Sampling & Data Cleaning.

- Prepared by: Sagun Shakya (https://github.com/sagsshakya)
- MSc. Data Science
- GITAM Institute of Science, Visakhapatnam.
- Email: sags.shakya@gmail.com (mailto:sags.shakya@gmail.com)

- We will take randomly 20,000 samples from both the previously pickled dataset
- Cleaning Process includes:
  - Removing punctuations (in both scripts - Nepali and English).
  - Removing numbers (in both scripts - Nepali and English).
  - Removing unprocessed UNICODE characters.

Importing necessary libraries.

In [1]:
```
import os
import pickle
import pandas as pd
import numpy as np
import re
```

In [2]:
```
os.chdir(r'C:\Users\acer\Desktop\PythonProgramming\Nepali_Hindi Language Classifi
```

Loading up the files.

In [3]:
```
nepali_raw = pickle.load(open('Nepali_Language.pkl', 'rb'))
hindi_raw = pickle.load(open('hindi_Language.pkl', 'rb'))
```

In [4]:
```
print('Total number of rows in Nepali: ',len(nepali_raw))
print('Total number of rows in Hindi: ',len(hindi_raw))
```

```
Total number of rows in Nepali:  20871
Total number of rows in Hindi:  118957
```

Converting the series into dataframe.

In [5]:
```
nepali_raw = nepali_raw.to_frame().reset_index(drop = True)
hindi_raw = hindi_raw.to_frame().reset_index(drop = True)
```

Taking 20k random samples from both datasets.

In [6]:
```
nepali_sample = nepali_raw.sample(n = 20000, random_state=100).reset_index(drop =
hindi_sample = hindi_raw.sample(n = 20000, random_state=100).reset_index(drop = T
```

In [7]:
```
nepali_sample.head(10)
```

	0
0	१२ औं राष्ट्रिय जनगणना २०७८ को तयारी सरकारले थ...
1	सत्ता घटक कांग्रेस र माओवादी केन्द्रले जिल्लाम...
2	नयाँ शैक्षिक सत्र सुरु हुन महिना दिन बाँकी छ ।...
3	हाम्रो देशमा नदीनाला, झरना, ताल, पोखरी गरी पान...
4	अन्य प्रदेशको दाँजोमा सबैभन्दा ढिलो कोरोना संक...
5	हाम्रो देशमा नदीनाला, झरना, ताल, पोखरी गरी पान...
6	प्रतिबन्धित चितुवाको छालासहित प्रहरीले तीन युव...
7	हाम्रा जनप्रतिनिधि र सरकारी संयन्त्र कतिसम्म अ...
8	केही दिनअघि दक्षिण भारतीय राज्य केरलामा जनावरम...
9	ललिता निवास प्रकरणमा पूर्वप्रधानमन्त्रीहरु माध...

Text cleaning.

In [8]:
```
import string
```

The fastest way to clean the text is to use the translation table (https://www.geeksfc
maketrans-translate-functions/).

Cleaning Hindi texts.

In [9]:
```
cleaner = lambda y: y[0].translate(str.maketrans('', '', string.punctuation + str
```

In [10]:
```
hindi_sample = hindi_sample.apply(cleaner, axis = 1)
hindi_sample = hindi_sample.to_frame()
```

In [11]:
```
hindi_sample.tail(7)
```

	0
19993	शहीद नरेंद्र सिंह के घर पहुंचे बिट्टा
19994	नगर निगम चुनाव सभी नामांकन पत्र सही
19995	फिटनेस में पिछड़ जाते हैं भारतीय खिलाड़ी धोनी
19996	पारा डिग्री लुढ़का मौसम हुआ सुहावना गर्मी से ...
19997	परिवहन मंत्री ने डग्गेमार बसों को पकड़ा
19998	संसद घेराव को लेकर की कार्यकर्ता बैठक
19999	पालिका प्रशासन ने निकाली जागरूकता रैली

Cleaning Nepali texts.

In [12]:
```
def cleaner_nepali(y):
    y = y[0]
    y = ' '.join(y.split())
    nepali_digits = ''.join([chr(2406+ii) for ii in range(10)])   # '०१२३४५६७८९'
    y = y.translate(str.maketrans('', '', string.punctuation + nepali_digits +
                                  '।‘’.—' + '!"#$%&\'()*+,-./:;<=>?@[\\]^_`{|}~')
    return y
```

In [13]:
```
nepali_sample = nepali_sample.apply(cleaner_nepali, axis = 1)
nepali_sample = nepali_sample.to_frame()
```

In [14]:
```
nepali_sample.head(7)
```

	0
0	औं राष्ट्रिय जनगणना को तयारी सरकारले थालेको ...
1	सत्ता घटक कांग्रेस र माओवादी केन्द्रले जिल्लाम...
2	नयाँ शैक्षिक सत्र सुरु हुन महिना दिन बाँकी छ ...
3	हाम्रो देशमा नदीनाला झरना ताल पोखरी गरी पानीका...
4	अन्य प्रदेशको दाँजोमा सबैभन्दा ढिलो कोरोना संक...
5	हाम्रो देशमा नदीनाला झरना ताल पोखरी गरी पानीका...
6	प्रतिबन्धित चितुवाको छालासहित प्रहरीले तीन युव...

Pickling the DataFrame.

In [17]:
```
nepali_sample.to_pickle('nepali_sample_cleaned.pkl')
```

In [18]:
```
hindi_sample.to_pickle('hindi_sample_cleaned.pkl')
```

The End.