**Name :afnan alhassni**

**Id:438006383**

**1-Extracting articles from HTML:**

import nltk

from nltk import sent\_tokenize, word\_tokenize,pos\_tag

from urllib.request import urlopen

from bs4 import BeautifulSoup

import requests

url=requests.get('https://feeds.simplecast.com/qm\_9xx0g')

soup= BeautifulSoup(url.text, 'html.parser')

import re

CLEANR = re.compile('<.\*?>|&([a-z0-9]+|#[0-9]{1,6}|#x[0-9a f]{1,6});')

s = soup.find\_all('item')

text=""

for story in s:

title = story.find('title')

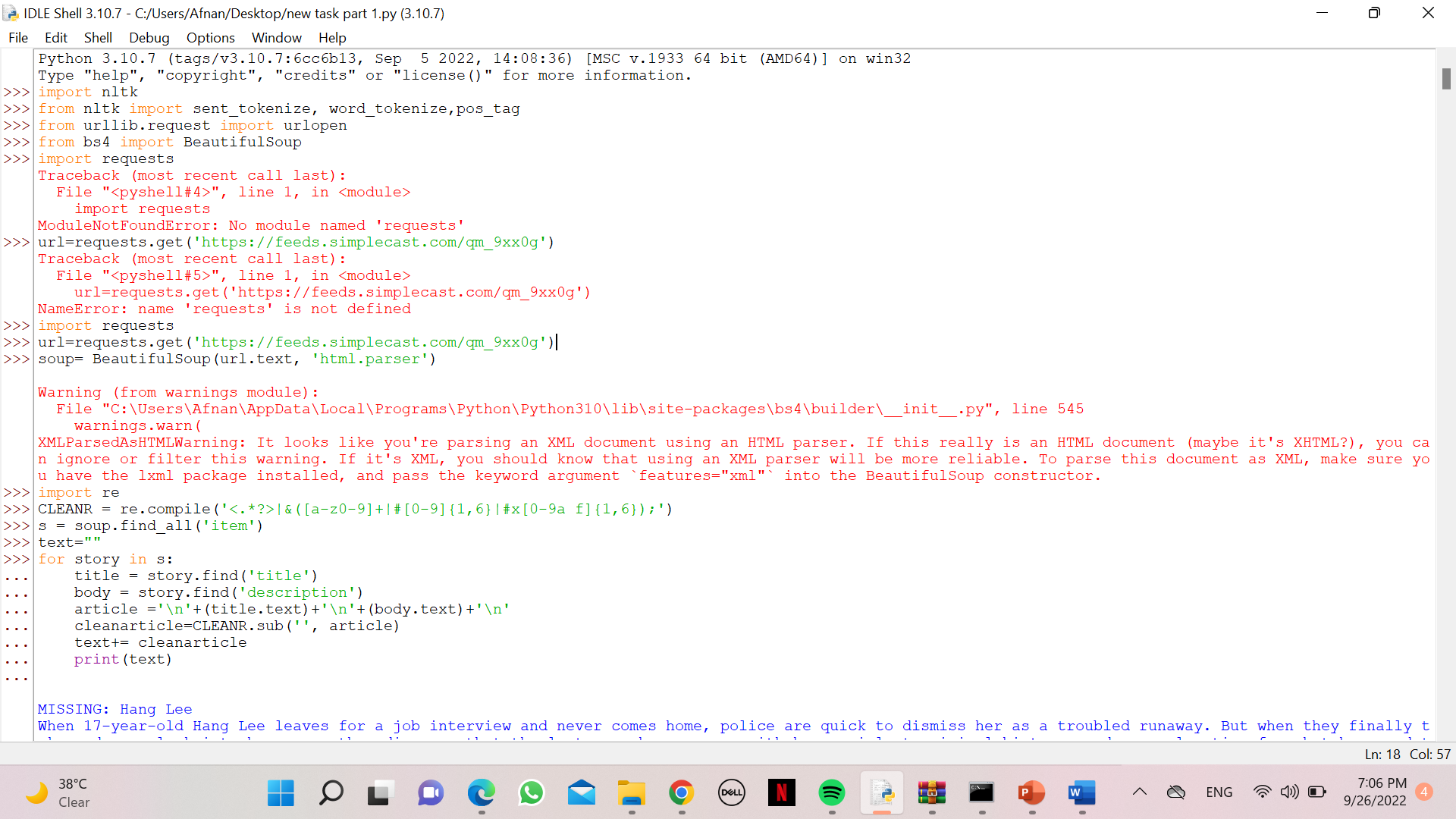
body = story.find('description')

article ='\n'+(title.text)+'\n'+(body.text)+'\n'

cleanarticle=CLEANR.sub('', article)

text+= cleanarticle

print(text)



Text

Description automatically generated

**2- tokenization:**

nltk.download('punkt')

[nltk\_data] Downloading package punkt to

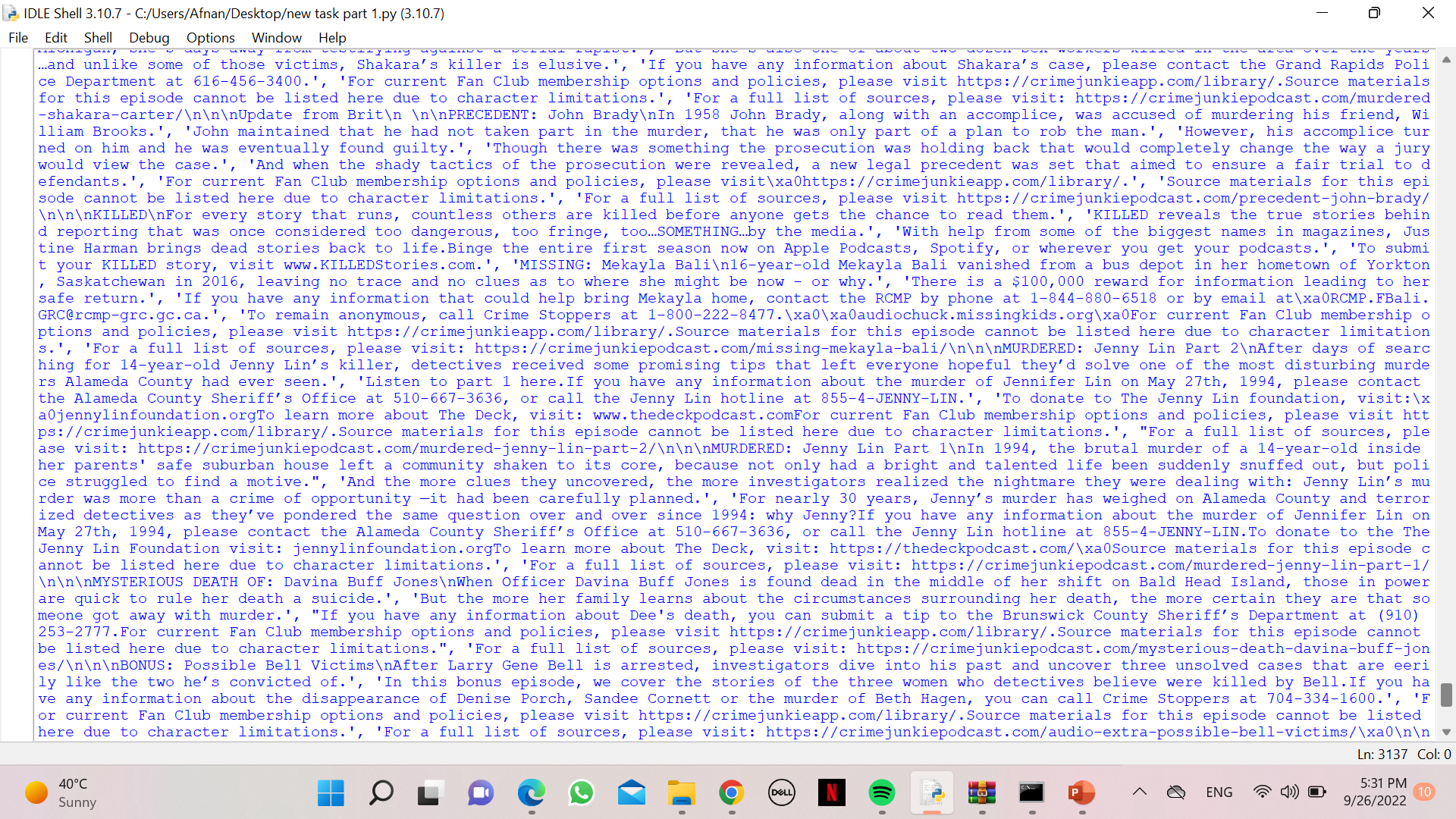
[nltk\_data] C:\Users\Afnan\AppData\Roaming\nltk\_data...

[nltk\_data] Unzipping tokenizers\punkt.zip.

True

s= sent\_tokenize(text)

print('Sentence :', s,'\n')



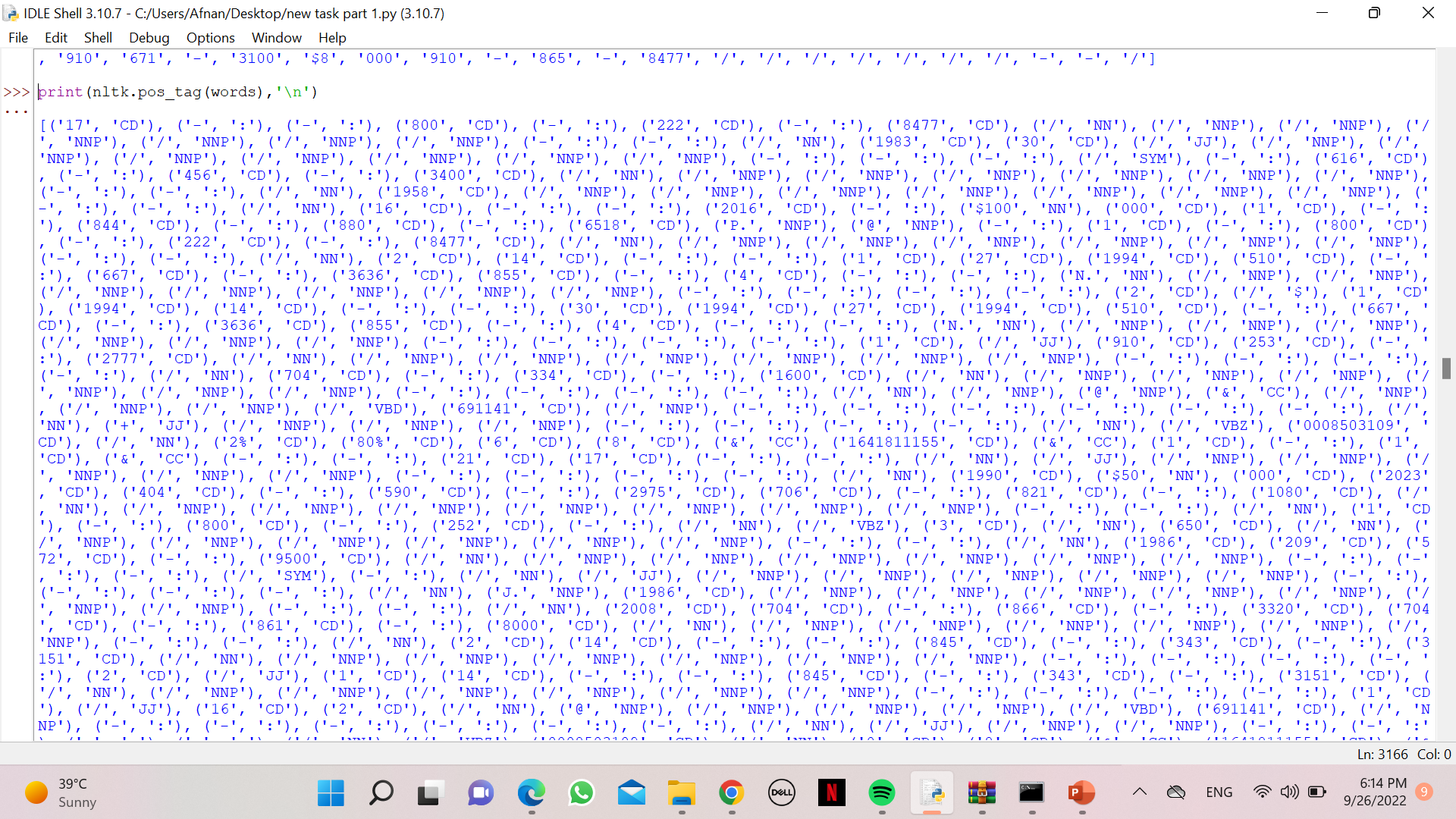
Word1= word\_tokenize(text)

print('Words:', word1 ,'\n')

A screenshot of a computer

Description automatically generated with medium confidence

**3-tag :**

****

**3-stop word :**

nltk.download('stopwords')

[nltk\_data] Downloading package stopwords to

[nltk\_data] C:\Users\Afnan\AppData\Roaming\nltk\_data...

[nltk\_data] Unzipping corpora\stopwords.zip.

True

stopwords= set(stopwords.words('english'))

wordsFiltered=[]

for w in words:

if w not in stopwords:

wordsFiltered.append(w);



**4- morphological :**

from nltk.stem import PorterStemmer, WordNetLemmatizer

for word in wordsFiltered:

print(PorterStemmer().stem(word))

print(WordNetLemmatizer().lemmatize(word))

Graphical user interface, text, application, Word

Description automatically generated

Graphical user interface, text, application, Word

Description automatically generated

Graphical user interface, text, application, Word

Description automatically generated