NLP PROGRAM-3

A program to tokenize Non-English Languages

Soundarya G_ 2048057

Different Indian Language Texts used for the Program

- Tamil_text = '''வணக்கம் எப்படி இருக்கிறாய்'''
- Hindi_text = '''नमस्ते कैसी हो तुम'''
- Punjabi_text = '''ਹੈਲੋ ਤੁਸੀ ਕਿਵੇਂ ਹੋ'''
- Telugu_text = '''సంస్కృతానికి'''
- Gujarati_text = '''હેલો કેમ છો?'''
- Kannada_text = '''ನಮಸ್ಕಾರ ಹೇಗಿದ್ದೀರಾ'''
- Malayalam_text = '''ഹലോ, നിങ്ങൾക്ക് സുഖമാണോ'''
- Nepali_text = '''नमस्ते तपाइँ कसरी हुनुहुन्छ?'''
- Odia_text = '''ହେ ତମେ କେମିତି ଅଛ'''
- Marathi_text = '''नमस्कार तुम्ही कसे आहात'''
- Bengali_text = '''হ্যালো, আপনি কেমন আছেন'''
- Urdu_text = '''ہیلو آپ کیسے ہیں'''

Import necessary libraries

```
# pip install inltk
# pip install langdetect
# pip install googletrans
```

```
from inltk.inltk import setup
from inltk.inltk import tokenize
from langdetect import detect
import pandas as pd
```

Different Indian Languages for tokenization

```
indian_languages={'hi': 'Hindi','pa': 'Punjabi','te': 'Telugu','gu': 'Gujarati',
                  'kn': 'Kannada','ml': 'Malayalam','ne': 'Nepali','or': 'Odia',
                  'mr': 'Marathi','bn': 'Bengali','ta': 'Tamil','ur': 'Urdu'}

df = pd.DataFrame(list(indian_languages.items()),columns = ['Code','Language'])
df
```

	Code	Language
0	hi	Hindi
1	pa	Punjabi
2	te	Telugu
3	gu	Gujarati
4	kn	Kannada
5	ml	Malayalam
6	ne	Nepali
7	or	Odia
8	mr	Marathi
9	bn	Bengali
10	ta	Tamil
11	ur	Urdu

User Text Function

This function takes input from user for the tokenization.

```
def user_text_func():
  text = input('\t')
  return text
```

```
user_text = user_text_func()
```

नमस्ते कैसी हो तुम

Language Detector

Detects which language is given as text.

```
print(detect(user_text), '-',indian_languages[detect(user_text)])
```

```
hi - Hindi
```

Setup for the user text's language

```
def setup_for_lang(text):
  try:
    setup(detect(text))
  except RuntimeError:
    #print('language setup is successful')
    return detect(text)
  else:
    print('error')
```

```
Lang = setup_for_lang(user_text)
Lang
```

```
'hi'  Done!
```

Language Translator

```
# User input translated to English
from googletrans import Translator
translator = Translator()
translator.translate('வணக்கம்',src='ta',dest='en')
```

Tokenization

```
# tokenize(input text, language code)
tokenize(user_text,Lang)
```

```
['_नमस्ते', '_कैसी', '_हो', '_तुम']
```

```
Lang_token_list = tokenize(user_text,Lang)
check = '_'
token_list = [idx for idx in Lang_token_list if idx[0]==check]
token_list
```

```
['_नमस्ते', '_कैसी', '_हो', '_तुम']
```

```
new_token_list = [s.replace('_','') for s in token_list]
new_token_list
```

```
['नमस्ते', 'कैसी', 'हो', 'तुम']
```

Main Function

```
def Non_English_Language_Translator():
  indian_languages={'hi': 'Hindi','pa': 'Punjabi','te': 'Telugu','gu': 'Gujarati',
                 'kn': 'Kannada','ml': 'Malayalam','ne': 'Nepali','or': 'Odia',
                 'mr': 'Marathi','bn': 'Bengali','ta': 'Tamil','ur': 'Urdu'}

  df = pd.DataFrame(list(indian_languages.items()),columns = ['Language','Code'])
  print(df)
  option = 'yes'
  while(option == 'yes'):
    print('\nType your text to be tokenized')
    user_text = user_text_func()
    Lang = setup_for_lang(user_text)
    Lang_token_list = tokenize(user_text,Lang)
    check = '_'
    token_list = [idx for idx in Lang_token_list if idx[0]==check]
    new_token_list = [s.replace('_','') for s in token_list]
    print('\n\tUser Input \t\t :', user_text)
    print('\tLanguage Identified \t :', indian_languages[Lang])
    print('\tLanguage Code \t\t :', Lang)
    print('\tTokens \t\t\t :', new_token_list)
    print('\tNo. of Tokens \t\t :', len(new_token_list), 'Tokens')
    print('\tEnglish Translation \t : On progress!!!')

    print('\n\t\tDo you want to continue the program')
    option = input('\t\t\t[yes/no]:')
```

```
Non_English_Language_Translator()
```

```
   Language       Code
0        hi      Hindi
1        pa    Punjabi
2        te     Telugu
3        gu   Gujarati
4        kn    Kannada
5        ml  Malayalam
6        ne     Nepali
7        or       Odia
8        mr    Marathi
9        bn    Bengali
10       ta      Tamil
11       ur       Urdu
```

```
Type your text to be tokenized
        வணக்கம் எப்படி இருக்கிறாய்

        User Input              : வணக்கம் எப்படி இருக்கிறாய்
        Language Identified     : Tamil
        Language Code           : ta
        Tokens                  : ['வணக்க', 'எப்படி', 'இருக்கிற']
        No. of Tokens           : 3 Tokens
        English Translation     : On progress!!!

                Do you want to continue the program
                        [yes/no]:yes

Type your text to be tokenized
        नमस्ते कैसी हो तुम

        User Input              : नमस्ते कैसी हो तुम
        Language Identified     : Hindi
        Language Code           : hi
        Tokens                  : ['नमस्ते', 'कैसी', 'हो', 'तुम']
        No. of Tokens           : 4 Tokens
        English Translation     : On progress!!!

                Do you want to continue the program
                        [yes/no]:yes

Type your text to be tokenized
        ನಮಸ್ಕಾರ ಹೇಗಿದ್ದೀರಾ

        User Input              : ನಮಸ್ಕಾರ ಹೇಗಿದ್ದೀರಾ
        Language Identified     : Kannada
        Language Code           : kn
        Tokens                  : ['ನಮಸ್ಕಾರ', 'ಹೇಗ']
        No. of Tokens           : 2 Tokens
        English Translation     : On progress!!!

                Do you want to continue the program
                        [yes/no]:no
```