XML Scrapping

```
# XML scrapping for xml sheet


import os

os.chdir(r"D:\Samson - All Data\Naresh IT Institute\New folder\xml_single articles")


import xml.etree.ElementTree as ET


tree = ET.parse("769952.xml")

root = tree.getroot()


root=ET.tostring(root, encoding='utf8').decode('utf8')


root


import re, string, unicodedata

import nltk


from bs4 import BeautifulSoup

from nltk import word_tokenize, sent_tokenize

from nltk.corpus import stopwords

from nltk.stem import LancasterStemmer, WordNetLemmatizer


def strip_html(text):

    soup = BeautifulSoup(text, "html.parser")

    return soup.get_text()
```

```
def remove_between_square_brackets(text):
    return re.sub('\[[^]]*\]', '', text)


def denoise_text(text):
    text = strip_html(text)
    text = remove_between_square_brackets(text)
    text=re.sub('  ','',text)
    return text


sample = denoise_text(root)
#print(sample)
```