**Ideas for Data CleanUp Code**

***Removing dirty code from BBC.com Yoruba & Igbo (foreign languages)***

Parse through all text characters

Where next character = <

Set i = 1

Count next character

if next character /= < or >

add character to string of previous characters

Then

Count next character

elif next character = <

i+1

And add character to previous string

Count next character

elif next first letter = >

i-1

Add to previous string

Count next character

When i = 0

Replace previous string with " "

Rewrite the same code with subject {

**Removing dirty code from BBC.com English**

if next character = (

Check next Character

if next character = r

append to previous string

Check next character

if next character = e

append to previous string

Check next character

if next character = q

append to previous string

i = 1

Initiate < type protocol above

else abort

**Separating words joined by Punctuation**

letters = { list of all unicode letters }

punctuation = { list of all Punctuation strings }

numbers = { 0 1 2 3 4 5 6 7 8 9 }

letters\_or\_numbers = letters + numbers

Check next character

if character is element of punctuation

Check next character

if character is element of letters\_or\_numbers

add " " before it

elif next character = " " abort

**Using NLKT**

*To take out html code using Beautifulsoup html.parser*

specific\_doc = Loadtxt

from bs4 import BeautifulSoup

raw = BeautifulSoup(specific\_doc, 'html.parser').get\_text()

*To transform to list of tokens (words)*

raw = tokenize()

for i in raw:

for char in i:

for j in punctuations:

if char == j:

if i.index(char) < (len(i) - 1):

i.split(char)

else: abort

elif char /= j abort