**AIM: Implement a simple web crawle**r

**Program**

import requests

import lxml

from bs4 import BeautifulSoup

url = "https://www.rottentomatoes.com/top/bestofrt/"

headers = {'User-Agent': 'Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.36 '

'(KHTML,like Gecko) Chrome/63.0.3239.132 Safari/537.36 QIHU 360SE' }

f = requests.get(url,headers = headers)

movies\_lst = []

soup =BeautifulSoup(f.content,'html.parser')

movies =soup.find('table', { 'class': 'table' }) .find\_all('a')

print(movies)

num = 0

for anchor in movies:

urls = 'https://www.rottentomatoes.com' + anchor['href']

movies\_lst.append(urls)

print(movies\_lst)

num += 1

movie\_url = urls

movie\_f = requests.get(movie\_url,headers = headers)

movie\_soup = BeautifulSoup(movie\_f.content,'lxml')

movie\_content = movie\_soup.find('div', {'class' : 'movies\_synopsis clamp clamp-6 js-clamp'

})

print(num, urls, '\n', 'Movie:' + anchor.string.strip())

print('Movie Info:' + movie\_content.string.strip())

**Output**

[<a class="unstyled articleLink" href="/m/it\_happened\_one\_night">

It Happened One Night (1934)</a>, <a class="unstyled articleLink" href="/m/citizen\_kane">

Citizen Kane (1941)</a>, <a class="unstyled articleLink" href="/m/the\_wizard\_of\_oz\_1939">

The Wizard of Oz (1939)</a>, <a class="unstyled articleLink" href="/m/modern\_times">

Modern Times (1936)</a>, <a class="unstyled articleLink" href="/m/black\_panther\_2018">

Black Panther (2018)</a>, <a class="unstyled articleLink" href="/m/parasite\_2019">

Parasite (Gisaengchung) (2019)</a>, <a class="unstyled articleLink" href="/m/avengers\_endgame">

Avengers: Endgame (2019)</a>, <a class="unstyled articleLink" href="/m/1003707-casablanca">

Casablanca (1942)</a>, <a class="unstyled articleLink" href="/m/knives\_out">

Knives Out (2019)</a>, <a class="unstyled articleLink" href="/m/us\_2019">

Us (2019)</a>, <a class="unstyled articleLink" href="/m/toy\_story\_4">

Toy Story 4 (2019)</a>, <a class="unstyled articleLink" href="/m/lady\_bird">

Lady Bird (2017)</a>, <a class="unstyled articleLink" href="/m/mission\_impossible\_fallout">

Mission: Impossible - Fallout (2018)</a>, <a class="unstyled articleLink" href="/m/blackkklansman">

BlacKkKlansman (2018)</a>, <a class="unstyled articleLink" href="/m/get\_out">

Get Out (2017)</a>, <a class="unstyled articleLink" href="/m/the\_irishman">

The Irishman (2019)</a>, <a class="unstyled articleLink" href="/m/godfather">

The Godfather (1972)</a>, <a class="unstyled articleLink" href="/m/mad\_max\_fury\_road">

Mad Max: Fury Road (2015)</a>, <a class="unstyled articleLink" href="/m/spider\_man\_into\_the\_spider\_verse">

Spider-Man: Into the Spider-Verse (2018)</a>, <a class="unstyled articleLink" href="/m/moonlight\_2016">

Moonlight (2016)</a>, <a class="unstyled articleLink" href="/m/sunset\_boulevard">

Sunset Boulevard (1950)</a>, <a class="unstyled articleLink" href="/m/1000626-all\_about\_eve">

All About Eve (1950)</a>, <a class="unstyled articleLink" href="/m/the\_cabinet\_of\_dr\_caligari">

The Cabinet of Dr. Caligari (Das Cabinet des Dr. Caligari) (1920)</a>, <a class="unstyled articleLink" href="/m/philadelphia\_story">

The Philadelphia Story (1940)</a>, <a class="unstyled articleLink" href="/m/roma\_2018">

Roma (2018)</a>, <a class="unstyled articleLink" href="/m/wonder\_woman\_2017">

Wonder Woman (2017)</a>, <a class="unstyled articleLink" href="/m/a\_star\_is\_born\_2018">

A Star Is Born (2018)</a>, <a class="unstyled articleLink" href="/m/inside\_out\_2015">

Inside Out (2015)</a>, <a class="unstyled articleLink" href="/m/a\_quiet\_place\_2018">

A Quiet Place (2018)</a>, <a class="unstyled articleLink" href="/m/one\_night\_in\_miami">

One Night in Miami (2020)</a>, <a class="unstyled articleLink" href="/m/eighth\_grade">

Eighth Grade (2018)</a>, <a class="unstyled articleLink" href="/m/1017293-rebecca">

Rebecca (1940)</a>, <a class="unstyled articleLink" href="/m/booksmart">

Booksmart (2019)</a>, <a class="unstyled articleLink" href="/m/logan\_2017">

Logan (2017)</a>, <a class="unstyled articleLink" href="/m/his\_girl\_friday">

His Girl Friday (1940)</a>, <a class="unstyled articleLink" href="/m/portrait\_of\_a\_lady\_on\_fire">

Portrait of a Lady on Fire (Portrait de la jeune fille en feu) (2020)</a>, <a class="unstyled articleLink" href="/m/coco\_2017">

Coco (2017)</a>, <a class="unstyled articleLink" href="/m/dunkirk\_2017">

Dunkirk (2017)</a>, <a class="unstyled articleLink" href="/m/star\_wars\_the\_last\_jedi">

Star Wars: The Last Jedi (2017)</a>, <a class="unstyled articleLink" href="/m/1015002-night\_at\_the\_opera">

A Night at the Opera (1935)</a>, <a class="unstyled articleLink" href="/m/the\_shape\_of\_water\_2017">

The Shape of Water (2017)</a>, <a class="unstyled articleLink" href="/m/thor\_ragnarok\_2017">

Thor: Ragnarok (2017)</a>, <a class="unstyled articleLink" href="/m/spotlight\_2015">

Spotlight (2015)</a>, <a class="unstyled articleLink" href="/m/the\_farewell\_2019">

The Farewell (2019)</a>, <a class="unstyled articleLink" href="/m/selma">

Selma (2014)</a>, <a class="unstyled articleLink" href="/m/the\_third\_man">

The Third Man (1949)</a>, <a class="unstyled articleLink" href="/m/1017289-rear\_window">

Rear Window (1954)</a>, <a class="unstyled articleLink" href="/m/et\_the\_extraterrestrial">

E.T. The Extra-Terrestrial (1982)</a>, <a class="unstyled articleLink" href="/m/seven\_samurai\_1956">

Seven Samurai (Shichinin no Samurai) (1956)</a>, <a class="unstyled articleLink" href="/m/la\_grande\_illusion">

La Grande illusion (Grand Illusion) (1938)</a>, <a class="unstyled articleLink" href="/m/arrival\_2016">

Arrival (2016)</a>, <a class="unstyled articleLink" href="/m/singin\_in\_the\_rain">

Singin' in the Rain (1952)</a>, <a class="unstyled articleLink" href="/m/the\_favourite\_2018">

The Favourite (2018)</a>, <a class="unstyled articleLink" href="/m/double\_indemnity">

Double Indemnity (1944)</a>, <a class="unstyled articleLink" href="/m/1000642-all\_quiet\_on\_the\_western\_front">

All Quiet on the Western Front (1930)</a>, <a class="unstyled articleLink" href="/m/1048445-snow\_white\_and\_the\_seven\_dwarfs">

Snow White and the Seven Dwarfs (1937)</a>, <a class="unstyled articleLink" href="/m/marriage\_story\_2019">

Marriage Story (2019)</a>, <a class="unstyled articleLink" href="/m/the\_big\_sick">

The Big Sick (2017)</a>, <a class="unstyled articleLink" href="/m/on\_the\_waterfront">

On the Waterfront (1954)</a>, <a class="unstyled articleLink" href="/m/star\_wars\_episode\_vii\_the\_force\_awakens">

Star Wars: Episode VII - The Force Awakens (2015)</a>, <a class="unstyled articleLink" href="/m/american\_in\_paris">

An American in Paris (1951)</a>, <a class="unstyled articleLink" href="/m/best\_years\_of\_our\_lives">

The Best Years of Our Lives (1946)</a>, <a class="unstyled articleLink" href="/m/1013775-metropolis">

Metropolis (1927)</a>, <a class="unstyled articleLink" href="/m/boyhood">

Boyhood (2014)</a>, <a class="unstyled articleLink" href="/m/gravity\_2013">

Gravity (2013)</a>, <a class="unstyled articleLink" href="/m/leave\_no\_trace">

Leave No Trace (2018)</a>, <a class="unstyled articleLink" href="/m/1013139-maltese\_falcon">

The Maltese Falcon (1941)</a>, <a class="unstyled articleLink" href="/m/the\_invisible\_man\_2020">

The Invisible Man (2020)</a>, <a class="unstyled articleLink" href="/m/12\_years\_a\_slave">

12 Years a Slave (2013)</a>, <a class="unstyled articleLink" href="/m/once\_upon\_a\_time\_in\_hollywood">

Once Upon a Time In Hollywood (2019)</a>, <a class="unstyled articleLink" href="/m/argo\_2012">

Argo (2012)</a>, <a class="unstyled articleLink" href="/m/soul\_2020">

Soul (2020)</a>, <a class="unstyled articleLink" href="/m/ma\_raineys\_black\_bottom">

Ma Rainey's Black Bottom (2020)</a>, <a class="unstyled articleLink" href="/m/1052609-kid">

The Kid (1921)</a>, <a class="unstyled articleLink" href="/m/manchester\_by\_the\_sea">

Manchester by the Sea (2016)</a>, <a class="unstyled articleLink" href="/m/nosferatu">

Nosferatu, a Symphony of Horror (Nosferatu, eine Symphonie des Grauens) (Nosferatu the Vampire) (1922)</a>, <a class="unstyled articleLink" href="/m/1000355-adventures\_of\_robin\_hood">

The Adventures of Robin Hood (1938)</a>, <a class="unstyled articleLink" href="/m/la\_la\_land">

La La Land (2016)</a>, <a class="unstyled articleLink" href="/m/north-by-northwest">

North by Northwest (1959)</a>, <a class="unstyled articleLink" href="/m/1012007-laura">

Laura (1944)</a>, <a class="unstyled articleLink" href="/m/spider\_man\_far\_from\_home">

Spider-Man: Far From Home (2019)</a>, <a class="unstyled articleLink" href="/m/incredibles\_2">

Incredibles 2 (2018)</a>, <a class="unstyled articleLink" href="/m/zootopia">

Zootopia (2016)</a>, <a class="unstyled articleLink" href="/m/alien">

Alien (1979)</a>, <a class="unstyled articleLink" href="/m/1011615-king\_kong">

King Kong (1933)</a>, <a class="unstyled articleLink" href="/m/1018688-shadow\_of\_a\_doubt">

Shadow of a Doubt (1943)</a>, <a class="unstyled articleLink" href="/m/call\_me\_by\_your\_name">

Call Me by Your Name (2018)</a>, <a class="unstyled articleLink" href="/m/psycho">

Psycho (1960)</a>, <a class="unstyled articleLink" href="/m/1917\_2019">

1917 (2020)</a>, <a class="unstyled articleLink" href="/m/la\_confidential">

L.A. Confidential (1997)</a>, <a class="unstyled articleLink" href="/m/the\_florida\_project">

The Florida Project (2017)</a>, <a class="unstyled articleLink" href="/m/war\_for\_the\_planet\_of\_the\_apes">

War for the Planet of the Apes (2017)</a>, <a class="unstyled articleLink" href="/m/paddington\_2">

Paddington 2 (2018)</a>, <a class="unstyled articleLink" href="/m/beatles\_a\_hard\_days\_night">

A Hard Day's Night (1964)</a>, <a class="unstyled articleLink" href="/m/widows\_2018">

Widows (2018)</a>, <a class="unstyled articleLink" href="/m/never\_rarely\_sometimes\_always">

Never Rarely Sometimes Always (2020)</a>, <a class="unstyled articleLink" href="/m/baby\_driver">

Baby Driver (2017)</a>, <a class="unstyled articleLink" href="/m/spider\_man\_homecoming">

Spider-Man: Homecoming (2017)</a>, <a class="unstyled articleLink" href="/m/godfather\_part\_ii">

The Godfather, Part II (1974)</a>, <a class="unstyled articleLink" href="/m/the\_battle\_of\_algiers">

The Battle of Algiers (La Battaglia di Algeri) (1967)</a>]

['https://www.rottentomatoes.com/m/it\_happened\_one\_night']

['https://www.rottentomatoes.com/m/it\_happened\_one\_night', 'https://www.rottentomatoes.com/m/citizen\_kane']

['https://www.rottentomatoes.com/m/it\_happened\_one\_night', 'https://www.rottentomatoes.com/m/citizen\_kane', 'https://www.rottentomatoes.com/m/the\_wizard\_of\_oz\_

**AIM: Implement simple web crawler (importing data to csv file)**

**Program**

from bs4 import BeautifulSoup

import requests

pages\_crawled = []

def crawler(url):

page = requests.get(url)

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

links = soup.find\_all('a')

for link in links:

if 'href' in link.attrs:

if link['href'].startswith('/wiki') and ':' not in link['href']:

if link['href'] not in pages\_crawled:

new\_link = f"https://en.wikipedia.org{link['href']}"

pages\_crawled.append(link['href'])

try:

with open('data.csv','a') as file:

file.write(f'{soup.title.text}; {soup.h1.text}; {link["href"]}\n')

crawler(new\_link)

except:

continue

crawler('https://en.wikipedia.org')

crawler()

**Output**

**AIM:Implement a program to scrap the web page of any popular website.**

**Program**

import requests

from bs4 import BeautifulSoup

import csv

URL = "http://www.values.com/inspirational-quotes"

r = requests.get(URL)

print(r.content)

soup = BeautifulSoup(r.content, 'lxml')

print(soup.prettify())

quotes = []

table = soup.find('div',attrs = {'id': 'all\_quotes'})

for row in table.findAll('div',attrs = {'class': 'col-6 col-lg-3 text-center margin-30px-bottom sm-margin-30px-top'}):

quote = {}

quote['theme'] = row.h5.text

quote['url'] = row.a['href']

quote['img'] = row.img['src']

quote['lines'] = row.img['alt'].split(" #")[0]

quote['author'] = row.img['alt'].split("#")[1]

quotes.append(quote)

filename = 'inspirtaional\_quotes.csv'

with open(filename, 'w', newline='') as f:

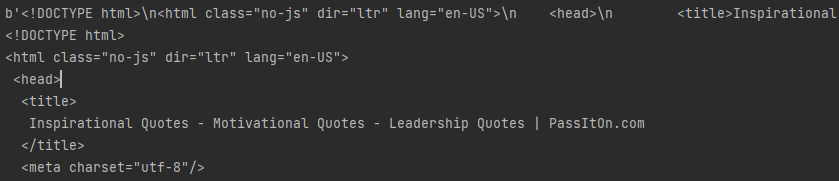
w = csv.DictWriter(f, ['theme','url','img','lines','author'])

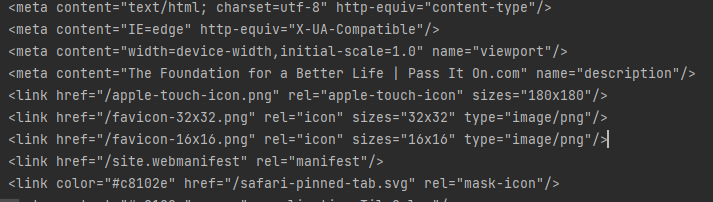
w.writeheader()

for quote in quotes:

w.writerow(quote)

Output







**AIM:Python program for natural program processing (ngram without using inbuilt functions)**

**Program**

def generate\_ngrams(text,WordsToCombine):

words = text.split()

output =[]

for i in range(len(words) - WordsToCombine + 1):

output.append(words[i:i + WordsToCombine])

return output

x = generate\_ngrams (text ='this is very good book to study',WordsToCombine=3)

print(x)

**Output**

C:\Users\mca\PycharmProjects\exam\venv\Scripts\python.exe "C:/Users/mca/PycharmProjects/exam/n gram.py"

[['this', 'is', 'very']]

**AIM:Python natural language processing for speech tag**

**Program**

import nltk

from nltk.corpus import stopwords

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

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

txt ="hello how are you."\

"This is my frind."\

"This is my notebook."\

"Russians celebrate the october revaluation in the month of november."\

"frienship is scared bond between people."\

"but never found the right one."

tokenized = sent\_tokenize(txt)

for i in tokenized:

wordsList = nltk.word\_tokenize(i)

wordsList = [w for w in wordsList if not w in stop\_words]

tagged = nltk.pos\_tag(wordsList)

print(tagged)

**Output**

**AIM: Python program for natural language processing N-gram(with using inbuilt function)**

**Program**

import nltk

from nltk.util import ngrams

nltk.download('punkt')

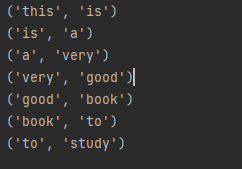
samplText = 'this is a very good book to study'

NGRAMS = ngrams(sequence=nltk.word\_tokenize(samplText), n=2)

for grams in NGRAMS:

print(grams)

**Output**

****

**import requests**

**r = requests.get('https://www.geeksforgeeks.org/python-programming-language/')**

**print(r)**

**print(r.content)**