Detecting an Unknown Language

NLP(Natural Language Processing) Project

Table of contents

01

Introduction

Language Detection

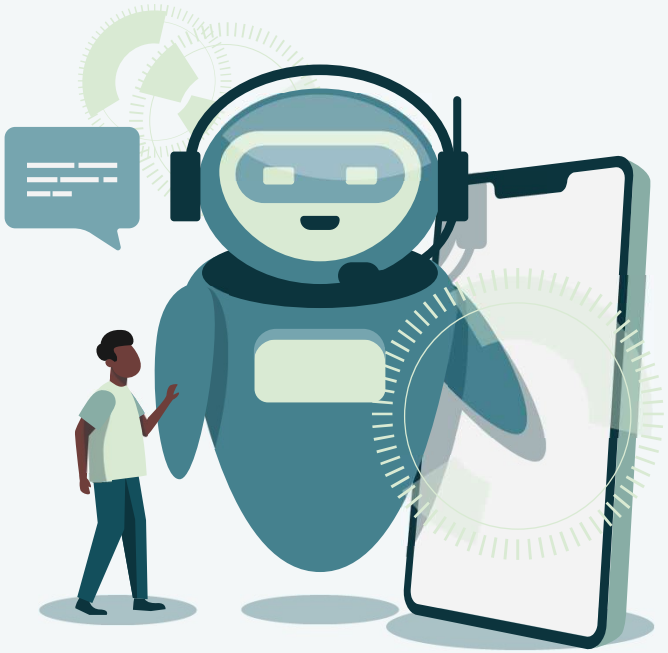
Introduction

Language detection, also known as language identification, is a fundamental task in Natural Language Processing (NLP) that involves determining the language of a given piece of text.

In today's globalized world, where users communicate in multiple languages across various platforms, language detection has become a crucial component of many NLP applications.

Language detection is the process of identifying the language of a given text. With Python, there are several powerful libraries that can automate this task. This presentation will introduce the concept of language detection and demonstrate how Python can be used to identify an unknown language in text.

—Let's begin...



02

Libraries Used

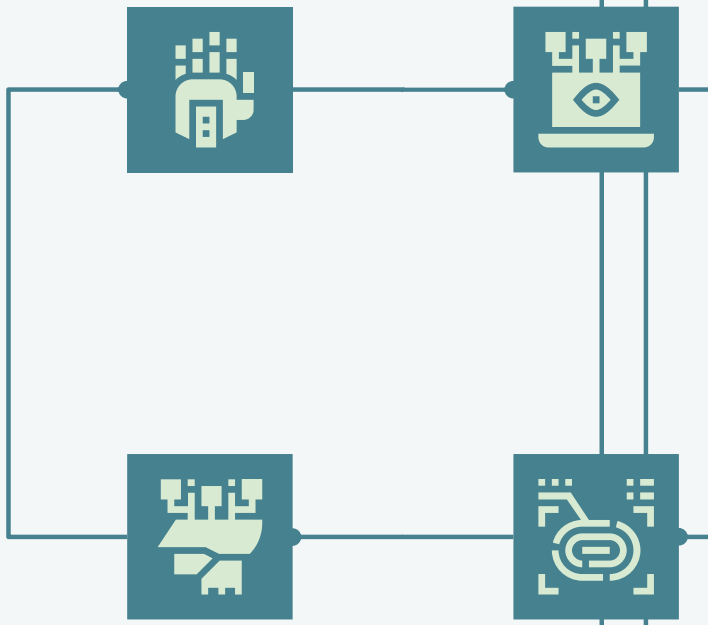
Libraries used for Language Detection

1. Langdetect – The **langdetect** library is popular Python library for detecting the language of a given text. It is based on the Google's Languge-detection library and uses statistical models to identify languages.

2. Textblob - **TextBlob** is a popular Python library for processing textual data. It provides simple and intuitive APIs for performing various Natural Language Processing (NLP) tasks, such as part-of-speech tagging, noun phrase extraction, sentiment analysis, and language translation.

3. googletrans==4.0.0-rcl – **googletrans** is a Python library that provides a simple interface to interact with Google Translate's API, allowing to translate text between various languages.

4. Langid - The library automatically detects the language of the input text by analyzing the frequency of characters, n-grams, and other linguistic patterns.



03

Working of Project

Code written for detecting the Language

Input:

```
# Python program to demonstrate langdetect


from langdetect import detect


# Specifying the language for detection

print(detect("AI refers to the systems that can perform tasks automatically."))
print(detect("Искусственный интеллект относится к системам, которые могут выполнять задачи автоматически."))
print(detect("La IA se refiere a los sistemas que pueden realizar tareas automáticamente."))
print(detect("AI refererer til systemene som kan utføre oppgaver automatisk."))
print(detect("एआई उन सिस्टमों को संदर्भित करता है जो स्वचालित रूप से कार्य कर सकते हैं।"))
print(detect("AIは自動的にタスクを実行できるシステムを指します。"))
```

Output:

```
en
ru
es
no
hi
ja
```

These are ISO 639-1 language codes representing different languages. Here's their meaning:

- **en** : English
- **ru** : Russian
- **es** : Spanish
- **no** : Norwegian
- **hi** : Hindi
- **ja** : Japanese
- **zh-CN** : Simplified Chinese (as used in mainland China)

These codes are widely used in localization, software, and linguistics to identify languages.

Code written for translating the Language

Input:

```
from textblob import TextBlob
from googletrans import Translator

L = ["AI refers to the systems that can perform tasks automatically.",
  "Искусственный интеллект относится к системам, которые могут выполнять задачи автоматически.",
  "La IA se refiere a los sistemas que pueden realizar tareas automáticamente.",
  "AI refererer til systemene som kan utføre oppgaver automatisk.",
  "एआई उन सिस्टमों को संदर्भित करता है जो स्वचालित रूप से कार्य कर सकते हैं।",
  "AIは自動的にタスクを実行できるシステムを指します。",
  ]

translator = Translator() # Create a Translator object

for i in L:
  # Language Detection using Google Translate API
  detected = translator.detect(i)
  print(detected.lang)
```

Output:

```
en
ru
es
no
hi
ja
```

These are ISO 639-1 language codes representing different languages. Here's their meaning:

- **en** : English
- **ru** : Russian
- **es** : Spanish
- **no** : Norwegian
- **hi** : Hindi
- **ja** : Japanese
- **zh-CN** : Simplified Chinese (as used in mainland China)

These codes are widely used in localization, software, and linguistics to identify languages.

Code written for detecting the Language (Taking Input from the user)

Input:

```
from textblob import TextBlob
from googletrans import Translator

word = input("Enter any word or sentence of any Language: ")
print(word)

translator = Translator() # Create a Translator object

# Language Detection using Google Translate API
detected = translator.detect(word)
print(detected.lang)
```

Note:- Here if we type any Language word or sentence, it will identify it also.
Langdetect and TextBlob support 50 languages.

Output:

```
Enter any word or sentence of any Language: 自動
自動
zh-TW
```

Code written for detecting the Language (Taking Input from the user)

Input:

```
from textblob import TextBlob
from googletrans import Translator

word = input("Enter any word or sentence of any Language: ")
print(word)

translator = Translator() # Create a Translator object

# Language Detection using Google Translate API
detected = translator.detect(word)
print(detected.lang)
```

Output:

```
Enter any word or sentence of any Language: abcd
abcd
en
```

Code written for detecting the Language with confidence

Input:

```
# Python program to demonstrate
# langid


import langid


L = ["AI refers to the systems that can perform tasks automatically.",
  "Искусственный интеллект относится к системам, которые могут выполнять задачи автоматически.",
  "La IA se refiere a los sistemas que pueden realizar tareas automáticamente.",
  "AI refererer til systemene som kan utføre oppgaver automatisk.",
  "एआई उन सिस्टमों को संदर्भित करता है जो स्वचालित रूप से कार्य कर सकते हैं।",
  "AIは自動的にタスクを実行できるシステムを指します。",
  ]

for i in L:

    # Language detection
    print(langid.classify(i))
```

Output:

```
('en', -149.66131234169006)
('ru', -1781.6450893878937)
('es', -228.5087342262268)
('nb', -103.85199356079102)
('hi', -426.2300704717636)
('ja', -635.0262796878815)
```

Scores (Log-Probabilities):

The numbers (e.g., -119.93 for en, -641.34 for ru) are log-probabilities. They indicate how likely the input text is to be in the corresponding language.

Higher (less negative) scores mean higher confidence.

The language with the highest score (closest to 0) is the most likely match.

Example Interpretation:

For the input you tested:

en: -119.93

ru: -641.34

es: -191.01

zh: -199.18

hi: -286.99

ja: -875.66

The Langid tool would classify the input as English (en) because it has the highest score (-119.93). Lower scores (e.g., -875.66 for ja) indicate much lower confidence for those languages.

Examples

Detecting language of user input in web applications.

Sorting content based on language.

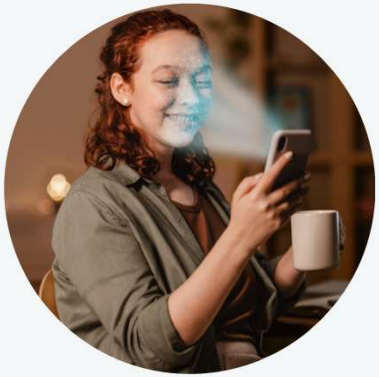
Analyzing multilingual data sets.

Detecting fake or spam content based on language patterns.

Applications

Localization

Machine translation and localization.



Communication

Real-time language detection in chatbots and voice assistants.

Social

Sentiment analysis for multi-lingual social media data.

Thanks!

Do you have any questions?

By –
Prachi Saroj (215/UAI/001)
Bhumika Dutta (215/UAI/035)
Anil Bhaskar (215/UAI/027)
Manish Rajput(215/UAI/028)