Rana

* Corpus consists of 2320742 words.
* To remove all non Arabic and unwanted characters using sed

$ sed 's/[a-z|A-Z|0-9!”@#$%-^&\*\/\]//g' L165\_diacritized\_unprocessed.txt > L165\_diacritized\_processed\_complete.txt

* To remove white spaces and extra lines

$ sed '/^ \*$/d

* Download pyArabic to use its library so we can use various methods
  + pip install pyarabic
  + import pyarabic.araby as araby
  + import pyarabic.number as number
* Strip out all diacritics using pyArabic

import pyarabic.araby as araby

input\_file = open ("L165\_diacritized\_processed\_complete.txt", "r")

words = input\_file.read()

input\_file.close()

output\_words = araby.strip\_tashkeel(words)

output\_file = open ("L165\_undiacritized.txt", "w")

output\_file.write(output\_words)

* Strip last diacritic only:

strip\_lastharaka(text)

* Removing white spaces

import re

f = open ("/Users/rana/Desktop/L165\_corpus1.txt", "r")

file = f.read()

z = file.lstrip()

trim\_space = re.sub(' +',' ', z)

output\_file = open ("/Users/rana/Desktop/L165\_diacritized\_corpus2.txt", 'w')

output\_file.write(trim\_space)

output\_file.close()