stemming

August 17, 2021

[1]:
```
#importing Bimba's Algorithsm
import hausastemmer as bimba

#importing Suraj's Improved Algorithsm
import nbimporter
from improved.HausaStemmer import HausaStemmer

#import pandas dataframe library
import pandas as pd
```

[2]:
```
#Using pandas to import dataset containing Normal Words and Already Stemmed␣
↪Words
data = pd.read_csv('test_words.csv')
actual = data.actual_words
expected = data.expected_stem
data
```

[2]:

	actual_words	expected_stem
0	ababen	ababe
1	abin	abin
2	abinci	abinci
3	abincin	abinci
4	abinda	abi
…	…	…
1715	ziyarci	ziyar
1716	zumunci	zumu
1717	zumuncin	zumu
1718	zumunta	zumu
1719	zumuntan	zumu

[1720 rows x 2 columns]

[3]:
```
#Calling Suraj's improved algorithms
suraj = HausaStemmer()

#looping through the Actual words columns to read data
for  item in data.actual_words:
```

```
    #using suraj's improved algorithsm to stem words with the improved lookup␣
 ↪words
    suraj_algo = suraj.stem(item, lookup=True)

    #using bimba's algorithsm to stem words with bimba's lookup words
    bimba_algo = bimba.stem(item, lookup=True)

    #-------exporting the stem output and merge with initial words--------

    # out_put = pd.DataFrame({'Actual_Words': actual, 'Expected_Stem':␣
 ↪expected,
    # 'Suraj_Algorithsm': suraj_algo, 'Bimba_Algorithsm': bimba_algo})
    # out_put.to_csv('stem_output.csv')

#reading the exported stemmmed words output file
stem_data = pd.read_csv('stem_output.csv')
stem_data
```

[3]:

	Actual_Words	Expected_Stem	Suraj_Algorithsm	Bimba_Algorithsm
0	ababen	ababe	ababe	ababe
1	abin	abin	abin	abin
2	abinci	abinci	abinci	abinci
3	abincin	abinci	abinci	abinci
4	abinda	abi	abi	abi
…	…	…	…	…
1715	ziyarci	ziyar	ziya	ziyar
1716	zumunci	zumu	zumu	zumu
1717	zumuncin	zumu	zumu	zumu
1718	zumunta	zumu	zumu	zumu
1719	zumuntan	zumu	zumu	zumu

[1720 rows x 4 columns]

[4]:
```
#comparing both Bimba's algorithsm and Suraj's improved algorithms with␣
 ↪expected_stem
stem_data["Surajs_Correct_Stemming"] =␣
 ↪(stem_data['Expected_Stem']==stem_data['Suraj_Algorithsm'])
stem_data["Bimbas_Correct_Stemming"] =␣
 ↪(stem_data['Expected_Stem']==stem_data['Bimba_Algorithsm'])

#comparing both Bimba's algorithsm and Suraj's improved algorithms with␣
 ↪actual_words
stem_data["Surajs_Unstemm"] =␣
 ↪(stem_data['Actual_Words']==stem_data['Suraj_Algorithsm'])
stem_data["Bimbas_Unstemm"] =␣
 ↪(stem_data['Actual_Words']==stem_data['Bimba_Algorithsm'])
```

```
#Output summary with Suraj's Algorithms
print("----------Surja's Algorithms-------------")
scs = stem_data.Surajs_Correct_Stemming.sum()
suw = stem_data.Surajs_Unstemm.sum()
tscs = scs - suw
sncs = len(stem_data) - stem_data.Surajs_Correct_Stemming.sum()
s_total = suw+tscs+sncs

ptscs = 100/s_total*tscs
psuw = 100/s_total*suw
psncs = 100/s_total*sncs
ps_total = ptscs+psuw+psncs

print("Correctly Stemmed Words =",tscs)
print("Un Correctly Stemmed Words =",sncs)
print("Un Stemmed Words =",suw)
print("Total number of words", s_total)

print("Percentage of Correctly Stemmed Words =",ptscs)
print("Percentage of Un Correctly Stemmed Words =",psncs)
print("Percentage of Un Stemmed Words =",psuw)
print("Percentage of Total number of words", ps_total)

#Output summary with Bimba's Algorithms
print("----------Bimba's Algorithms-------------")
bcs = stem_data.Bimbas_Correct_Stemming.sum()
buw = stem_data.Bimbas_Unstemm.sum()
tbcs = bcs - buw
bscs = len(stem_data) - stem_data.Bimbas_Correct_Stemming.sum()
b_total = buw+tbcs+bscs

ptbcs = 100/s_total*tbcs
pbuw = 100/s_total*buw
pbncs = 100/s_total*bscs
pb_total = ptbcs+pbuw+pbncs

print("Correctly Stemmed Words =",tbcs)
print("Un Correctly Stemmed Words =",bscs)
print("Un Stemmed Words =",buw)
print("Total number of words", b_total)

print("Percentage of Correctly Stemmed Words =",ptbcs)
print("Percentage of Un Correctly Stemmed Words =",pbncs)
print("Percentage of Un Stemmed Words =",pbuw)
```

```
print("Percentage of Total number of words", pb_total)
```

```
----------Surja's Algorithms-------------
Correctly Stemmed Words = 1120
Un Correctly Stemmed Words = 127
Un Stemmed Words = 473
Total number of words 1720
Percentage of Correctly Stemmed Words = 65.11627906976744
Percentage of Un Correctly Stemmed Words = 7.383720930232558
Percentage of Un Stemmed Words = 27.5
Percentage of Total number of words 100.0
----------Bimba's Algorithms-------------
Correctly Stemmed Words = 1187
Un Correctly Stemmed Words = 2
Un Stemmed Words = 531
Total number of words 1720
Percentage of Correctly Stemmed Words = 69.01162790697674
Percentage of Un Correctly Stemmed Words = 0.11627906976744186
Percentage of Un Stemmed Words = 30.872093023255815
Percentage of Total number of words 100.0
```