Name: Hifza Majeed

Roll No: 19P-1652

Section: BS(SE) 8A

Assignment No 1
Tokenization and segmentation

Introduction

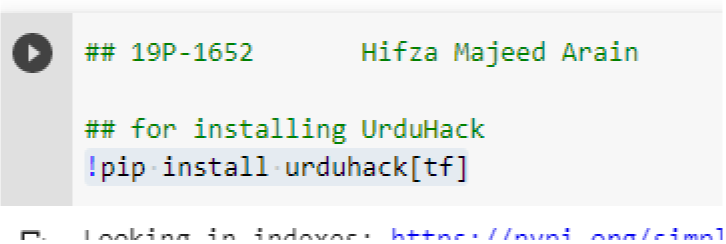
The language I worked with for this project was Urdu, and I separated words and sentences from passages.

Steps to solve the Problem

Step: 1

To install the Urdu hack and import the Urdu hack.

Install the urdu hack through `!pip install urduhack[tf]`.



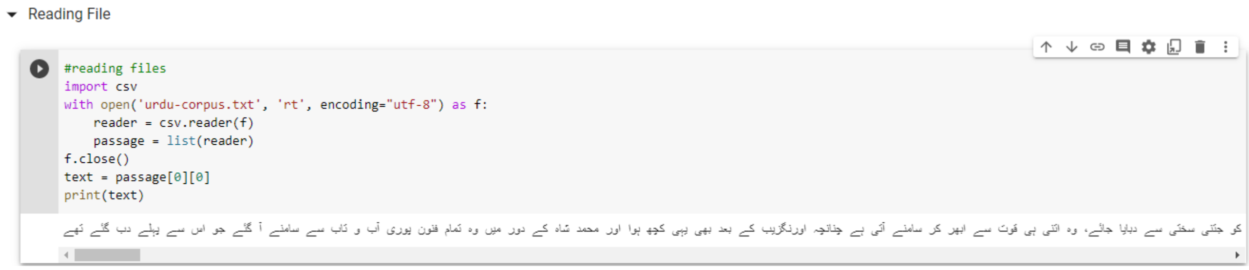
```
## 19P-1652       Hifza Majeed Arain

## for installing UrduHack
!pip install urduhack[tf]
```

```
Looking in indexes: https://pypi.org/simpl
```

Step: 2

Reading file:



Reading File

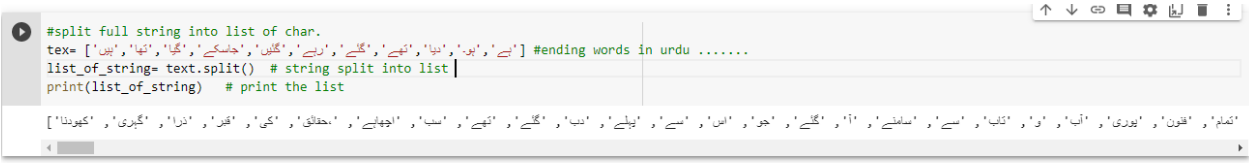
```
#reading files
import csv
with open('urdu-corpus.txt', 'rt', encoding="utf-8") as f:
    reader = csv.reader(f)
    passage = list(reader)
f.close()
text = passage[0][0]
print(text)
```

```
کو جتنی سختی سے دبایا جائے، وہ اتنی ہی قوت سے ابھر کر سامنے آتی ہے چنانچہ اورنگزیب کے بعد بھی یہی کچھ ہوا اور محمد شاہ کے دور میں وہ تمام فنون پوری آب و تاب سے سامنے آ گئے جو اس سے پہلے دب گئے تھے
```

Step: 3

Split full string into list of char

We splits the paragraph into the words.



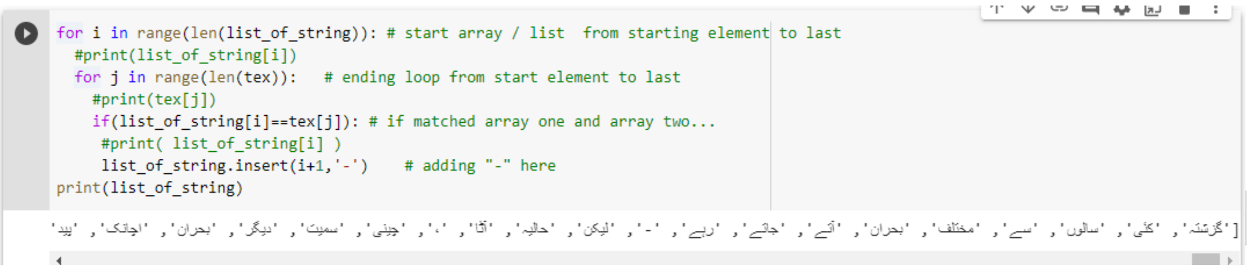
```
#split full string into list of char.
tex= ['ہیں','تھا','گیا','جاسکے','گئیں','رہے','گئے','تھے','دیا','ہو۔','ہے'] #ending words in urdu .......
list_of_string= text.split()  # string split into list
print(list_of_string)   # print the list
```

```
['تمام', 'فنون', 'پوری', 'آپ', 'و', 'ثاب', 'سے', 'سامنے', 'آ', 'گئے', 'جو', 'اس', 'سے', 'پہلے', 'دب', 'گئے', 'تھے', 'سب', 'اچھاہے', '،حقائق', 'کی', 'قبر', 'ذرا', 'گہری', 'کھودنا'
```

Step: 4

Adding the '–' at the end of sentences

we are comparing the paragraph with end sentences word list than add the '–' at end of sentences .



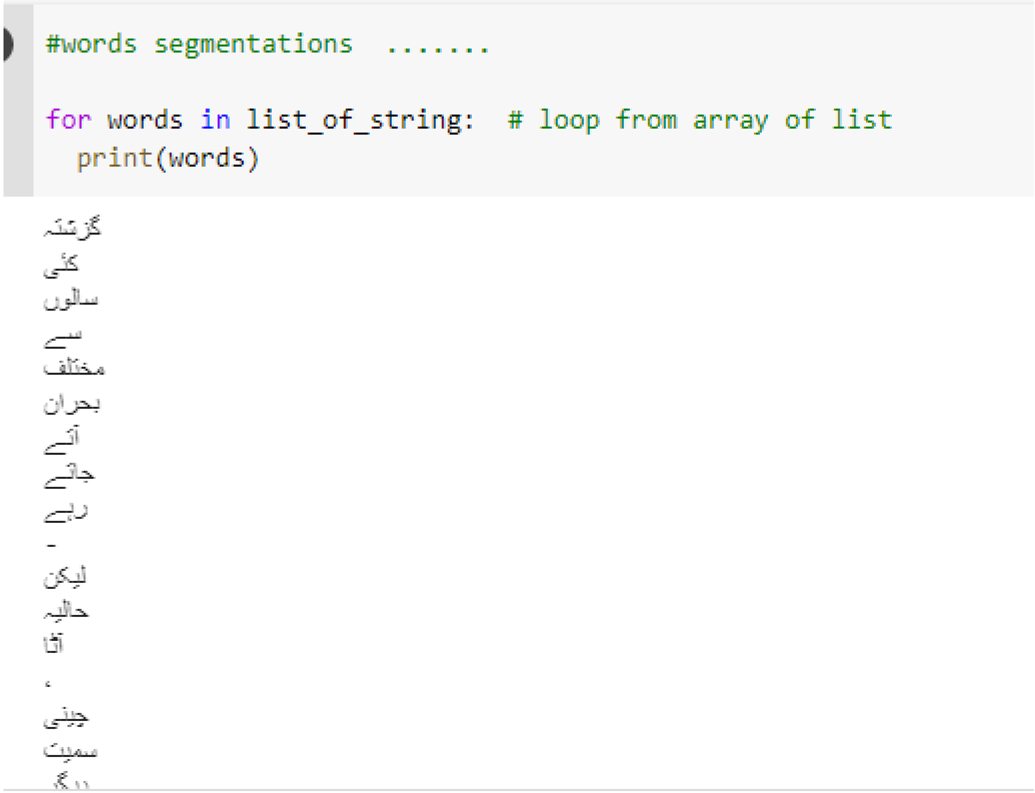
```
for i in range(len(list_of_string)): # start array / list  from starting element to last
  #print(list_of_string[i])
  for j in range(len(tex)):   # ending loop from start element to last
    #print(tex[j])
    if(list_of_string[i]==tex[j]): # if matched array one and array two...
     #print( list_of_string[i] )
     list_of_string.insert(i+1,'-')    # adding "-" here
print(list_of_string)
```

```
['گزشتہ', 'کئی', 'سالوں', 'سے', 'مختلف', 'بحران', 'آتے', 'جاتے', 'رہے', '-', 'لیکن', 'حالیہ', 'آٹا', '،', 'چینی', 'سمیت', 'دیگر', 'بحران', 'اچانک', 'پید'
```

Step: 5

Words tokenization

in list of string having all the word just we apply the loop on the list of string and print the words



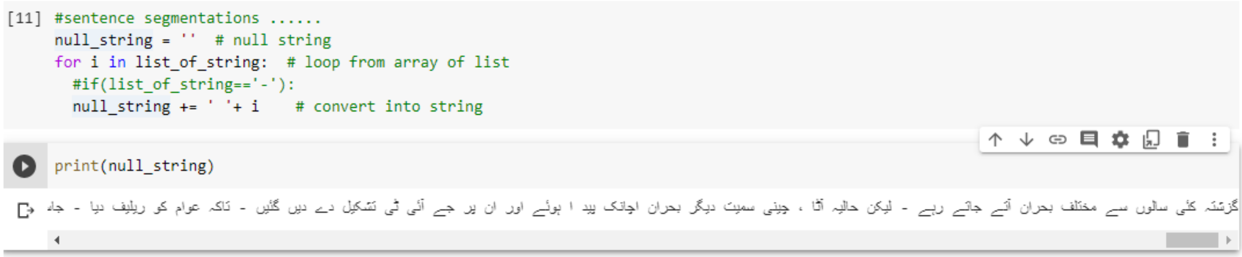
```
#words segmentations  .......

for words in list_of_string:  # loop from array of list
  print(words)
```

```
گزشتہ
کئی
سالوں
سے
مختلف
بحران
آتے
جاتے
رہے
-
لیکن
حالیہ
آٹا
،
چینی
سمیت
دیگر
```

Step: 6

Sentences tokenization



```
[11] #sentence segmentations ......
     null_string = ''  # null string
     for i in list_of_string:  # loop from array of list
       #if(list_of_string=='-'):
       null_string += ' '+ i    # convert into string
```

```
print(null_string)
```

```
گزشتہ کئی سالوں سے مختلف بحران آتے جاتے رہے - لیکن حالیہ آٹا ، چینی سمیت دیگر بحران اچانک پید ا ہوئے اور ان پر جے آئی ٹی تشکیل دے دیں گئیں - تاکہ عوام کو ریلیف دیا - جاہ
```