python

# coding: utf-8

import string

import csv

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.pipeline import Pipeline
from sklearn.model_selection import train_test_split
from sklearn import metrics
import nltk
#in[2]:
 lang=pd.read_csv("D:\PYTHON\dataset.csv")
# In[3]:
lang=lang[0:2000]
# In[4]:
lang
# In[5]:
lang.isnull().sum()

# In[70]:
for char in string.punctuation:
    print(char,end=" ")
translate_table=dict((ord(char),None) for char in string.punctuation)
# In[71]:
data_lang=[]
for i,line in lang.iterrows():
    line=line['Text']
    if len(line)!=0:
        line=re.sub(r"\d+", "",line)
        line=re.sub(r"[a-zA-Z]+","",line)
        line=line.translate(translate_table)
        data_lang.append(line)

languag=[]
for i,line in lang.iterrows():
    line=line['language']
    if len(line)!=0:
        languag.append(line)

# In[72]:
df=pd.DataFrame({"Text":data_lang,"language":languag} )

# In[73]:
df



# In[74]:
df.shape
# In[75]:
X,y=df['Text'],df['language']
y
# In[76]:
X_train,X_test,y_train,y_test=train_test_split(X,y,test_size=0.2,random_state=0)
# In[77]:
print(X_train.shape)
print(X_test.shape)
print(y_train.shape)
print(y_test.shape)
# In[78]:
vectorizer=feature_extraction.text.TfidfVectorizer(ngram_range=(1,3),analyzer='char')
# In[79]:
pipeline_clf=Pipeline([('vectorizer',vectorizer),('clf',linear_model.LogisticRegression())])
# In[80]:
pipeline_clf.fit(X_train,y_train)
# In[81]:
y_predicted=pipeline_clf.predict(X_test)
# In[82]:
```

```
acc=(metrics.accuracy_score(y_test,y_predicted))


# In[83]:
print(acc)
# In[84]:
import pickle
lrfile=open('Langmodel.pckl','wb')
pickle.dump(pipeline_clf,lrfile)
lrfile.close()
# In[85]:
def lang_detect(text):
    import pickle
    import numpy as np
    import string
    import re
    translate_table=dict((ord(char),None) for char in string.punctuation)

    global LanguagedetectionModel
    Languagedetectionfile=open('Langmodel.pckl','rb')
    LanguagedetectionModel=pickle.load(Languagedetectionfile)
    Languagedetectionfile.close()

    text=" ".join(text.split())
    text=text.lower()
    text=re.sub(r"\d+", "",text)
    text=text.translate( translate_table)
    pred=LanguagedetectionModel.predict([text])
    return pred

# In[86]:
  df['language'].value_counts()
```

```
# In[87]:

df.drop(df.loc[df['language']=='indonesian'].index,inplace=True)

# In[88]:

df['language'].value_counts()

# In[89]:

lang_detect("Bonne Année")

# In[90]:

lang_detect("¡Feliz Año Nuevo!")

# In[91]:

lang_detect("あけましておめでとう")

# In[92]:

lang_detect("Mutlu yıllar")

# In[93]:

lang_detect("Sťastný nový rok")

# In[94]:

lang_detect("S novym godom")

# In[95]:

lang_detect("Feliz Ano Novo")
```

## 3.3 Screen Shot


Downloads/
nlpminiproject - Jupyter Notebo
Untitled
localhost:8947/notebooks/Downloads/nlpminiproject.ipynb#
jupyter nlpminiproject Last Checkpoint: 6 hours ago (autosaved)
Logout
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Not Trusted
Python 3
Run
Code
Type here to search
ENG
02:28
19-04-2021

```
Chinese        77
Romanian       75
Name: language, dtype: int64

In [89]: lang_detect("Bonne Année")
Out[89]: array(['French'], dtype=object)

In [90]: lang_detect("¡Feliz Año Nuevo!")
Out[90]: array(['Latin'], dtype=object)

In [91]: lang_detect("あけましておめでとう")
Out[91]: array(['Japanese'], dtype=object)

In [92]: lang_detect("Mutlu yıllar")
Out[92]: array(['Turkish'], dtype=object)

In [93]: lang_detect("Sťastný nový rok")
Out[93]: array(['Indonesian'], dtype=object)

In [94]: lang_detect("S novym godom")
Out[94]: array(['Indonesian'], dtype=object)

In [95]: lang_detect("Feliz Ano Novo")
Out[95]: array(['Indonesian'], dtype=object)
```

# Chapter 4

# Conclusion and Future Scope

## 4.1 Conclusion

Thus we have successfully Implemented designing part And developed a machine model that can successfully predict a language of a given input successfully.

## 4.2 Future Scope

- Can also be implemented in speech recognition
- Speech to speech translation.

# Acknowledgement

We would like to take this opportunity to thank one and all.

It is our immense pleasure to express our gratitude to our Guide, Lect. Mamta patil for providing us with constructive and positive feedback during the preparation of this project.

We would like to express our thanks to the Head of Computer Department, Dr. Sanjay N. Sharma and all other staff members for their encouragement and suggestions.

Last but not the least, we are thankful to our friends for their support and coordination. We are also thankful to our parents for their constant support and best wishes.

________________________
Vishal Singh

________________________
Arun borale

________________________
Hritik Madke

Date: 22rd April 2021