Web Mining Lab Assignment 1

Name: Om Ashish Mishra

Registration Number: 16BCE0789

Slot: F2

Stop word Removal

1. Write a program to remove the stopwords for any given paragraph. Create a set of stop words given below and print the output stop\_words =['.',',','a','they','the','his','so','and','were','from','that','of','in','only','with','to']

The Code:

from nltk.corpus import stopwords

from nltk.tokenize import word\_tokenize

example\_sent = "This is a sample sentence, showing off the stop words filtration."

stop\_words = set(stopwords.words('english'))

word\_tokens = word\_tokenize(example\_sent)

filtered\_sentence = []

for w in word\_tokens:

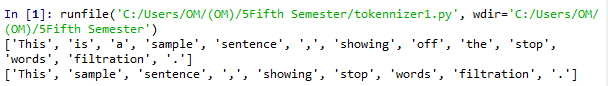
if w not in stop\_words:

filtered\_sentence.append(w)

print(word\_tokens)

print(filtered\_sentence)

The Output:



2. Write a program to tokenize

a) A sentence

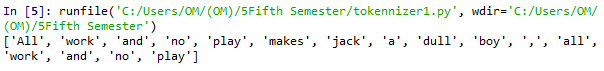
The Code:

from nltk.tokenize import sent\_tokenize, word\_tokenize

data = "All work and no play makes jack a dull boy, all work and no play"

print(word\_tokenize(data))

The Output:



b) Multiple sentences

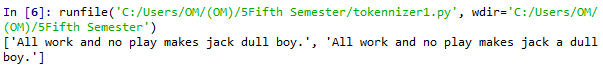
The Code:

from nltk.tokenize import sent\_tokenize, word\_tokenize

data = "All work and no play makes jack dull boy. All work and no play makes jack a dull boy."

print(sent\_tokenize(data))

The Output:



3. Write a program (using nltk toolkit in python environment) to tokenize

a) Sentence

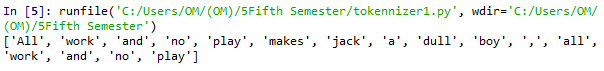
The Code:

from nltk.tokenize import sent\_tokenize, word\_tokenize

data = "All work and no play makes jack a dull boy, all work and no play"

print(word\_tokenize(data))

The Output:



b) Multiple sentences

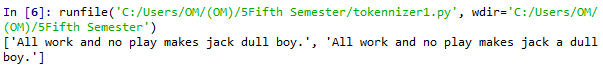
The Code:

from nltk.tokenize import sent\_tokenize, word\_tokenize

data = "All work and no play makes jack dull boy. All work and no play makes jack a dull boy."

print(sent\_tokenize(data))

The Output:



c) A paragraphe

The Code

import re

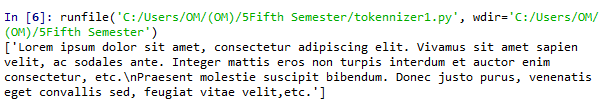
from nltk.tokenize import sent\_tokenize, word\_tokenize

x = """Lorem ipsum dolor sit amet, consectetur adipiscing elit. Vivamus sit amet sapien velit, ac sodales ante. Integer mattis eros non turpis interdum et auctor enim consectetur, etc.

Praesent molestie suscipit bibendum. Donec justo purus, venenatis eget convallis sed, feugiat vitae velit,etc."""

print(re.split('\s{4,}',x))

The Output:



d) Information of a complete web page

The Code:

from urllib import request

from nltk.tokenize import sent\_tokenize, word\_tokenize

url = "http://news.bbc.co.uk/2/hi/health/2284783.stm"

html = request.urlopen(url).read().decode('utf8')

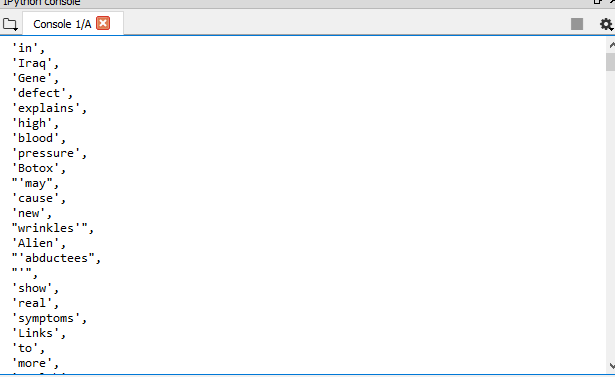
html[:60]

from bs4 import BeautifulSoup

raw = BeautifulSoup(html).get\_text()

tokens = word\_tokenize(raw)

tokens

The Output: 

Stemming

The Code:

from nltk.stem import PorterStemmer

from nltk.tokenize import sent\_tokenize, word\_tokenize

ps = PorterStemmer()

example\_words = ["python","pythoner","pythoning","pythoned","pythonly"]

for w in example\_words:

print(ps.stem(w))

The Output:

