stemming

August 16, 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 algorithsm 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'])
stem_data
```

[4]:

	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

      Surajs_Correct_Stemming  Bimbas_Correct_Stemming
0                        True                    True
1                        True                    True
2                        True                    True
3                        True                    True
4                        True                    True
…                           …                       …
1715                     False                   True
1716                      True                   True
1717                      True                   True
1718                      True                   True
1719                      True                   True

[1720 rows x 6 columns]
```

[5]:
```
#counting correct stemmed words with Suraj's Improved Algorithsm
surajs_correct_stemmed_words = stem_data['Surajs_Correct_Stemming'].
 ↪value_counts()
surajs_correct_stemmed_words

#"True" means correctly stemmed
#"False" means wrongly stemmed
```

[5]:
```
True     1593
False     127
Name: Surajs_Correct_Stemming, dtype: int64
```

[6]:
```
#counting correct stemmed words with Bimba's Improved Algorithsm
bimbas_correct_stemmed_words = stem_data['Bimbas_Correct_Stemming'].
 ↪value_counts()
bimbas_correct_stemmed_words

#"True" means correctly stemmed
#"False" means wrongly stemmed
```

[6]:
```
True     1718
False       2
Name: Bimbas_Correct_Stemming, dtype: int64
```