UNIVERSITY OF MUMBAI

DEPARTMENT OF COMPUTER SCIENCE

M.SC (Computer Science)

CERTIFICATE

Certified that the work entered in this journal was done in the computer laboratory by the student

Mr. Karankumar Yadav

Seat No.112002

Of the Class - MSc computer Science

First Year Sem 2

Semester during the year **2021-2022** in a Satisfactory manner.

Practical 1

AIM :- Scrape an online E-Commerce Site for Data.

1. Extract product data from Amazon - be it any product and put these details in the MySQL database. One can use pipeline. Like 1 pipeline to process the scraped data and other to put data in the database and since Amazon has some restrictions on scraping of data, ask them to work on small set of requests otherwise proxies and all would have to be used.
2. Scrape the details like color, dimensions, material etc. Or customer ratings by features.

THEORY :-

Web scraping is the process of collecting structured web data in an automated fashion. It's also called web data extraction. Some of the main use cases of web scraping include price monitoring, price intelligence, news monitoring, lead generation, and market research among many others.

CODE / OUTPUT :-

```
pip install kora -q


import csv

from bs4 import BeautifulSoup


from kora.selenium import wd

wd.get('https://www.amazon.in/')



wd.page_source
```

```
Out[19]: '<html lang="en-in" class=" a-js a-audio a-video a-canvas a-svg a-drag-drop a-geolocation a-history a-webworker a-autofocus
         a-input-placeholder a-textarea-placeholder a-local-storage a-gradients a-transform3d a-touch-scrolling a-text-shadow a-text-
         stroke a-box-shadow a-border-radius a-border-image a-opacity a-transform a-transition a-ember" data-19ax5a9jf="dingo" data-a
         ui-build-date="3.22.1-2022-05-18" data-useragent="Mozilla/5.0 (X11; Linux x86_64) AppleWebKit/537.36 (KHTML, like Gecko) Hea
         dlessChrome/101.0.4951.64 Safari/537.36" data-platform="Linux x86_64"><!-- sp:feature:head-start --><head><script async="" s
         rc="https://c.amazon-adsystem.com/bao-csm/forensics/a9-tq-forensics-incremental.min.js" crossorigin="anonymous"></script><sc
         ript async="" src="https://images-eu.ssl-images-amazon.com/images/I/31YXrY93hfL.js" crossorigin="anonymous"></script><script
         >var aPageStart = (new Date()).getTime();</script><meta charset="utf-8">\n<!-- sp:end-feature:head-start -->\n\n<script type
         ="text/javascript">var ue_t0=ue_t0||+new Date();</script>\n<!-- sp:feature:cs-optimization -->\n<meta http-equiv="x-dns-pref
         etch-control" content="on">\n<link rel="dns-prefetch" href="https://images-eu.ssl-images-amazon.com">\n<link rel="dns-prefet
         ch" href="https://m.media-amazon.com">\n<link rel="dns-prefetch" href="https://completion.amazon.com">\n<!-- sp:end-feature:
         cs-optimization -->\n<script type="text/javascript">\nwindow.ue_ihb = (window.ue_ihb || window.ueinit || 0) + 1;\nif (windo
         w.ue_ihb === 1) {\n\nvar ue_csm = window,\n    ue_hob = +new Date();\n(function(d){var e=d.ue=d.ue||{},f=Date.now||function
         (){return+new Date};e.d=function(b){return f()-(b?0:d.ue_t0)};e.stub=function(b,a){if(!b[a]){var c=[];b[a]=function(){c.push
         ([c.slice.call(arguments),e.d(),d.ue_id])};b[a].replay=function(b){for(var a;a=c.shift();)b(a[0],a[1],a[2])};b[a].isStub=
         1}};e.exec=function(b,a){return function(){try{return b.apply(this,arguments)}catch(c){ueLogError(c,{attribution:a||"undefin
         ed",logLevel:"WARN"})}}}})(ue_csm);\n\n\n    var ue_err_chan = \'jserr-rw\';\n(function(d,e){function h(f,b){if(!(a.ec>a.mx
         e)&&f){a.ter.push(f);b=b||{};var c=f.logLevel||b.logLevel;c&&c!==k&&c!==m&&c!==n&&c!==p||a.ec++;c&&c!=k||a.ecf++;b.pageURL
         =""+(e.location?e.location.href:"");b.logLevel=c;b.attribution=f.attribution||b.attribution;a.erl.push({ex:f,info:b})}}funct
```

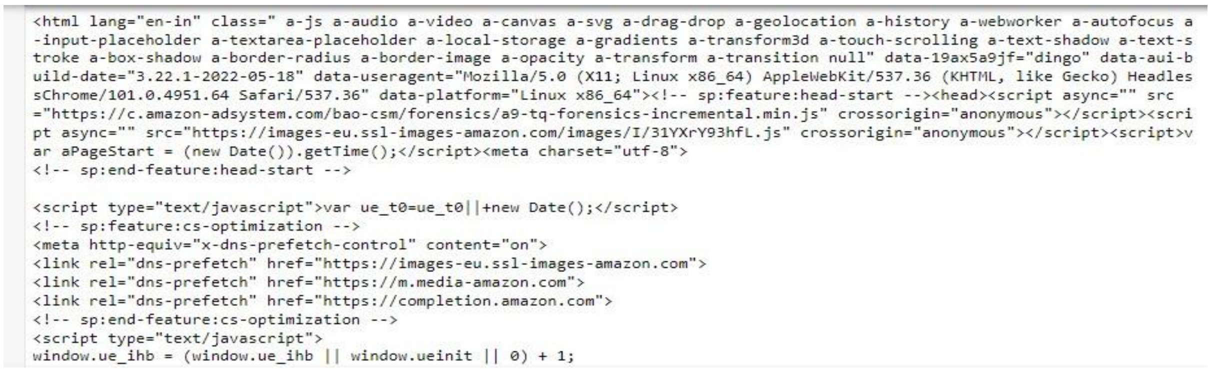
```
ion l(a,b,c,e,g){d.ueLogError({m:a,f:b,l:c,c:""+e,err:g,fromOnError:1,args:arguments},g?{attribution:g.attribution,logLevel:
g.logLevel}:void 0);return!1}var k="FATAL",m="ERROR",n="WARN",p="DOWNGRADED",a={ec:0,ecf:0,\npec:0,ts:0,erl:[],ter:[],mxe:5
0,startTimer:function(){a.ts++;setInterval(function(){d.ue&&a.pec<a.ec&&d.uex("at");a.pec=a.ec},1E4)}};l.skipTrace=1;h.skipT
race=1;h.isStub=1;d.ueLogError=h;d.ue_err=a;e.onerror=l})(ue_csm,window);\n\n\nvar ue_id = \'NBPNAXCMTDD9TMDWTBD8\',\n    ue
_url = \'/rd/uedata\',\n    ue_navtiming = 1,\n    ue_mid = \'A21TJRUUN4KGV\',\n    ue_sid = \'257-8565408-9856116\',\n    u
e_sn = \'www.amazon.in\',\n    ue_furl = \'fls-eu.amazon.in\',\n    ue_surl = \'https://unagi-eu.amazon.com/1/events/com.ama
zon.csm.nexusclient.prod\',\n    ue_int = 0,\n    ue_fcsn = 1,\n    ue_urt = 3,\n    ue_rpl_ns = \'cel-rpl\',\n    ue_ddq =
1,\n    ue_fpf = \'//fls-eu.amazon.in/1/batch/1/OP/A21TJRUUN4KGV:257-8565408-9856116:NBPNAXCMTDD9TMDWTBD8$uedata=s:\',\n
ue_sbuimp = 1,\n    ue_cel_lclia = 1,\n    ue_ibft = 0,\n    ue_sswmts = 0,\n    ue_jsmtf = 0,\n    ue_fnt = 0,\n\n    ue_sw
i = 1;\nvar ue_viz=function(){(function(b,e,a){function k(c){if(b.ue.viz.length<p&&!l){var a=c.type;c=c.originalEvent;/^focu
s./.test(a)&&c&&(c.toElement||c.fromElement||c.relatedTarget)||(a=e[m]||("blur"==a||"focusout"==a?"hidden":"visible"),b.ue.v
iz.push(a+":"+(+new Date-b.ue.t0)),"visible"==a&&(b.ue.isl&&q("at"),l=1))}}for(var l=0,q=b.uex,f,g,m,n=["","webkit","o","m
s","moz"],d=0,p=20,h=0;h<n.length&&!d;h++)if(a=n[h],f=(a?a+"H":"h")+"idden",d="boolean"==typeof e[f])g=a+"visibilitychange",
m=(a?a+"V":"v")+\n"isibilityState";k({});d&&e.addEventListener(g,k,0);b.ue&&d&&(b.ue.pageViz={event:g,propHid:f})})(ue_csm,u
e_csm.document,ue_csm.window)};\n\n(function(d,h,N){function H(a){return a&&a.replace&&a.replace(/^\\s+|\\s+$/g,"")}function
u(a){return"undefined"===typeof a}function B(a,b){for(var c in b)b[v](c)&&(a[c]=b[c])}function I(a){try{var b=N.cookie.match
(RegExp("(^| )"+a+"=([^;]+)"));if(b)return b[2].trim()}catch(c){}}function O(m,b,c){var q=(x||{}).type;if("device"!==c||2!==
q&&1!==q)m&&(d.ue_id=a.id=a.rid=m,w&&(w=w.replace(/((.*?:){2})(\\w+)/,function(a,b){return b+m})),D&&(e("id",D,m),D=0)),b&&
(w&&(w=w.replace(/(.*?:)(\\w|-)+/,function(a,\nc){return c+b})),d.ue_sid=b),c&&a.tag("page-source:"+c),d.ue_fpf=w}function P
```

```
var f,l;if(b||u(c)){if(d)for(l in f=b?e("t",b)||e("t",b,{}):a.t,f[d]=q,c)c[v](l)&&e(l,b,c[l]);return q}}function e(d,b,c){va
r e=b&&b!=a.id?a.sc[b]:a;e||(e=a.sc[b]={});"id"===d&&c&&(Q=1);return e[d]=c||e[d]}function R(d,b,c,e,f){c="on"+c;var l=b
[c];"function"===typeof l?d&&(a.h[d]=l):l=function(){};b[c]=\nfunction(a){f?(e(a),l(a)):(l(a),e(a))};b[c]&&(b[c].isUeh=1)}fu
nction S(m,b,c,q){function p(b,c){var d=[b],g=0,f={},l,h;c?(d.push("m=1"),f[c]=1):f=a.sc;for(h in f)if(f[v](h)){var q=e("w
b",h),p=e("t",h)||{},n=e("t0",h)||a.t0,k;if(c||2==q){q=q?g++:"";d.push("sc"+q+"="+h);for(k in p)u(p[k])||null===p[k]||d.push
(k+q+"="+(p[k]-n));d.push("t"+q+"="+p[m]);if(e("ctb",h)||e("wb",h))l=1}}!J&&l&&d.push("ctb=1");return d.join("&")}function l
(b,c,g,e){if(b){var f=d.ue_err;d.ue_url&&!e&&b&&0<b.length&&(e=new Image,\na.iel.push(e),e.src=b,a.count&&a.count("postbackI
mageSize",b.length));if(w){var m=h.encodeURIComponent;m&&b&&(e=new Image,b=""+d.ue_fpf+m(b)+":"+(+new E-d.ue_t0),a.iel.push
(e),e.src=b)}else a.log&&(a.log(b,"uedata",{n:1}),a.ielf.push(b));f&&!f.ts&&f.startTimer();a.b&&(f=a.b,a.b="",l(f,c,g,1))}}f
unction A(b){var c=x?x.type:F,d=2==c||a.isBFonMshop,c=c&&!d,e=a.bfini;Q||(e&&1<e&&(b+="&bfform=1",c||(a.isBFT=e-1)),d&&(b+
="&bfnt=1",a.isBFT=a.isBFT||1),a.ssw&&a.isBFT&&(a.isBFonMshop&&(a.isNRBF=0),u(a.isNRBF)&&\n(d=a.ssw(a.oid),d.e||u(d.val)||
(a.isNRBF=1<d.val?0:1)),u(a.isNRBF)||(b+="&nrbf="+a.isNRBF)),a.isBFT&&!a.isNRBF&&(b+="&bft="+a.isBFT));return b}if(!a.paused
&&(b||u(c))){for(var k in c)c[v](k)&&e(k,b,c[k]);a.isBFonMshop||y("pc",b,c);k="ld"===m&&b&&e("wb",b);var s=e("id",b)||a.id;k
||s===a.oid||(D=b,ba(s,(e("t",b)||{}).tc||+e("t0",b),+e("t0",b)));var s=e("id",b)||a.id,t=e("id2",b),g=a.url+"?"+m+"&v="+a.v
+"&id="+s,J=e("ctb",b)||e("wb",b),z;J&&(g+="&ctb="+J);t&&(g+="&id2="+t);1<d.ueinit&&(g+="&ic="+d.ueinit);\nif(!("ld"!=m&&"u
l"!=m||b&&b!=s)){if("ld"==m){try{h[K]&&h[K].isUeh&&(h[K]=null)}catch(I){}if(h.chrome)for(t=0;t<L.length;t++)T(G,L[t]);(t=N.u
e_backdetect)&&t.ue_back&&t.ue_back.value++;d._uess&&(z=d._uess());a.isl=1}a._bf&&(g+="&bf="+a._bf());d.ue_navtiming&&f&&(e
("ctb",s,"1"),a.isBFonMshop||y("tc",F,F,M));!C||a.isBFonMshop||U||(f&&B(a.t,{na_:f.navigationStart,ul_:f.unloadEventStart,_u
l:f.unloadEventEnd,rd_:f.redirectStart,_rd:f.redirectEnd,fe_:f.fetchStart,lk_:f.domainLookupStart,_lk:f.domainLookupEnd,\nco
```

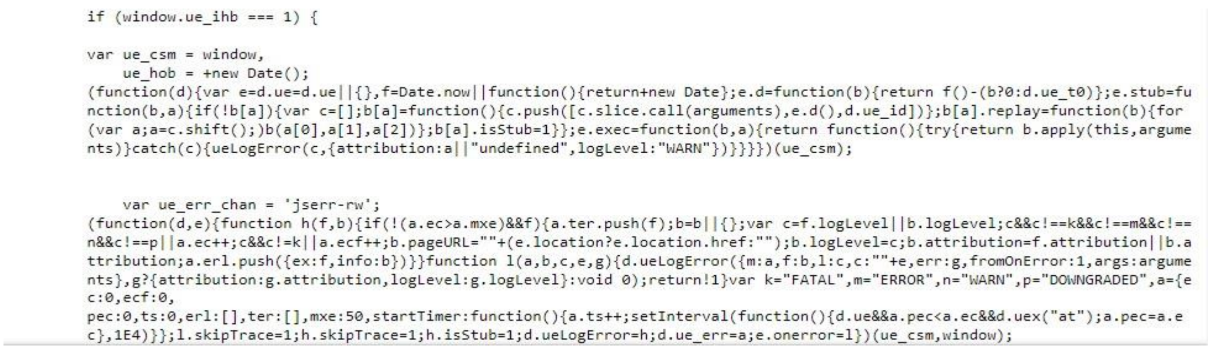
```
n d?c(d.downlink||0):null,downlinkMax:"downlinkMax"in d?c(d.downlinkMax||0):null,rtt:"rtt"in d?(d.rtt||0).toFixed():null,typ
e:d.type||null,effectiveType:d.effectiveType||null,saveData:"saveData"in d?d.saveData:null})}function k(){v(),u()}function w
(){l(),u()}l(),u(),e.on("$afterPageTransition",w),e.on(d,"change",k)});\nif (window.ue && window.ue.uels) {\n    ue.uels("ht
tps://c.amazon-adsystem.com/bao-csm/forensics/a9-tq-forensics-incremental.min.js");\n}\n\n\nue.exec(function(d,c){function g
(e,c){e&&ue.tag(e+c);return!!e}function n(){for(var e=RegExp("^https://(.*\\.(images|ssl-images|media)-amazon\\.com|"+c.loca
tion.hostname+")/images/","i"),d={},h=0,k=c.performance.getEntriesByType("resource"),l=!1,b,a,m,f=0;f<k.length;f++)if(a=k
[f],0<a.transferSize&&a.transferSize>=a.encodedBodySize&&(b=e.exec(String(a.name)))&&3===b.length){a:{b=a.serverTiming||[];f
or(a=0;a<b.length;a++)if("provider"===b[a].name){b=b[a].description;break a}b=void 0}b&&(l||(l=g(b,"_cdn_fr")),\na=d[b]=(d
[b]||0)+1,a>h&&(m=b,h=a))}g(m,"_cdn_mp")}d.ue&&"function"===typeof d.ue.tag&&c.performance&&c.location&&n()},"cdnTagging")(u
e_csm,window);\n\n\n}\n/* ⚠ */\n</script>\n\n</div>\n\n<noscript>\n    <img height="1" width="1" style=\'display:none;visibi
lity:hidden;\' src=\'//fls-eu.amazon.in/1/batch/1/OP/A21TJRUUN4KGV:257-8565408-9856116:NBPNAXCMTDD9TMDWTBD8$uedata=s:%2Frd%2
Fuedata%3Fnoscript%26id%3DNBPNAXCMTDD9TMDWTBD8:0\' alt=""/>\n</noscript>\n\n<script>window.ue && ue.count && ue.count(\'CSML
ibrarySize\', 59905)</script>\n</div>\n<!--        _\n       .__(.)< (MEOW)\n        \\___)   \n ~~~~~~~~~~~~~~~~~~-->\n<!--
sp:eh:XI/YVNnEbNeOe4aGwO7Nyg7p3L7sZiyNnRV6BuQPPzSnoiwusSwYsINoYu0ZZwZUFR+QxCH01tz+hDG9AenLXiKHW2xYfodimq/1xO6PcnaS+tI6y0V1k+
fQxdg= -->\n<div id="a-popover-root" style="z-index:-1;position:absolute;"></div><iframe id="3gb7vpf722s" height="5px" width
="5px" src="javascript:\'1\'" style="position: absolute; top: -10000px; left: -10000px; visibility: hidden; opacity: 0;"></i
frame><iframe id="sdff40rer8f" height="5px" width="5px" src="javascript:\'1\'" style="position: absolute; top: -10000px; lef
t: -10000px; visibility: hidden; opacity: 0;"></iframe></body></html>'
```

```
from selenium import webdriver

options = webdriver.ChromeOptions()

options.add_argument('-headless')

options.add_argument('-no-sandbox')
```

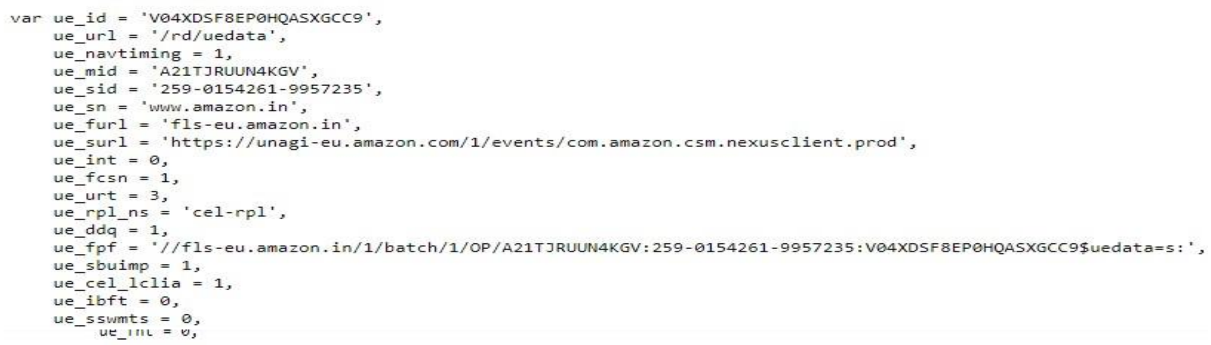
```
options.add_argument('-disable-dev-shm-usage')


wd = webdriver.Chrome('chromedriver',options=options)

wd.get("https://www.amazon.in/")

print(wd.page_source) # results
```



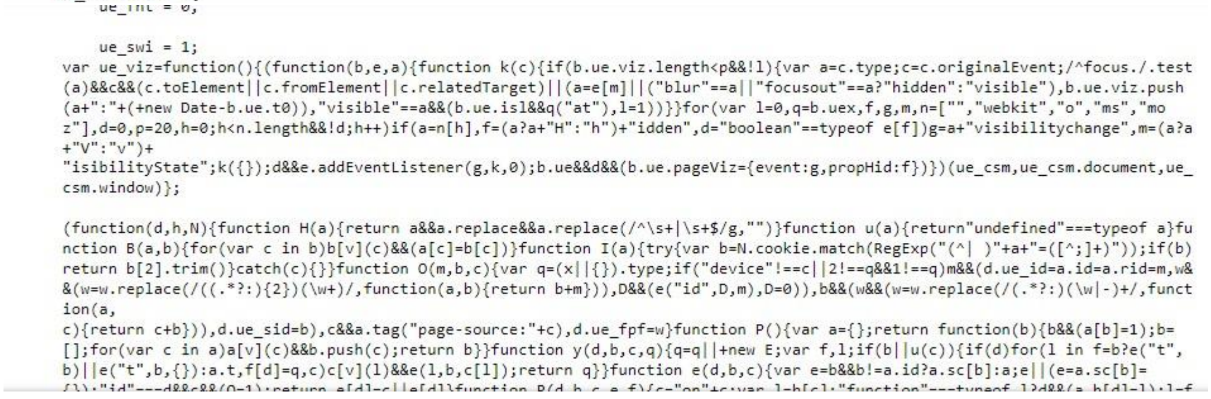
```
<html lang="en-in" class=" a-js a-audio a-video a-canvas a-svg a-drag-drop a-geolocation a-history a-webworker a-autofocus a
-input-placeholder a-textarea-placeholder a-local-storage a-gradients a-transform3d a-touch-scrolling a-text-shadow a-text-s
troke a-box-shadow a-border-radius a-border-image a-opacity a-transform a-transition null" data-19ax5a9jf="dingo" data-aui-b
uild-date="3.22.1-2022-05-18" data-useragent="Mozilla/5.0 (X11; Linux x86_64) AppleWebKit/537.36 (KHTML, like Gecko) Headles
sChrome/101.0.4951.64 Safari/537.36" data-platform="Linux x86_64"><!-- sp:feature:head-start --><head><script async="" src
="https://c.amazon-adsystem.com/bao-csm/forensics/a9-tq-forensics-incremental.min.js" crossorigin="anonymous"></script><scri
pt async="" src="https://images-eu.ssl-images-amazon.com/images/I/31YXrY93hfL.js" crossorigin="anonymous"></script><script>v
ar aPageStart = (new Date()).getTime();</script><meta charset="utf-8">
<!-- sp:end-feature:head-start -->

<script type="text/javascript">var ue_t0=ue_t0||+new Date();</script>
<!-- sp:feature:cs-optimization -->
<meta http-equiv="x-dns-prefetch-control" content="on">
<link rel="dns-prefetch" href="https://images-eu.ssl-images-amazon.com">
<link rel="dns-prefetch" href="https://m.media-amazon.com">
<link rel="dns-prefetch" href="https://completion.amazon.com">
<!-- sp:end-feature:cs-optimization -->
<script type="text/javascript">
window.ue_ihb = (window.ue_ihb || window.ueinit || 0) + 1;
```



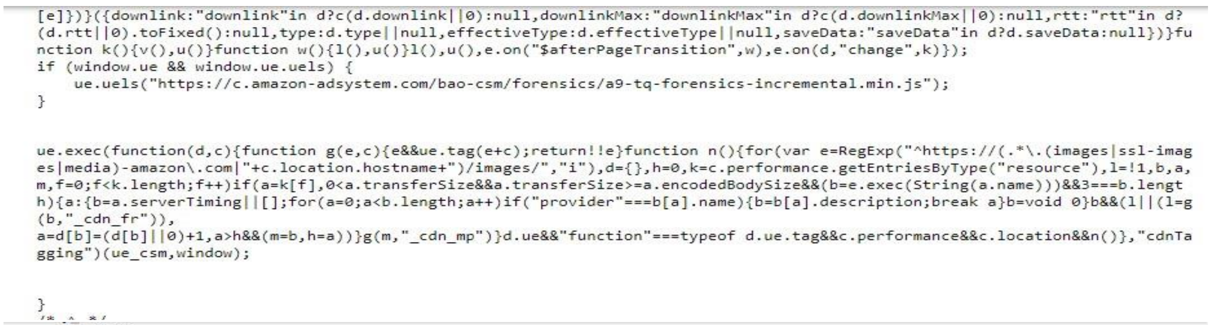
```
    if (window.ue_ihb === 1) {

    var ue_csm = window,
        ue_hob = +new Date();
    (function(d){var e=d.ue=d.ue||{},f=Date.now||function(){return+new Date};e.d=function(b){return f()-(b?0:d.ue_t0)};e.stub=fu
    nction(b,a){if(!b[a]){var c=[];b[a]=function(){c.push([c.slice.call(arguments),e.d(),d.ue_id])};b[a].replay=function(b){for
    (var a;a=c.shift();)b(a[0],a[1],a[2])};b[a].isStub=1}};e.exec=function(b,a){return function(){try{return b.apply(this,argume
    nts)}catch(c){ueLogError(c,{attribution:a||"undefined",logLevel:"WARN"})}}}})(ue_csm);


        var ue_err_chan = 'jserr-rw';
    (function(d,e){function h(f,b){if(!(a.ec>a.mxe)&&f){a.ter.push(f);b=b||{};var c=f.logLevel||b.logLevel;c&&c!==k&&c!==m&&c!==
    n&&c!==p||a.ec++;c&&c!=k||a.ecf++;b.pageURL=""+(e.location?e.location.href:"");b.logLevel=c;b.attribution=f.attribution||b.a
    ttribution;a.erl.push({ex:f,info:b})}}function l(a,b,c,e,g){d.ueLogError({m:a,f:b,l:c,c:""+e,err:g,fromOnError:1,args:argume
    nts},g?{attribution:g.attribution,logLevel:g.logLevel}:void 0);return!1}var k="FATAL",m="ERROR",n="WARN",p="DOWNGRADED",a={e
    c:0,ecf:0,
    pec:0,ts:0,erl:[],ter:[],mxe:50,startTimer:function(){a.ts++;setInterval(function(){d.ue&&a.pec<a.ec&&d.uex("at");a.pec=a.e
    c},1E4)}};l.skipTrace=1;h.skipTrace=1;h.isStub=1;d.ueLogError=h;d.ue_err=a;e.onerror=l})(ue_csm,window);
```



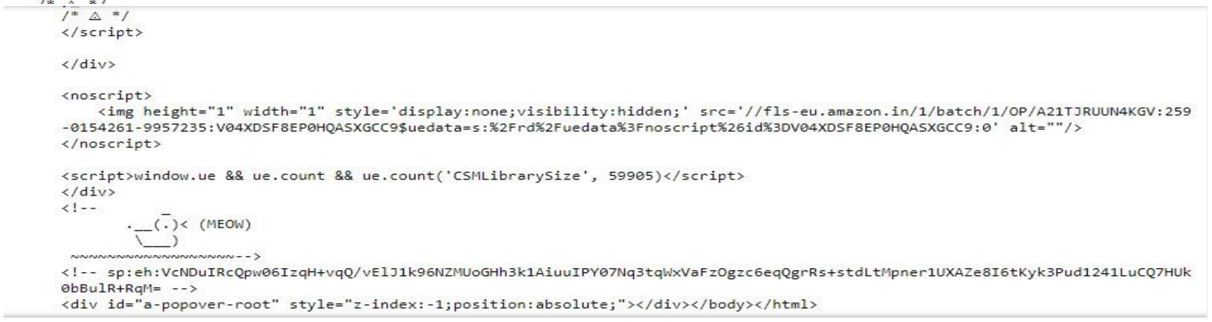
```
var ue_id = 'V04XDSF8EP0HQASXGCC9',
    ue_url = '/rd/uedata',
    ue_navtiming = 1,
    ue_mid = 'A21TJRUUN4KGV',
    ue_sid = '259-0154261-9957235',
    ue_sn = 'www.amazon.in',
    ue_furl = 'fls-eu.amazon.in',
    ue_surl = 'https://unagi-eu.amazon.com/1/events/com.amazon.csm.nexusclient.prod',
    ue_int = 0,
    ue_fcsn = 1,
    ue_urt = 3,
    ue_rpl_ns = 'cel-rpl',
    ue_ddq = 1,
    ue_fpf = '//fls-eu.amazon.in/1/batch/1/OP/A21TJRUUN4KGV:259-0154261-9957235:V04XDSF8EP0HQASXGCC9$uedata=s:',
    ue_sbuimp = 1,
    ue_cel_lclia = 1,
    ue_ibft = 0,
    ue_sswmts = 0,
```



```
      ue_swi = 1;
  var ue_viz=function(){(function(b,e,a){function k(c){if(b.ue.viz.length<p&&!l){var a=c.type;c=c.originalEvent;/^focus./.test
  (a)&&c&&(c.toElement||c.fromElement||c.relatedTarget)||(a=e[m]||("blur"==a||"focusout"==a?"hidden":"visible"),b.ue.viz.push
  (a+":"+(+new Date-b.ue.t0)),"visible"==a&&(b.ue.isl&&q("at"),l=1))}}for(var l=0,q=b.uex,f,g,m,n=["","webkit","o","ms","mo
  z"],d=0,p=20,h=0;h<n.length&&!d;h++)if(a=n[h],f=(a?a+"H":"h")+"idden",d="boolean"==typeof e[f])g=a+"visibilitychange",m=(a?a
  +"V":"v")+
  "isibilityState";k({});d&&e.addEventListener(g,k,0);b.ue&&d&&(b.ue.pageViz={event:g,propHid:f})})(ue_csm,ue_csm.document,ue_
  csm.window)};

  (function(d,h,N){function H(a){return a&&a.replace&&a.replace(/^\s+|\s+$/g,"")}function u(a){return"undefined"===typeof a}fu
  nction B(a,b){for(var c in b)b[v](c)&&(a[c]=b[c])}function I(a){try{var b=N.cookie.match(RegExp("(^| )"+a+"=([^;]+)"));if(b)
  return b[2].trim()}catch(c){}}function O(m,b,c){var q=(x||{}).type;if("device"!==c||2!==q&&1!==q)m&&(d.ue_id=a.id=a.rid=m,w&
  &(w=w.replace(/((.*?:){2})(\w+)/,function(a,b){return b+m})),D&&(e("id",D,m),D=0)),b&&(w&&(w=w.replace(/(.*?:)(\w|-)+/,funct
  ion(a,
  c){return c+b})),d.ue_sid=b),c&&a.tag("page-source:"+c),d.ue_fpf=w}function P(){var a={};return function(b){b&&(a[b]=1);b=
  [];for(var c in a)a[v](c)&&b.push(c);return b}}function y(d,b,c,q){q=q||+new E;var f,l;if(b||u(c)){if(d)for(l in f=b?e("t",
  b)||e("t",b,{}):a.t,f[d]=q,c)c[v](l)&&e(l,b,c[l]);return q}}function e(d,b,c){var e=b&&b!=a.id?a.sc[b]:a;e||(e=a.sc[b]=
```

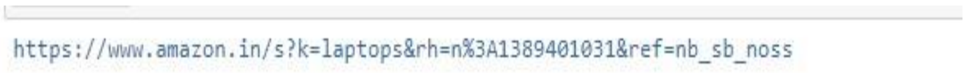


```
[e]})}({downlink:"downlink"in d?c(d.downlink||0):null,downlinkMax:"downlinkMax"in d?c(d.downlinkMax||0):null,rtt:"rtt"in d?
(d.rtt||0).toFixed():null,type:d.type||null,effectiveType:d.effectiveType||null,saveData:"saveData"in d?d.saveData:null})}fu
nction k(){v(),u()}function w(){l(),u()}l(),u(),e.on("$afterPageTransition",w),e.on(d,"change",k)});
if (window.ue && window.ue.uels) {
    ue.uels("https://c.amazon-adsystem.com/bao-csm/forensics/a9-tq-forensics-incremental.min.js");
}


ue.exec(function(d,c){function g(e,c){e&&ue.tag(e+c);return!!e}function n(){for(var e=RegExp("^https://(.*\.(images|ssl-imag
es|media)-amazon\.com|"+c.location.hostname+")/images/","i"),d={},h=0,k=c.performance.getEntriesByType("resource"),l=!1,b,a,
m,f=0;f<k.length;f++)if(a=k[f],0<a.transferSize&&a.transferSize>=a.encodedBodySize&&(b=e.exec(String(a.name)))&&3===b.lengt
h){a:{b=a.serverTiming||[];for(a=0;a<b.length;a++)if("provider"===b[a].name){b=b[a].description;break a}b=void 0}b&&(l||(l=g
(b,"_cdn_fr")),
a=d[b]=(d[b]||0)+1,a>h&&(m=b,h=a))}g(m,"_cdn_mp")}d.ue&&"function"===typeof d.ue.tag&&c.performance&&c.location&&n()},"cdnTa
gging")(ue_csm,window);


}
```



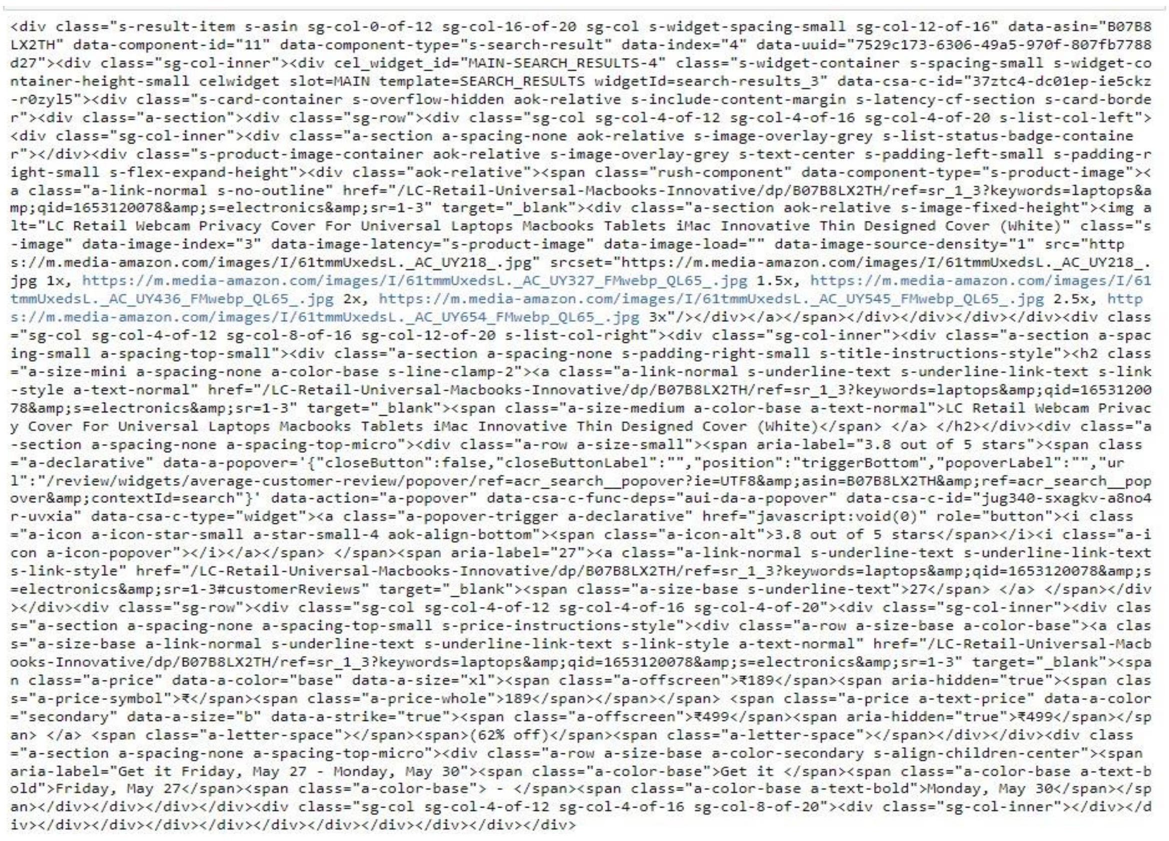
```
  /* △ */
  </script>

  </div>

  <noscript>
      <img height="1" width="1" style='display:none;visibility:hidden;' src='//fls-eu.amazon.in/1/batch/1/OP/A21TJRUUN4KGV:259
  -0154261-9957235:V04XDSF8EP0HQASXGCC9$uedata=s:%2Frd%2Fuedata%3Fnoscript%26id%3DV04XDSF8EP0HQASXGCC9:0' alt=""/>
  </noscript>

  <script>window.ue && ue.count && ue.count('CSMLibrarySize', 59905)</script>
  </div>
  <!--
             _
         .__(.)< (MEOW)
          \___)
   ~~~~~~~~~~~~~~~~~~-->
  <!-- sp:eh:VcNDuIRcQpw06IzqH+vqQ/vElJ1k96NZMUoGHh3k1AiuuIPY07Nq3tqWxVaFzOgzc6eqQgrRs+stdLtMpner1UXAZe8I6tKyk3Pud1241LuCQ7HUk
  0bBulR+RqM= -->
  <div id="a-popover-root" style="z-index:-1;position:absolute;"></div></body></html>
```

```
def get_url(search_term):

  template =
"https://www.amazon.in/s?k={}&rh=n%3A1389401031&ref=nb_sb_noss"

  search_term = search_term.replace(' ','+')

  return template.format(search_term)


url = get_url('laptops')

print(url)
```



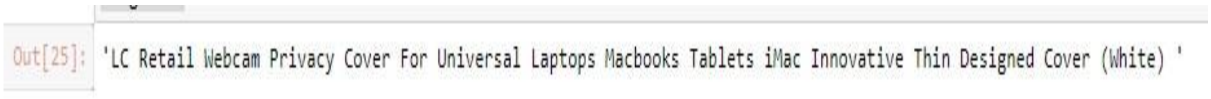
```
https://www.amazon.in/s?k=laptops&rh=n%3A1389401031&ref=nb_sb_noss
```

```
wd.get(url)

soup = BeautifulSoup(wd.page_source, 'html.parser')


result = soup.find_all('div',{'data-component-type':'s-search-result'})

len(result)

print(result[2])
```



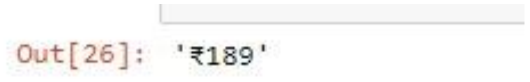
```
<div class="s-result-item s-asin sg-col-0-of-12 sg-col-16-of-20 sg-col s-widget-spacing-small sg-col-12-of-16" data-asin="B07B8
LX2TH" data-component-id="11" data-component-type="s-search-result" data-index="4" data-uuid="7529c173-6306-49a5-970f-807fb7788
d27"><div class="sg-col-inner"><div cel_widget_id="MAIN-SEARCH_RESULTS-4" class="s-widget-container s-spacing-small s-widget-co
ntainer-height-small celwidget slot=MAIN template=SEARCH_RESULTS widgetId=search-results_3" data-csa-c-id="37ztc4-dc01ep-ie5ckz
-r0zyl5"><div class="s-card-container s-overflow-hidden aok-relative s-include-content-margin s-latency-cf-section s-card-borde
r"><div class="a-section"><div class="sg-row"><div class="sg-col sg-col-4-of-12 sg-col-4-of-16 sg-col-4-of-20 s-list-col-left">
<div class="sg-col-inner"><div class="a-section a-spacing-none aok-relative s-image-overlay-grey s-list-status-badge-containe
r"></div><div class="s-product-image-container aok-relative s-image-overlay-grey s-text-center s-padding-left-small s-padding-r
ight-small s-flex-expand-height"><div class="aok-relative"><span class="rush-component" data-component-type="s-product-image"><
a class="a-link-normal s-no-outline" href="/LC-Retail-Universal-Macbooks-Innovative/dp/B07B8LX2TH/ref=sr_1_3?keywords=laptops&a
mp;qid=165312007&amp;s=electronics&amp;sr=1-3" target="_blank"><div class="a-section aok-relative s-image-fixed-height"><img a
lt="LC Retail Webcam Privacy Cover For Universal Laptops Macbooks Tablets iMac Innovative Thin Designed Cover (White)" class="s
-image" data-image-index="3" data-image-latency="s-product-image" data-image-load="" data-image-source-density="1" src="http
s://m.media-amazon.com/images/I/61tmmUxedsL._AC_UY218_.jpg" srcset="https://m.media-amazon.com/images/I/61tmmUxedsL._AC_UY218_.
jpg 1x, https://m.media-amazon.com/images/I/61tmmUxedsL._AC_UY327_FMwebp_QL65_.jpg 1.5x, https://m.media-amazon.com/images/I/61
tmmUxedsL._AC_UY436_FMwebp_QL65_.jpg 2x, https://m.media-amazon.com/images/I/61tmmUxedsL._AC_UY545_FMwebp_QL65_.jpg 2.5x, http
s://m.media-amazon.com/images/I/61tmmUxedsL._AC_UY654_FMwebp_QL65_.jpg 3x"/></div></a></span></div></div></div></div><div class
="sg-col sg-col-4-of-12 sg-col-8-of-16 sg-col-12-of-20 s-list-col-right"><div class="sg-col-inner"><div class="a-section a-spac
ing-small a-spacing-top-small"><div class="a-section a-spacing-none s-padding-right-small s-title-instructions-style"><h2 class
="a-size-mini a-spacing-none a-color-base s-line-clamp-2"><a class="a-link-normal s-underline-text s-underline-link-text s-link
-style a-text-normal" href="/LC-Retail-Universal-Macbooks-Innovative/dp/B07B8LX2TH/ref=sr_1_3?keywords=laptops&amp;qid=16531200
78&amp;s=electronics&amp;sr=1-3" target="_blank"><span class="a-size-medium a-color-base a-text-normal">LC Retail Webcam Privac
y Cover For Universal Laptops Macbooks Tablets iMac Innovative Thin Designed Cover (White)</span> </a> </h2></div><div class="a
-section a-spacing-none a-spacing-top-micro"><div class="a-row a-size-small"><span aria-label="3.8 out of 5 stars"><span class
="a-declarative" data-a-popover='{"closeButton":false,"closeButtonLabel":"","position":"triggerBottom","popoverLabel":"","ur
l":"/review/widgets/average-customer-review/popover/ref=acr_search__popover?ie=UTF8&amp;asin=B07B8LX2TH&amp;ref=acr_search__pop
over&amp;contextId=search"}' data-action="a-popover" data-csa-c-func-deps="aui-da-a-popover" data-csa-c-id="jug340-sxagkv-a8no4
r-uvxia" data-csa-c-type="widget"><a class="a-popover-trigger a-declarative" href="javascript:void(0)" role="button"><i class
="a-icon a-icon-star-small a-star-small-4 aok-align-bottom"><span class="a-icon-alt">3.8 out of 5 stars</span></i><i class="a-i
con a-icon-popover"></i></a></span> </span><span aria-label="27"><a class="a-link-normal s-underline-text s-underline-link-text
s-link-style" href="/LC-Retail-Universal-Macbooks-Innovative/dp/B07B8LX2TH/ref=sr_1_3?keywords=laptops&amp;qid=1653120078&amp;s
=electronics&amp;sr=1-3#customerReviews" target="_blank"><span class="a-size-base s-underline-text">27</span> </a> </span></div
></div><div class="sg-row"><div class="sg-col sg-col-4-of-12 sg-col-4-of-16 sg-col-4-of-20"><div class="sg-col-inner"><div clas
s="a-section a-spacing-none a-spacing-top-small s-price-instructions-style"><div class="a-row a-size-base a-color-base"><a clas
s="a-size-base a-link-normal s-underline-text s-underline-link-text s-link-style a-text-normal" href="/LC-Retail-Universal-Macb
ooks-Innovative/dp/B07B8LX2TH/ref=sr_1_3?keywords=laptops&amp;qid=1653120078&amp;s=electronics&amp;sr=1-3" target="_blank"><spa
n class="a-price" data-a-color="base" data-a-size="xl"><span class="a-offscreen">₹189</span><span aria-hidden="true"><span clas
s="a-price-symbol">₹</span><span class="a-price-whole">189</span></span></span> <span class="a-price a-text-price" data-a-color
="secondary" data-a-size="b" data-a-strike="true"><span class="a-offscreen">₹499</span><span aria-hidden="true">₹499</span></sp
an> </a> <span class="a-letter-space"></span><span>(62% off)</span><span class="a-letter-space"></span></div></div><div class
="a-section a-spacing-none a-spacing-top-micro"><div class="a-row a-size-base a-color-secondary s-align-children-center"><span
aria-label="Get it Friday, May 27 - Monday, May 30"><span class="a-color-base">Get it </span><span class="a-color-base a-text-b
old">Friday, May 27</span><span class="a-color-base"> - </span><span class="a-color-base a-text-bold">Monday, May 30</span></sp
an></div></div></div></div><div class="sg-col sg-col-4-of-12 sg-col-4-of-16 sg-col-8-of-20"><div class="sg-col-inner"></div></d
iv></div></div></div></div></div></div></div></div></div></div>
```

```
item = result[2]

atag = item.h2.a

atag.text
```



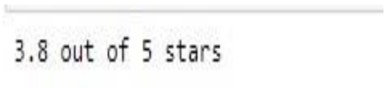
```
Out[25]: 'LC Retail Webcam Privacy Cover For Universal Laptops Macbooks Tablets iMac Innovative Thin Designed Cover (White) '
```

```
price_parent = item.find('span','a-price')

price_parent.find('span','a-offscreen').text
```



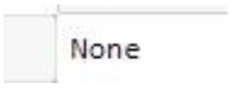
```
Out[26]: '₹189'
```

```
rating = item.i.text
print(rating)
```



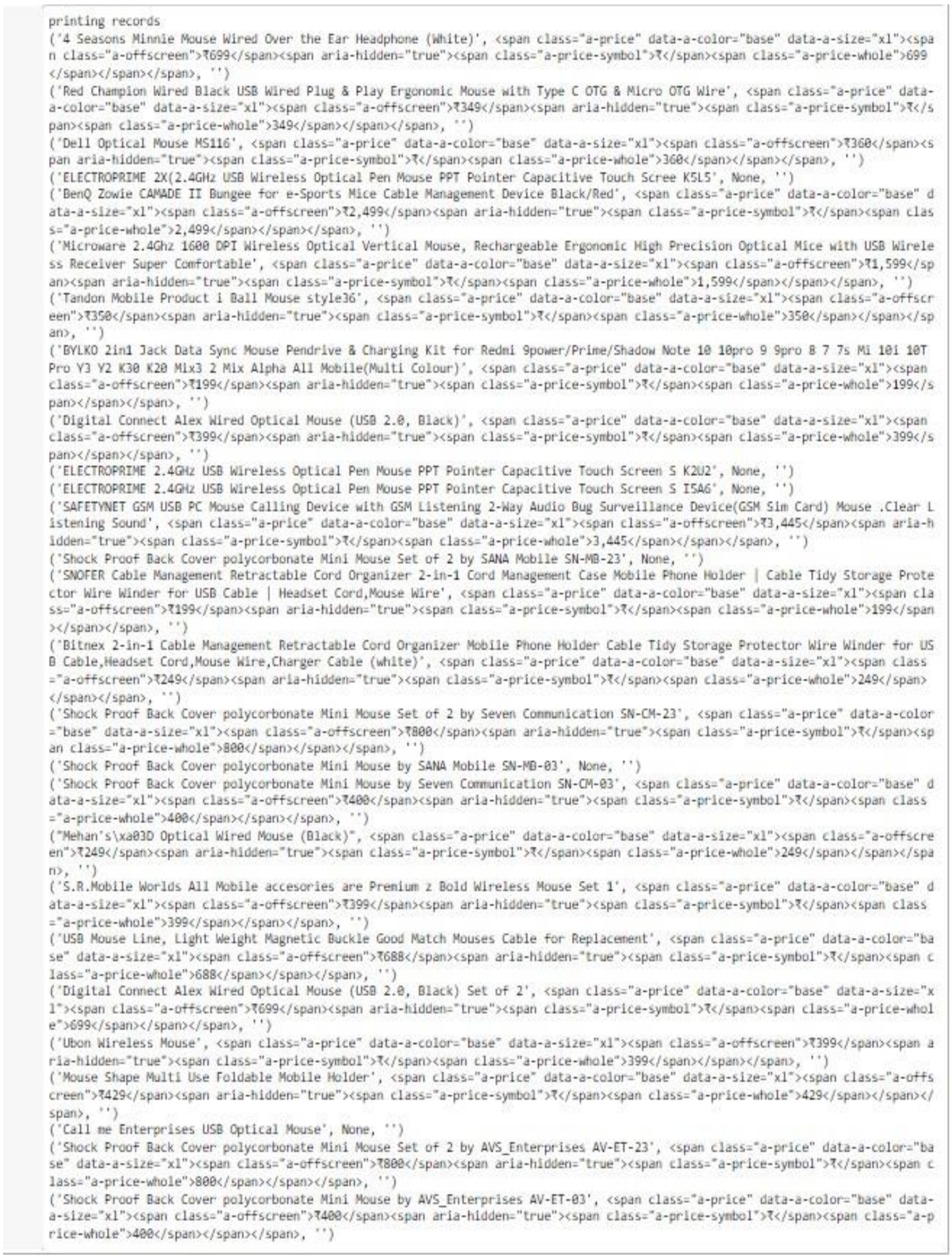
```
3.8 out of 5 stars
```

```
review_count = item.find('span', {'class':'a-size-base','dir':'auto'})

print(review_count)
```



```
None
```

```
def extract_record(item1):
  atag = item1.h2.a
  description = atag.text.strip()
  url = "https://www.amazon.in/" + atag.get('href')

  price_parent = item1.find('span','a-price')
  #price_parent.find('span','a-offscreen').text

  rating = ""
  result = (description, price_parent, rating)
  return result



url = get_url('mouse')
wd.get(url)
soup = BeautifulSoup(wd.page_source, 'html.parser')
records = []
results = soup.find_all('div',{'data-component-type':'s-search-result'})



for item in results:
  records.append(extract_record(item))
records[0]
print("printing records")
for x in range(len(records)):
  print(records[x])
```



```
printing records
('4 Seasons Minnie Mouse Wired Over the Ear Headphone (White)', <span class="a-price" data-a-color="base" data-a-size="xl"><spa
n class="a-offscreen">₹699</span><span aria-hidden="true"><span class="a-price-symbol">₹</span><span class="a-price-whole">699
</span></span></span>, '')
('Red Champion Wired Black USB Wired Plug & Play Ergonomic Mouse with Type C OTG & Micro OTG Wire', <span class="a-price" data-
a-color="base" data-a-size="xl"><span class="a-offscreen">₹349</span><span aria-hidden="true"><span class="a-price-symbol">₹</s
pan><span class="a-price-whole">349</span></span></span>, '')
('Dell Optical Mouse MS116', <span class="a-price" data-a-color="base" data-a-size="xl"><span class="a-offscreen">₹360</span><s
pan aria-hidden="true"><span class="a-price-symbol">₹</span><span class="a-price-whole">360</span></span></span>, '')
('ELECTROPRIME 2X(2.4GHz USB Wireless Optical Pen Mouse PPT Pointer Capacitive Touch Scree K5L5', None, '')
('BenQ Zowie CAMADE II Bungee for e-Sports Mice Cable Management Device Black/Red', <span class="a-price" data-a-color="base" d
ata-a-size="xl"><span class="a-offscreen">₹2,499</span><span aria-hidden="true"><span class="a-price-symbol">₹</span><span clas
s="a-price-whole">2,499</span></span></span>, '')
('Microware 2.4Ghz 1600 DPI Wireless Optical Vertical Mouse, Rechargeable Ergonomic High Precision Optical Mice with USB Wirele
ss Receiver Super Comfortable', <span class="a-price" data-a-color="base" data-a-size="xl"><span class="a-offscreen">₹1,599</sp
an><span aria-hidden="true"><span class="a-price-symbol">₹</span><span class="a-price-whole">1,599</span></span></span>, '')
('Tandon Mobile Product i Ball Mouse style36', <span class="a-price" data-a-color="base" data-a-size="xl"><span class="a-offscr
een">₹350</span><span aria-hidden="true"><span class="a-price-symbol">₹</span><span class="a-price-whole">350</span></span></sp
an>, '')
('BYLKO 2in1 Jack Data Sync Mouse Pendrive & Charging Kit for Redmi 9power/Prime/Shadow Note 10 10pro 9 9pro 8 7 7s Mi 10i 10T
Pro Y3 Y2 K30 K20 Mix3 2 Mix Alpha All Mobile(Multi Colour)', <span class="a-price" data-a-color="base" data-a-size="xl"><span
class="a-offscreen">₹199</span><span aria-hidden="true"><span class="a-price-symbol">₹</span><span class="a-price-whole">199</s
pan></span></span>, '')
('Digital Connect Alex Wired Optical Mouse (USB 2.0, Black)', <span class="a-price" data-a-color="base" data-a-size="xl"><span
class="a-offscreen">₹399</span><span aria-hidden="true"><span class="a-price-symbol">₹</span><span class="a-price-whole">399</s
pan></span></span>, '')
('ELECTROPRIME 2.4GHz USB Wireless Optical Pen Mouse PPT Pointer Capacitive Touch Screen S K2U2', None, '')
('ELECTROPRIME 2.4GHz USB Wireless Optical Pen Mouse PPT Pointer Capacitive Touch Screen S I5A6', None, '')
('SAFETYNET GSM USB PC Mouse Calling Device with GSM Listening 2-Way Audio Bug Surveillance Device(GSM Sim Card) Mouse .Clear L
istening Sound', <span class="a-price" data-a-color="base" data-a-size="xl"><span class="a-offscreen">₹3,445</span><span aria-h
idden="true"><span class="a-price-symbol">₹</span><span class="a-price-whole">3,445</span></span></span>, '')
('Shock Proof Back Cover polycorbonate Mini Mouse Set of 2 by SANA Mobile SN-MB-23', None, '')
('SNOFER Cable Management Retractable Cord Organizer 2-in-1 Cord Management Case Mobile Phone Holder | Cable Tidy Storage Prote
ctor Wire Winder for USB Cable | Headset Cord,Mouse Wire', <span class="a-price" data-a-color="base" data-a-size="xl"><span cla
ss="a-offscreen">₹199</span><span aria-hidden="true"><span class="a-price-symbol">₹</span><span class="a-price-whole">199</span
></span></span>, '')
('Bitnex 2-in-1 Cable Management Retractable Cord Organizer Mobile Phone Holder Cable Tidy Storage Protector Wire Winder for US
B Cable,Headset Cord,Mouse Wire,Charger Cable (white)', <span class="a-price" data-a-color="base" data-a-size="xl"><span class
="a-offscreen">₹249</span><span aria-hidden="true"><span class="a-price-symbol">₹</span><span class="a-price-whole">249</span>
</span></span>, '')
('Shock Proof Back Cover polycorbonate Mini Mouse Set of 2 by Seven Communication SN-CM-23', <span class="a-price" data-a-color
="base" data-a-size="xl"><span class="a-offscreen">₹800</span><span aria-hidden="true"><span class="a-price-symbol">₹</span><sp
an class="a-price-whole">800</span></span></span>, '')
('Shock Proof Back Cover polycorbonate Mini Mouse by SANA Mobile SN-MB-03', None, '')
('Shock Proof Back Cover polycorbonate Mini Mouse by Seven Communication SN-CM-03', <span class="a-price" data-a-color="base" d
ata-a-size="xl"><span class="a-offscreen">₹400</span><span aria-hidden="true"><span class="a-price-symbol">₹</span><span class
="a-price-whole">400</span></span></span>, '')
("Mehan's\xa03D Optical Wired Mouse (Black)", <span class="a-price" data-a-color="base" data-a-size="xl"><span class="a-offscre
en">₹249</span><span aria-hidden="true"><span class="a-price-symbol">₹</span><span class="a-price-whole">249</span></span></spa
n>, '')
('S.R.Mobile Worlds All Mobile accesories are Premium z Bold Wireless Mouse Set 1', <span class="a-price" data-a-color="base" d
ata-a-size="xl"><span class="a-offscreen">₹399</span><span aria-hidden="true"><span class="a-price-symbol">₹</span><span class
="a-price-whole">399</span></span></span>, '')
('USB Mouse Line, Light Weight Magnetic Buckle Good Match Mouses Cable for Replacement', <span class="a-price" data-a-color="ba
se" data-a-size="xl"><span class="a-offscreen">₹688</span><span aria-hidden="true"><span class="a-price-symbol">₹</span><span c
lass="a-price-whole">688</span></span></span>, '')
('Digital Connect Alex Wired Optical Mouse (USB 2.0, Black) Set of 2', <span class="a-price" data-a-color="base" data-a-size="x
l"><span class="a-offscreen">₹699</span><span aria-hidden="true"><span class="a-price-symbol">₹</span><span class="a-price-whol
e">699</span></span></span>, '')
('Ubon Wireless Mouse', <span class="a-price" data-a-color="base" data-a-size="xl"><span class="a-offscreen">₹399</span><span a
ria-hidden="true"><span class="a-price-symbol">₹</span><span class="a-price-whole">399</span></span></span>, '')
('Mouse Shape Multi Use Foldable Mobile Holder', <span class="a-price" data-a-color="base" data-a-size="xl"><span class="a-offs
creen">₹429</span><span aria-hidden="true"><span class="a-price-symbol">₹</span><span class="a-price-whole">429</span></span></
span>, '')
('Call me Enterprises USB Optical Mouse', None, '')
('Shock Proof Back Cover polycorbonate Mini Mouse Set of 2 by AVS_Enterprises AV-ET-23', <span class="a-price" data-a-color="ba
se" data-a-size="xl"><span class="a-offscreen">₹800</span><span aria-hidden="true"><span class="a-price-symbol">₹</span><span c
lass="a-price-whole">800</span></span></span>, '')
('Shock Proof Back Cover polycorbonate Mini Mouse by AVS_Enterprises AV-ET-03', <span class="a-price" data-a-color="base" data-
a-size="xl"><span class="a-offscreen">₹400</span><span aria-hidden="true"><span class="a-price-symbol">₹</span><span class="a-p
rice-whole">400</span></span></span>, '')
```

Practical 2

AIM :- Page Rank for link analysis using python

Create a small set of pages namely page1, page2, page3 and page4 apply random walk on the same.

THEORY : -

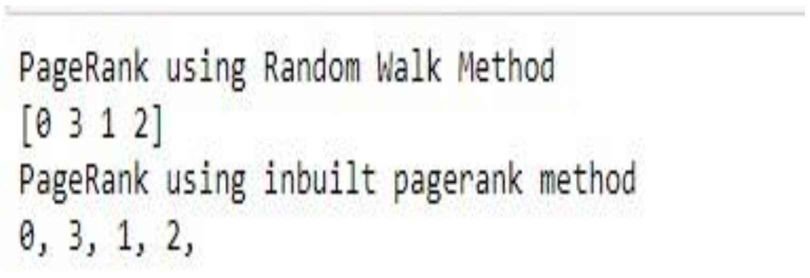
PageRank works by counting the number and quality of links to a page to determine a rough estimate of how important the website is. The underlying assumption is that more important websites are likely to receive more links from other websites.

CODE / OUTPUT : -

```
"""Random_Walk_Page_Rank"""
"Practical 3"

import networkx as nx
import random
import numpy as np

# Add directed edges in graph
def add_edges(g, pr):
      for each in g.nodes():
            for each1 in g.nodes():
                  if (each != each1):
                        ra = random.random()
                        if (ra < pr):
                              g.add_edge(each, each1)
                        else:
                              continue
      return g
```

```
# Sort the nodes
def nodes_sorted(g, points):
      t = np.array(points)
      t = np.argsort(-t)
      return t

# Distribute points randomly in a graph
def random_Walk(g):
      rwp = [0 for i in range(g.number_of_nodes())]
      nodes = list(g.nodes())
      r = random.choice(nodes)
      rwp[r] += 1
      neigh = list(g.out_edges(r))
      z = 0

      while (z != 10000):
            if (len(neigh) == 0):
                  focus = random.choice(nodes)
            else:
                  r1 = random.choice(neigh)
                  focus = r1[1]
            rwp[focus] += 1
            neigh = list(g.out_edges(focus))
            z += 1
      return rwp

g = nx.DiGraph()
N = 4
g.add_nodes_from(range(N))

# 2. Add directed edges in graph
```

```
g = add_edges(g, 0.4)

# 3. perform a random walk
points = random_Walk(g)

# 4. Get nodes rank according to their random walk points
sorted_by_points = nodes_sorted(g, points)
print("PageRank using Random Walk Method")
print(sorted_by_points)

# p_dict is dictionary of tuples
p_dict = nx.pagerank(g)
p_sort = sorted(p_dict.items(), key=lambda x: x[1], reverse=True)

print("PageRank using inbuilt pagerank method")
for i in p_sort:
      print(i[0], end=", ")
```



```
PageRank using Random Walk Method
[0 3 1 2]
PageRank using inbuilt pagerank method
0, 3, 1, 2,
```

Practical 3

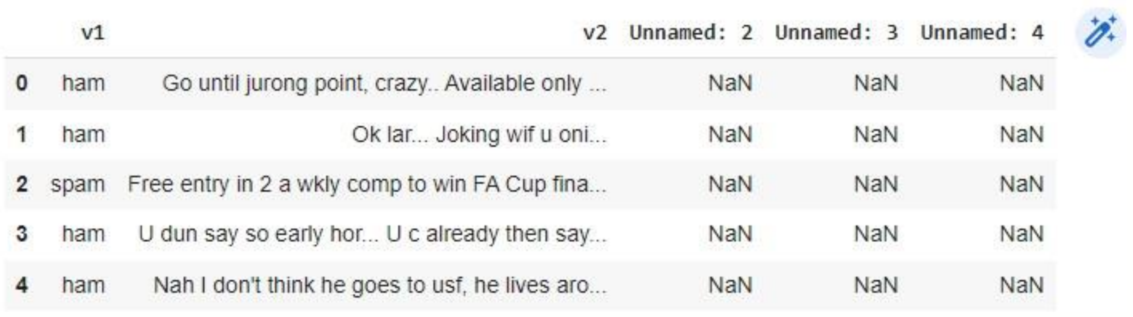
AIM :- Perform Spam Classifier

THEORY : -

A spam filter/classifier is a program used to detect unsolicited, unwanted and virus-infected emails and prevent those messages from getting to a user's inbox. Like other types of filtering programs, a spam filter looks for specific criteria on which to base its judgments.

CODE / OUTPUT : -

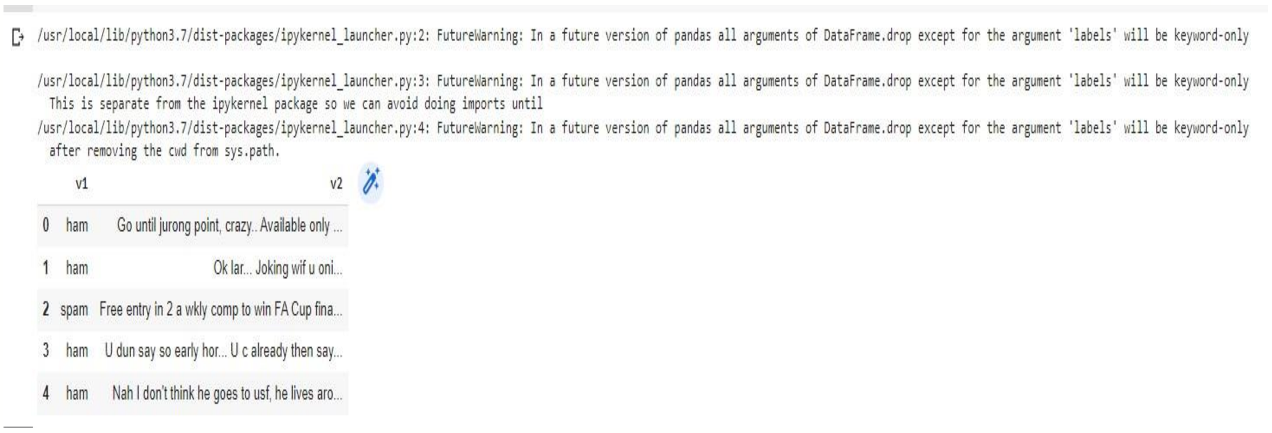
```
import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

from google.colab import drive

from sklearn import feature_extraction, naive_bayes, metrics,
model_selection

from sklearn.ensemble import RandomForestClassifier

from sklearn.model_selection import train_test_split

from sklearn.metrics import precision_recall_fscore_support as score

drive.mount('/content/drive')


dataset =pd.read_csv("/content/drive/MyDrive/spam.csv", encoding='latin-
1')


dataset.head()
```



	v1	v2	Unnamed: 2	Unnamed: 3	Unnamed: 4
0	ham	Go until jurong point, crazy.. Available only ...	NaN	NaN	NaN
1	ham	Ok lar... Joking wif u oni...	NaN	NaN	NaN
2	spam	Free entry in 2 a wkly comp to win FA Cup fina...	NaN	NaN	NaN
3	ham	U dun say so early hor... U c already then say...	NaN	NaN	NaN
4	ham	Nah I don't think he goes to usf, he lives aro...	NaN	NaN	NaN

```
#removing unnamed columns
```

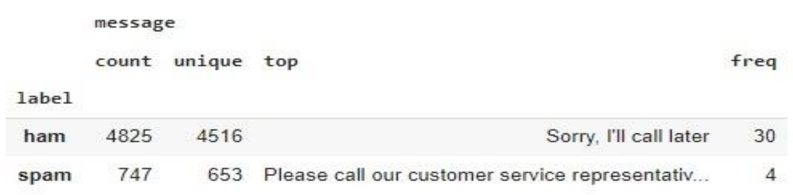
```
dataset = dataset.drop('Unnamed: 2', 1)
dataset = dataset.drop('Unnamed: 3', 1)
dataset = dataset.drop('Unnamed: 4', 1)
dataset.head()
```



```
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:2: FutureWarning: In a future version of pandas all arguments of DataFrame.drop except for the argument 'labels' will be keyword-only

/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:3: FutureWarning: In a future version of pandas all arguments of DataFrame.drop except for the argument 'labels' will be keyword-only
  This is separate from the ipykernel package so we can avoid doing imports until
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:4: FutureWarning: In a future version of pandas all arguments of DataFrame.drop except for the argument 'labels' will be keyword-only
  after removing the cwd from sys.path.
```

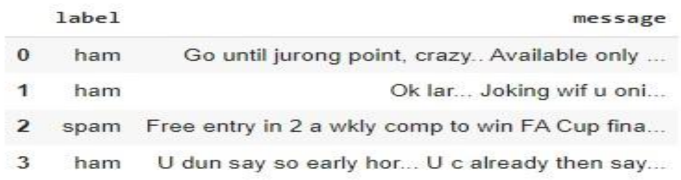
	v1	v2
0	ham	Go until jurong point, crazy.. Available only ...
1	ham	Ok lar... Joking wif u oni...
2	spam	Free entry in 2 a wkly comp to win FA Cup fina...
3	ham	U dun say so early hor... U c already then say...
4	ham	Nah I don't think he goes to usf, he lives aro...

```
dataset = dataset.rename(columns = {'v1':'label','v2':'message'})

dataset.groupby('label').describe()
```



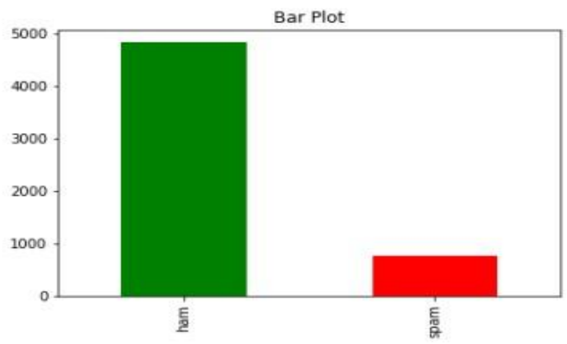
	message			
	count	unique	top	freq
label				
ham	4825	4516	Sorry, I'll call later	30
spam	747	653	Please call our customer service representativ...	4

```
dataset.head(4)
```



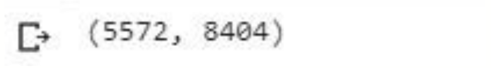
	label	message
0	ham	Go until jurong point, crazy.. Available only ...
1	ham	Ok lar... Joking wif u oni...
2	spam	Free entry in 2 a wkly comp to win FA Cup fina...
3	ham	U dun say so early hor... U c already then say...

```
count_Class=pd.value_counts(dataset["label"], sort= True)
count_Class.plot(kind = 'bar',color = ["green","red"])
plt.title('Bar Plot')
plt.show();
```



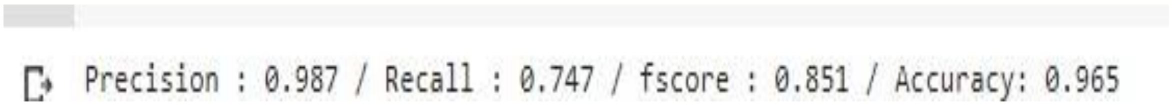
```
f = feature_extraction.text.CountVectorizer(stop_words = 'english')
```

```
X = f.fit_transform(dataset["message"])
np.shape(X)
```



```
(5572, 8404)
```

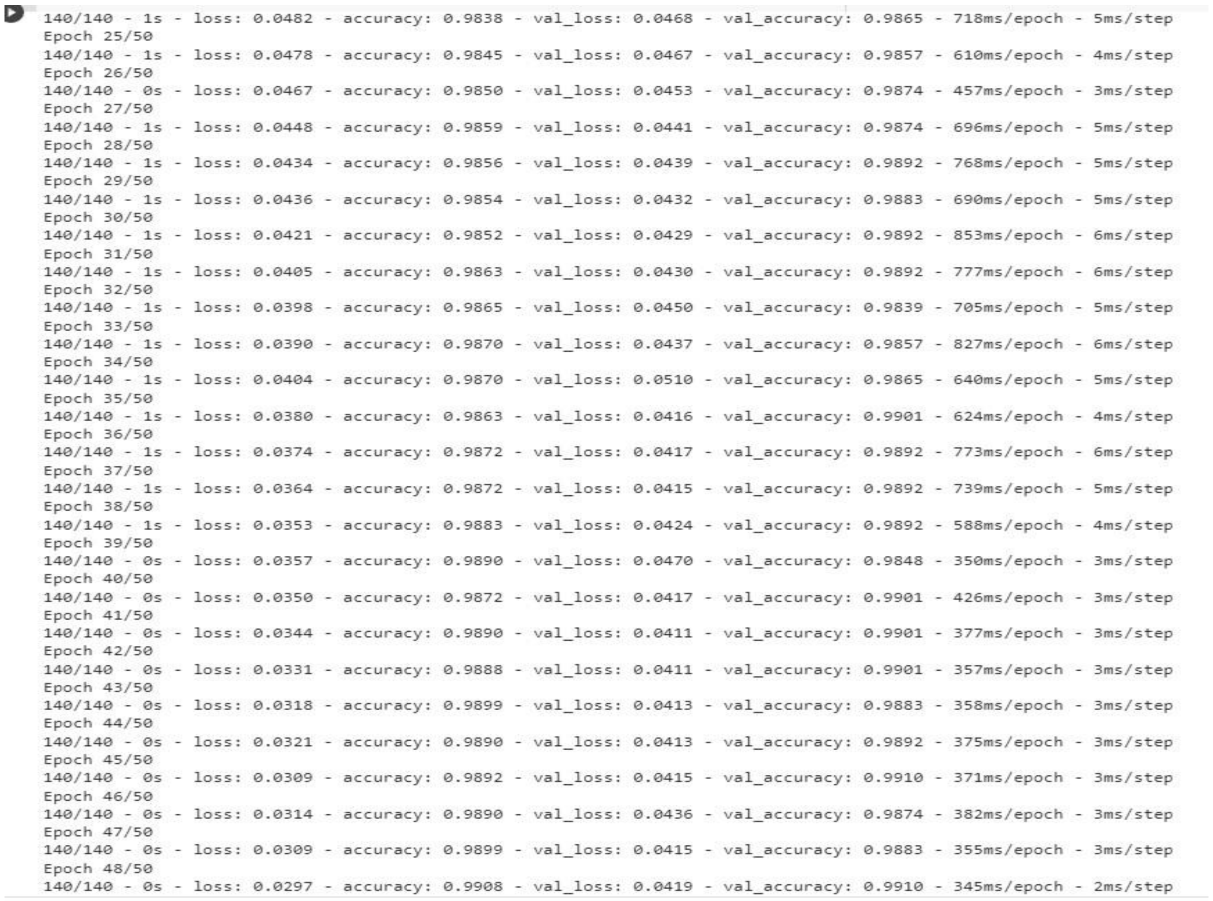
```
# Classifying spam and not spam msgs as 1 and 0

dataset["label"]=dataset["label"].map({'spam':1,'ham':0})
X_train, X_test, y_train, y_test = model_selection.train_test_split(X,
dataset['label'], test_size=0.70, random_state=42)

list_alpha = np.arange(1/100000, 20, 0.11)
score_train = np.zeros(len(list_alpha))
score_test = np.zeros(len(list_alpha))
recall_test = np.zeros(len(list_alpha))
precision_test= np.zeros(len(list_alpha))
count = 0
for alpha in list_alpha:
    bayes = naive_bayes.MultinomialNB(alpha=alpha)
    bayes.fit(X_train, y_train)
    score_train[count] = bayes.score(X_train, y_train)
    score_test[count]= bayes.score(X_test, y_test)
    recall_test[count] = metrics.recall_score(y_test,
bayes.predict(X_test))
    precision_test[count] = metrics.precision_score(y_test,
bayes.predict(X_test))
    count = count + 1

matrix = np.matrix(np.c_[list_alpha, score_train, score_test, recall_test,
precision_test])
models = pd.DataFrame(data = matrix, columns =
             ['alpha', 'Train Accuracy', 'Test Accuracy', 'Test Recall',
'Test Precision'])
models.head(n=10)

best_index = models['Test Precision'].idxmax()
models.iloc[best_index, :]

rf = RandomForestClassifier(n_estimators=100,max_depth=None,n_jobs=-1)
rf_model = rf.fit(X_train,y_train)

y_pred=rf_model.predict(X_test)
precision,recall,fscore,support =score(y_test,y_pred,pos_label=1, average
='binary')
print('Precision : {} / Recall : {} / fscore : {} / Accuracy:
{}'.format(round(precision,3),round(recall,3),round(fscore,3),round((y_pre
d==y_test).sum()/len(y_test),3)))
```



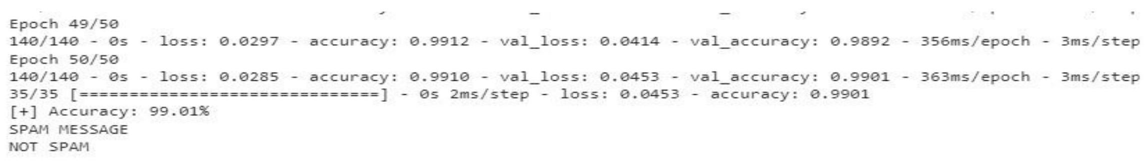
```
Precision : 0.987 / Recall : 0.747 / fscore : 0.851 / Accuracy: 0.965
```

```
import tensorflow as tf
from keras.preprocessing.text import Tokenizer
from keras.layers import Embedding, LSTM, Dropout, Dense
from keras.models import Sequential
from tensorflow.keras.utils import to_categorical
#from keras.utils import to_categorical
from keras.preprocessing.sequence import pad_sequences
import tensorflow as tf

vocab_size = 400
oov_tok = "<OOV>"
max_length = 250
embedding_dim = 16
encode = ({'ham': 0, 'spam': 1} )
#new dataset with replaced values
dataset = dataset.replace(encode)

X = dataset['message']
Y = dataset['label']
tokenizer = Tokenizer(num_words=vocab_size, oov_token=oov_tok)
tokenizer.fit_on_texts(X)
# convert to sequence of integers
X = tokenizer.texts_to_sequences(X)

X = np.array(X)
y = np.array(Y)

X = pad_sequences(X, maxlen=max_length)
```

```
model = tf.keras.Sequential([
    tf.keras.layers.Embedding(vocab_size, embedding_dim,
input_length=max_length),
    tf.keras.layers.GlobalAveragePooling1D(),
    tf.keras.layers.Dense(24, activation='relu'),
    tf.keras.layers.Dense(1, activation='sigmoid')
])
model.compile(loss='binary_crossentropy',optimizer='adam',metrics=['accura
cy'])
model.summary()


num_epochs = 50
X_train, X_test, y_train, y_test = train_test_split(X, Y, test_size=.20,
random_state=7)
history = model.fit(X_train, y_train, epochs=num_epochs,
validation_data=(X_test,y_test), verbose=2)


results = model.evaluate(X_test, y_test)
loss = results[0]
accuracy = results[1]


print(f"[+] Accuracy: {accuracy*100:.2f}%")


from keras.preprocessing import sequence


#Defining the function
def get_predictions(txts):
    txts = tokenizer.texts_to_sequences(txts)
    txts = sequence.pad_sequences(txts, maxlen=max_length)
    preds = model.predict(txts)
    if(preds[0] > 0.5):
        print("SPAM MESSAGE")
    else:

        print('NOT SPAM')


txts=["You have won a free ticket to las vegas. Contact now"]


get_predictions(txts)


txts=["Hey there call me asap!!"]

get_predictions(txts)
```



```
140/140 - 1s - loss: 0.0482 - accuracy: 0.9838 - val_loss: 0.0468 - val_accuracy: 0.9865 - 718ms/epoch - 5ms/step
Epoch 25/50
140/140 - 1s - loss: 0.0478 - accuracy: 0.9845 - val_loss: 0.0467 - val_accuracy: 0.9857 - 610ms/epoch - 4ms/step
Epoch 26/50
140/140 - 0s - loss: 0.0467 - accuracy: 0.9850 - val_loss: 0.0453 - val_accuracy: 0.9874 - 457ms/epoch - 3ms/step
Epoch 27/50
140/140 - 1s - loss: 0.0448 - accuracy: 0.9859 - val_loss: 0.0441 - val_accuracy: 0.9874 - 696ms/epoch - 5ms/step
Epoch 28/50
140/140 - 1s - loss: 0.0434 - accuracy: 0.9856 - val_loss: 0.0439 - val_accuracy: 0.9892 - 768ms/epoch - 5ms/step
Epoch 29/50
140/140 - 1s - loss: 0.0436 - accuracy: 0.9854 - val_loss: 0.0432 - val_accuracy: 0.9883 - 690ms/epoch - 5ms/step
Epoch 30/50
140/140 - 1s - loss: 0.0421 - accuracy: 0.9852 - val_loss: 0.0429 - val_accuracy: 0.9892 - 853ms/epoch - 6ms/step
Epoch 31/50
140/140 - 1s - loss: 0.0405 - accuracy: 0.9863 - val_loss: 0.0430 - val_accuracy: 0.9892 - 777ms/epoch - 6ms/step
Epoch 32/50
140/140 - 1s - loss: 0.0398 - accuracy: 0.9865 - val_loss: 0.0450 - val_accuracy: 0.9839 - 705ms/epoch - 5ms/step
Epoch 33/50
140/140 - 1s - loss: 0.0390 - accuracy: 0.9870 - val_loss: 0.0437 - val_accuracy: 0.9857 - 827ms/epoch - 6ms/step
Epoch 34/50
140/140 - 1s - loss: 0.0404 - accuracy: 0.9870 - val_loss: 0.0510 - val_accuracy: 0.9865 - 640ms/epoch - 5ms/step
Epoch 35/50
140/140 - 1s - loss: 0.0380 - accuracy: 0.9863 - val_loss: 0.0416 - val_accuracy: 0.9901 - 624ms/epoch - 4ms/step
Epoch 36/50
140/140 - 1s - loss: 0.0374 - accuracy: 0.9872 - val_loss: 0.0417 - val_accuracy: 0.9892 - 773ms/epoch - 6ms/step
Epoch 37/50
140/140 - 1s - loss: 0.0364 - accuracy: 0.9872 - val_loss: 0.0415 - val_accuracy: 0.9892 - 739ms/epoch - 5ms/step
Epoch 38/50
140/140 - 1s - loss: 0.0353 - accuracy: 0.9883 - val_loss: 0.0424 - val_accuracy: 0.9892 - 588ms/epoch - 4ms/step
Epoch 39/50
140/140 - 0s - loss: 0.0357 - accuracy: 0.9890 - val_loss: 0.0470 - val_accuracy: 0.9848 - 350ms/epoch - 3ms/step
Epoch 40/50
140/140 - 0s - loss: 0.0350 - accuracy: 0.9872 - val_loss: 0.0417 - val_accuracy: 0.9901 - 426ms/epoch - 3ms/step
Epoch 41/50
140/140 - 0s - loss: 0.0344 - accuracy: 0.9890 - val_loss: 0.0411 - val_accuracy: 0.9901 - 377ms/epoch - 3ms/step
Epoch 42/50
140/140 - 0s - loss: 0.0331 - accuracy: 0.9888 - val_loss: 0.0411 - val_accuracy: 0.9901 - 357ms/epoch - 3ms/step
Epoch 43/50
140/140 - 0s - loss: 0.0318 - accuracy: 0.9899 - val_loss: 0.0413 - val_accuracy: 0.9883 - 358ms/epoch - 3ms/step
Epoch 44/50
140/140 - 0s - loss: 0.0321 - accuracy: 0.9890 - val_loss: 0.0413 - val_accuracy: 0.9892 - 375ms/epoch - 3ms/step
Epoch 45/50
140/140 - 0s - loss: 0.0309 - accuracy: 0.9892 - val_loss: 0.0415 - val_accuracy: 0.9910 - 371ms/epoch - 3ms/step
Epoch 46/50
140/140 - 0s - loss: 0.0314 - accuracy: 0.9890 - val_loss: 0.0436 - val_accuracy: 0.9874 - 382ms/epoch - 3ms/step
Epoch 47/50
140/140 - 0s - loss: 0.0309 - accuracy: 0.9899 - val_loss: 0.0415 - val_accuracy: 0.9883 - 355ms/epoch - 3ms/step
Epoch 48/50
140/140 - 0s - loss: 0.0297 - accuracy: 0.9908 - val_loss: 0.0419 - val_accuracy: 0.9910 - 345ms/epoch - 2ms/step
```



```
Epoch 49/50
140/140 - 0s - loss: 0.0297 - accuracy: 0.9912 - val_loss: 0.0414 - val_accuracy: 0.9892 - 356ms/epoch - 3ms/step
Epoch 50/50
140/140 - 0s - loss: 0.0285 - accuracy: 0.9910 - val_loss: 0.0453 - val_accuracy: 0.9901 - 363ms/epoch - 3ms/step
35/35 [==============================] - 0s 2ms/step - loss: 0.0453 - accuracy: 0.9901
[+] Accuracy: 99.01%
SPAM MESSAGE
NOT SPAM
```

Practical 4

AIM :- Demonstrate Text Mining and Webpage Pre-processing using meta information from the web pages (Local/Online).

THEORY : -

Web mining is the application of data mining techniques to discover patterns from the World Wide Web. It uses automated methods to extract both structured and unstructured data from web pages, server logs and link structures. There are three main sub-categories of web mining.

Preprocessing stage helps to clean the records and determine the interesting user patterns and session creation. Data preprocessing is an important job of Web usage mining application. So, data must be processed before applying data mining methods to determine user access patterns from web log.

A meta description is an HTML element that provides a brief summary of a web page. A page's meta description tag is displayed as part of the search snippet in a search engine results page (SERP) and is meant to give the user an idea of the content that exists within the page and how it relates to their search query.

CODE / OUTPUT : -

```
"""Pract5.ipynb"""

# Commented out IPython magic to ensure Python compatibility.

# Imports


import requests

import numpy as np

import pandas as pd

from bs4 import BeautifulSoup

import matplotlib.pyplot as plt


# IMDB's homepage

imdb_url = 'https://www.imdb.com'
```

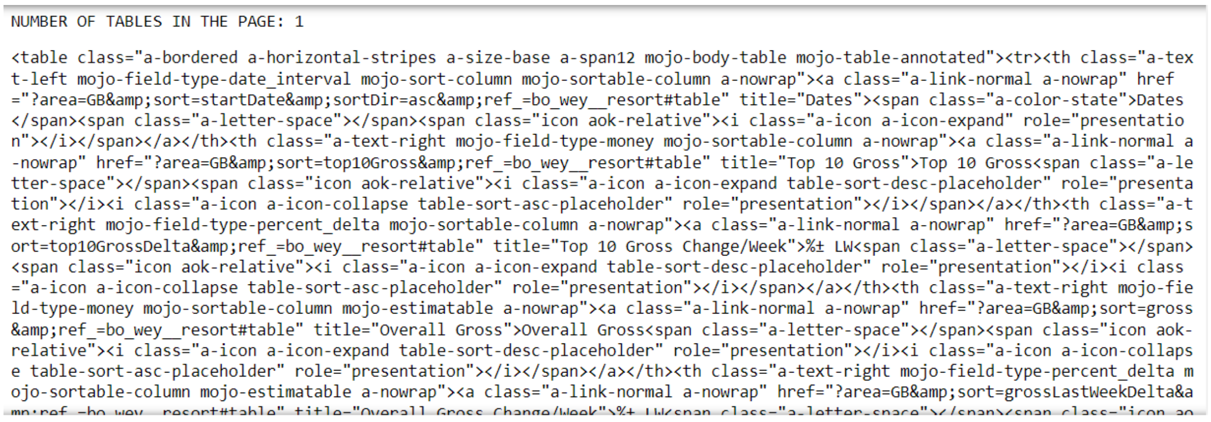
```
# Use requests to retrieve data from a given URL

imdb_response = requests.get(imdb_url)


# Parse the whole HTML page using BeautifulSoup

imdb_soup = BeautifulSoup(imdb_response.text, 'html.parser')


# Title of the parsed page

imdb_soup.title.text
```

```
Out[3]: 'IMDb: Ratings, Reviews, and Where to Watch the Best Movies & TV Shows'
```

```
# Find all links

links = [link.get('href') for link in imdb_soup.find_all('a')]


# Add homepage and keep the unique links

unique_links = []

for link in links:

    if not link in unique_links:

        unique_links.append(imdb_url + link)


# Box Office Mojo - UK Weekend box office

boxofficemojo_url =
'https://www.boxofficemojo.com/intl/uk/?yr=2019&wk=33&currency=local'


# Use requests to retrieve data from a given URL

bom_response = requests.get(boxofficemojo_url)


# Parse the whole HTML page using BeautifulSoup

bom_soup = BeautifulSoup(bom_response.text, 'html.parser')


print(f"NUMBER OF TABLES IN THE PAGE: {len(bom_soup.find_all('table'))}")
```

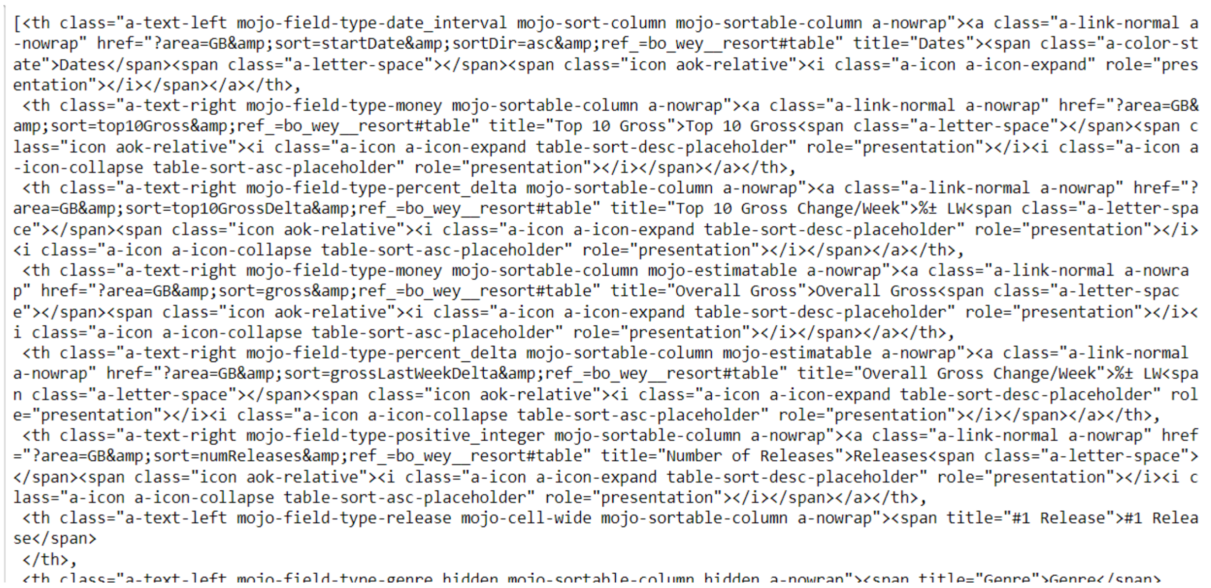
```
table = bom_soup.find_all('table')[0]
```

```
table
```



```
NUMBER OF TABLES IN THE PAGE: 1

<table class="a-bordered a-horizontal-stripes a-size-base a-span12 mojo-body-table mojo-table-annotated"><tr><th class="a-tex
t-left mojo-field-type-date_interval mojo-sort-column mojo-sortable-column a-nowrap"><a class="a-link-normal a-nowrap" href
="?area=GB&amp;sort=startDate&amp;sortDir=asc&amp;ref_=bo_wey__resort#table" title="Dates"><span class="a-color-state">Dates
</span><span class="a-letter-space"></span><span class="icon aok-relative"><i class="a-icon a-icon-expand" role="presentatio
n"></i></span></a></th><th class="a-text-right mojo-field-type-money mojo-sortable-column a-nowrap"><a class="a-link-normal a
-nowrap" href="?area=GB&amp;sort=top10Gross&amp;ref_=bo_wey__resort#table" title="Top 10 Gross">Top 10 Gross<span class="a-le
tter-space"></span><span class="icon aok-relative"><i class="a-icon a-icon-expand table-sort-desc-placeholder" role="presenta
tion"></i><i class="a-icon a-icon-collapse table-sort-asc-placeholder" role="presentation"></i></span></a></th><th class="a-t
ext-right mojo-field-type-percent_delta mojo-sortable-column a-nowrap"><a class="a-link-normal a-nowrap" href="?area=GB&amp;s
ort=top10GrossDelta&amp;ref_=bo_wey__resort#table" title="Top 10 Gross Change/Week">%± LW<span class="a-letter-space"></span>
<span class="icon aok-relative"><i class="a-icon a-icon-expand table-sort-desc-placeholder" role="presentation"></i><i class
="a-icon a-icon-collapse table-sort-asc-placeholder" role="presentation"></i></span></a></th><th class="a-text-right mojo-fie
ld-type-money mojo-sortable-column mojo-estimatable a-nowrap"><a class="a-link-normal a-nowrap" href="?area=GB&amp;sort=gross
&amp;ref_=bo_wey__resort#table" title="Overall Gross">Overall Gross<span class="a-letter-space"></span><span class="icon aok-
relative"><i class="a-icon a-icon-expand table-sort-desc-placeholder" role="presentation"></i><i class="a-icon a-icon-collaps
e table-sort-asc-placeholder" role="presentation"></i></span></a></th><th class="a-text-right mojo-field-type-percent_delta m
ojo-sortable-column mojo-estimatable a-nowrap"><a class="a-link-normal a-nowrap" href="?area=GB&amp;sort=grossLastWeekDelta&a
mp;ref_=bo_wey__resort#table" title="Overall Gross Change/Week">%± LW<span class="a-letter-space"></span><span class="icon ao
```

```
table.find_all('tr')[0].contents
```

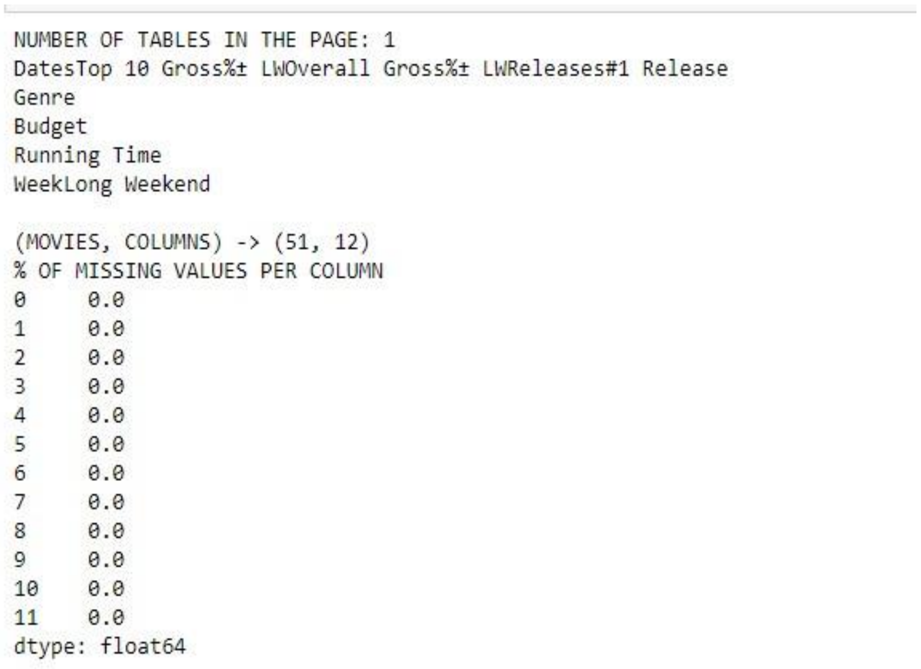


```
[<th class="a-text-left mojo-field-type-date_interval mojo-sort-column mojo-sortable-column a-nowrap"><a class="a-link-normal a
-nowrap" href="?area=GB&amp;sort=startDate&amp;sortDir=asc&amp;ref_=bo_wey__resort#table" title="Dates"><span class="a-color-st
ate">Dates</span><span class="a-letter-space"></span><span class="icon aok-relative"><i class="a-icon a-icon-expand" role="pres
entation"></i></span></a></th>,
 <th class="a-text-right mojo-field-type-money mojo-sortable-column a-nowrap"><a class="a-link-normal a-nowrap" href="?area=GB&
amp;sort=top10Gross&amp;ref_=bo_wey__resort#table" title="Top 10 Gross">Top 10 Gross<span class="a-letter-space"></span><span c
lass="icon aok-relative"><i class="a-icon a-icon-expand table-sort-desc-placeholder" role="presentation"></i><i class="a-icon a
-icon-collapse table-sort-asc-placeholder" role="presentation"></i></span></a></th>,
 <th class="a-text-right mojo-field-type-percent_delta mojo-sortable-column a-nowrap"><a class="a-link-normal a-nowrap" href="?
area=GB&amp;sort=top10GrossDelta&amp;ref_=bo_wey__resort#table" title="Top 10 Gross Change/Week">%± LW<span class="a-letter-spa
ce"></span><span class="icon aok-relative"><i class="a-icon a-icon-expand table-sort-desc-placeholder" role="presentation"></i>
<i class="a-icon a-icon-collapse table-sort-asc-placeholder" role="presentation"></i></span></a></th>,
 <th class="a-text-right mojo-field-type-money mojo-sortable-column mojo-estimatable a-nowrap"><a class="a-link-normal a-nowra
p" href="?area=GB&amp;sort=gross&amp;ref_=bo_wey__resort#table" title="Overall Gross">Overall Gross<span class="a-letter-spac
e"></span><span class="icon aok-relative"><i class="a-icon a-icon-expand table-sort-desc-placeholder" role="presentation"></i><
i class="a-icon a-icon-collapse table-sort-asc-placeholder" role="presentation"></i></span></a></th>,
 <th class="a-text-right mojo-field-type-percent_delta mojo-sortable-column mojo-estimatable a-nowrap"><a class="a-link-normal
a-nowrap" href="?area=GB&amp;sort=grossLastWeekDelta&amp;ref_=bo_wey__resort#table" title="Overall Gross Change/Week">%± LW<spa
n class="a-letter-space"></span><span class="icon aok-relative"><i class="a-icon a-icon-expand table-sort-desc-placeholder" rol
e="presentation"></i><i class="a-icon a-icon-collapse table-sort-asc-placeholder" role="presentation"></i></span></a></th>,
 <th class="a-text-right mojo-field-type-positive_integer mojo-sortable-column a-nowrap"><a class="a-link-normal a-nowrap" href
="?area=GB&amp;sort=numReleases&amp;ref_=bo_wey__resort#table" title="Number of Releases">Releases<span class="a-letter-space">
</span><span class="icon aok-relative"><i class="a-icon a-icon-expand table-sort-desc-placeholder" role="presentation"></i><i c
lass="a-icon a-icon-collapse table-sort-asc-placeholder" role="presentation"></i></span></a></th>,
 <th class="a-text-left mojo-field-type-release mojo-cell-wide mojo-sortable-column a-nowrap"><span title="#1 Release">#1 Relea
se</span>
 </th>,
 <th class="a-text-left mojo-field-type-genre hidden mojo-sortable-column hidden a-nowrap"><span title="Genre">Genre</span>
```

```
table.find_all('tr')[0].text.split('\n')
```

```
['DatesTop 10 Gross%± LWOverall Gross%± LWReleases#1 Release',
 'Genre',
 'Budget',
 'Running Time',
 'WeekLong Weekend',
 '']
```

```
lst = []

for row in table.find_all('tr')[1:-1]:

    s = pd.Series([data.text for data in row.find_all('td')])

    lst.append(s)


data = pd.concat(lst, axis=1).T


data.head(2)


print(f'(MOVIES, COLUMNS) -> {data.shape}')


print(f'% OF MISSING VALUES PER COLUMN\n{(data.isnull().sum() /
data.shape[0]) * 100}')
```



```
NUMBER OF TABLES IN THE PAGE: 1
DatesTop 10 Gross%± LWOverall Gross%± LWReleases#1 Release
Genre
Budget
Running Time
WeekLong Weekend

(MOVIES, COLUMNS) -> (51, 12)
% OF MISSING VALUES PER COLUMN
0     0.0
1     0.0
2     0.0
3     0.0
4     0.0
5     0.0
6     0.0
7     0.0
8     0.0
9     0.0
10    0.0
11    0.0
dtype: float64
```

Practical 5

AIM :- Apriori Algorithm implementation in case study.

THEORY : -

Apriori algorithm refers to an algorithm that is used in mining frequent products sets and relevant association rules. Generally, the apriori algorithm operates on a database containing a huge number of transactions. For example, the items customers but at a Big Bazar.

CODE / OUTPUT : -

Minimum Support = 50 %

Minimum Confidence = 70 %

Itemset ={ Bread, Chicken, Butter, Milk, Toast}

Transaction ID	**Items**
100	{Bread, Butter, Milk}
200	{Chicken, Butter, Toast}
300	{Bread, Chicken, Butter, Toast}
400	{Chicken, Toast}

Item	**Support**
Bread	**2 /4 = 0.5 = 50%**
Chicken	**3 /4 = 0.5 = 75%**
Butter	**3 /4 = 0.75 = 75%**
Milk	**1/4 = 0.25 = 25%**
Toast	**3/4 = 0.75 = 75%**

Itemset = { Bread, Chicken , Butter, Toast}

Item	**Support**
{Bread, Chicken}	**¼ = 0.25 =25%**
{Bread, Butter}	**2/4 =0.50 = 50%**
{Bread, Toast}	**¼ = 0.25 = 25%**
{Chicken, Butter}	**2/4 = 0.50 = 50 %**
{Chicken, Toast}	**¾ = 0.75 = 75%**
{Butter, Toast}	**2/4 = 0.50 = 50%**

Itemset = ({Bread, Butter}, {Chicken, Butter} , {Chicken, Toast}, {Butter, Toast})

Item	**Support**
{Bread, Butter, Toast}	1/ 4 = 0.25 = 25%
{Chicken, Butter, Toast}	2/4 =0.50 = 50 %
{Bread, Butter, Chicken}	¼ = 0.25 = 25%

Minimum Support = 50%

Minimum Confidence = 70%

Final Resultant Set based on Support = {Chicken, Butter, Toast}

Rules

1 . (Chicken & Butter) - > Toast 2 (50%)

2. (Butter & Toast) -> Chicken 2 (50%)

3. (Chicken & Toast) -> Butter 2 (50%)

4. Chicken - > (Butter & Toast) 2 (50%)

5. Toast -> (Chicken & Butter) 2 (50%)

6. Butter -> (Chicken & Toast) 2 (50%)

Confidence = S(A U B).count / S(A).count

1 . (Chicken & Butter) - > Toast 2 (50%)

S((Chicken &Butter) U (Toast))/ S(Chicken & Butter)

=2 / 2 = 1 = **<u>100%</u>**

2. . (Butter & Toast) -> Chicken

Confidence = S(A U B).count / S(A).count

S((Butter & Toast) U Chicken)) /S(Butter & Toast)

=2 / 2 = 1 = **<u>100%</u>**

3. (Chicken & Toast) -> Butter 2 (50%)

Confidence = S(A U B).count / S(A).count

S((Chicken & Toast) U (Butter))/S(Chicken & Toast)

=2/3 = 0.666 **<u>= 67%</u>**

4. Chicken - > (Butter & Toast) 2 (50%)

Confidence = S(A U B).count / S(A).count

S((Chicken) U (Butter & Toast))/S(Chicken)

=2/3 = 0.666 = **<u>67%</u>**

Minimum Support = 50%

Minimum Confidence = 70%

5. Toast -> (Chicken & Butter) 2 (50%)

Confidence = S(A U B).count / S(A).count

S((Toast) U (Chicken & Butter))/S(Toast)

=2/3 = 0.666 = **67%**

6. Butter -> (Chicken & Toast) 2 (50%)

Confidence = S(A U B).count / S(A).count

S((Butter) U (Chicken & Toast))/S(Butter)

=2/3 = 0.666 = **67%**

Final Associated Items rules are

1 . (Chicken & Butter) - > Toast 2 (50%)

2. (Butter & Toast) -> Chicken 2 (50%)

Practical 6

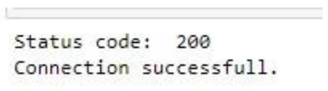
AIM :- Develop a basic crawler for the web search for user defined keywords.

THEORY : -

A Web crawler, sometimes called a spider or spiderbot and often shortened to crawler, is an Internet bot that systematically browses the World Wide Web and that is typically operated by search engines for the purpose of Web indexing (web spidering).

CODE / OUTPUT : -

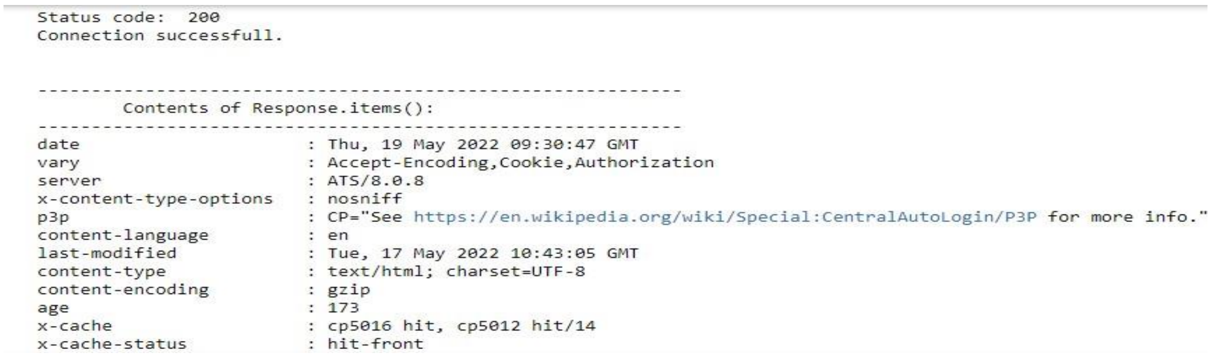
```
import requests
url = 'https://en.wikipedia.org/wiki/Stock_market'


response = requests.get(url, timeout=3) #timeout set to stop the request
action in case of hanging
print('Status code: ',response.status_code)
if response.status_code==200:
    print('Connection successfull.\n\n')
else:
    print('Error. Check status code table.\n\n')
```



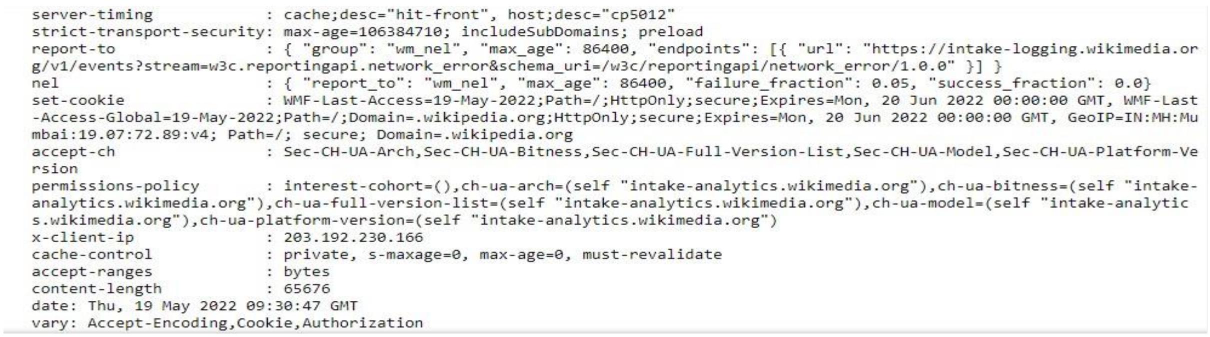
```
Status code:  200
Connection successfull.
```

```
print(f"{'---'*20}\n\tContents of Response.items():\n{'---'*20}")


for k,v in response.headers.items():
    print(f"{k:{25}}: {v:{40}}") # Note: add :{number} inside of a


print(f"{'---'*20}\nStatus code: {response.status_code}\n{'---'*20}\n")
```

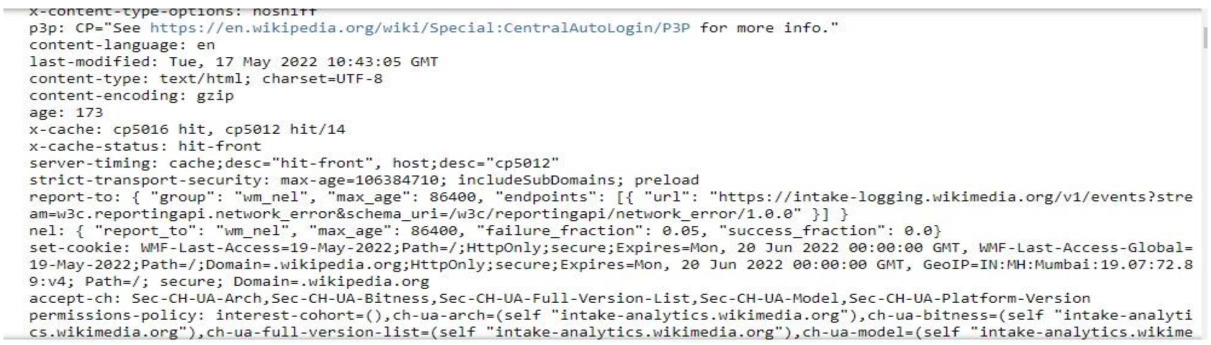
```
from bs4 import BeautifulSoup

# Feed the response's .content into BeauitfulSoup

page_content = response.content

soup = BeautifulSoup(page_content,'lxml') #'html.parser')


# Preview soup contents using .prettify()

print(soup.prettify()[:2000])


body = soup.body

for child in body.children:

  # print child if its not empty

  print(child if child is not None else ' ', '\n\n')  # '\n\n' for visual
separation


title = soup.head.title

print(title.parent.name)


results = soup.find_all()


results
```



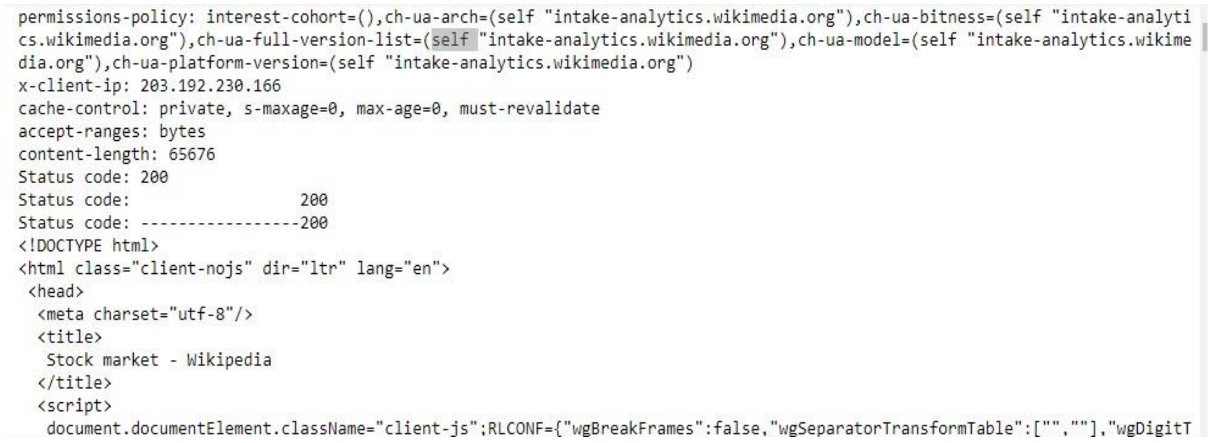
```
Status code:  200
Connection successfull.


------------------------------------------------------------
        Contents of Response.items():
------------------------------------------------------------
date                         : Thu, 19 May 2022 09:30:47 GMT
vary                         : Accept-Encoding,Cookie,Authorization
server                       : ATS/8.0.8
x-content-type-options       : nosniff
p3p                          : CP="See https://en.wikipedia.org/wiki/Special:CentralAutoLogin/P3P for more info."
content-language             : en
last-modified                : Tue, 17 May 2022 10:43:05 GMT
content-type                 : text/html; charset=UTF-8
content-encoding             : gzip
age                          : 173
x-cache                      : cp5016 hit, cp5012 hit/14
x-cache-status               : hit-front
```



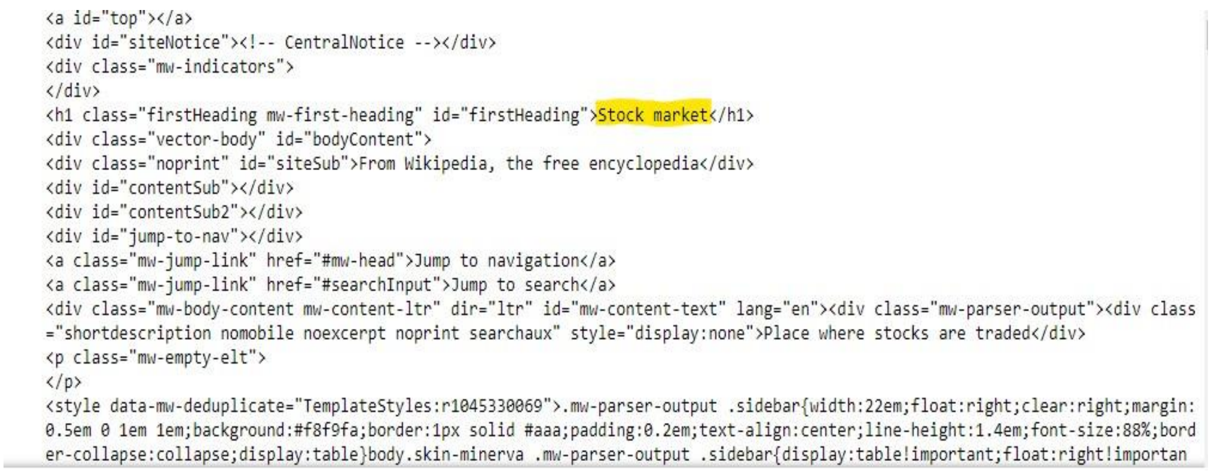
```
server-timing            : cache;desc="hit-front", host;desc="cp5012"
strict-transport-security: max-age=106384710; includeSubDomains; preload
report-to                : { "group": "wm_nel", "max_age": 86400, "endpoints": [{ "url": "https://intake-logging.wikimedia.or
g/v1/events?stream=w3c.reportingapi.network_error&schema_uri=/w3c/reportingapi/network_error/1.0.0" }] }
nel                      : { "report_to": "wm_nel", "max_age": 86400, "failure_fraction": 0.05, "success_fraction": 0.0}
set-cookie               : WMF-Last-Access=19-May-2022;Path=/;HttpOnly;secure;Expires=Mon, 20 Jun 2022 00:00:00 GMT, WMF-Last
-Access-Global=19-May-2022;Path=/;Domain=.wikipedia.org;HttpOnly;secure;Expires=Mon, 20 Jun 2022 00:00:00 GMT, GeoIP=IN:MH:Mu
mbai:19.07:72.89:v4; Path=/; secure; Domain=.wikipedia.org
accept-ch                : Sec-CH-UA-Arch,Sec-CH-UA-Bitness,Sec-CH-UA-Full-Version-List,Sec-CH-UA-Model,Sec-CH-UA-Platform-Ve
rsion
permissions-policy       : interest-cohort=(),ch-ua-arch=(self "intake-analytics.wikimedia.org"),ch-ua-bitness=(self "intake-
analytics.wikimedia.org"),ch-ua-full-version-list=(self "intake-analytics.wikimedia.org"),ch-ua-model=(self "intake-analytic
s.wikimedia.org"),ch-ua-platform-version=(self "intake-analytics.wikimedia.org")
x-client-ip              : 203.192.230.166
cache-control            : private, s-maxage=0, max-age=0, must-revalidate
accept-ranges            : bytes
content-length           : 65676
date: Thu, 19 May 2022 09:30:47 GMT
vary: Accept-Encoding,Cookie,Authorization
```



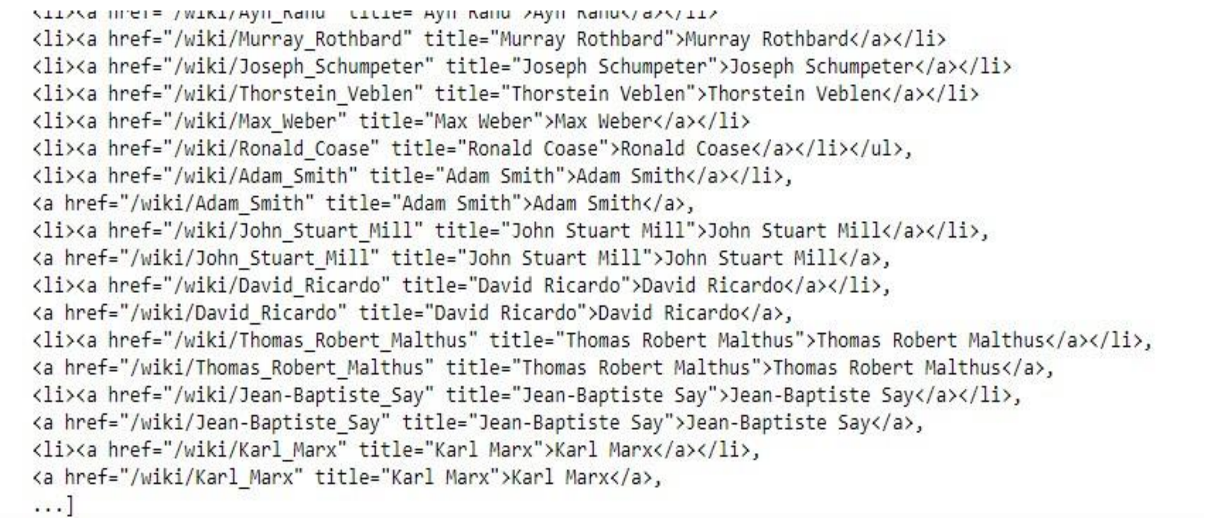
```
x-content-type-options: nosniff
p3p: CP="See https://en.wikipedia.org/wiki/Special:CentralAutoLogin/P3P for more info."
content-language: en
last-modified: Tue, 17 May 2022 10:43:05 GMT
content-type: text/html; charset=UTF-8
content-encoding: gzip
age: 173
x-cache: cp5016 hit, cp5012 hit/14
x-cache-status: hit-front
server-timing: cache;desc="hit-front", host;desc="cp5012"
strict-transport-security: max-age=106384710; includeSubDomains; preload
report-to: { "group": "wm_nel", "max_age": 86400, "endpoints": [{ "url": "https://intake-logging.wikimedia.org/v1/events?stre
am=w3c.reportingapi.network_error&schema_uri=/w3c/reportingapi/network_error/1.0.0" }] }
nel: { "report_to": "wm_nel", "max_age": 86400, "failure_fraction": 0.05, "success_fraction": 0.0}
set-cookie: WMF-Last-Access=19-May-2022;Path=/;HttpOnly;secure;Expires=Mon, 20 Jun 2022 00:00:00 GMT, WMF-Last-Access-Global=
19-May-2022;Path=/;Domain=.wikipedia.org;HttpOnly;secure;Expires=Mon, 20 Jun 2022 00:00:00 GMT, GeoIP=IN:MH:Mumbai:19.07:72.8
9:v4; Path=/; secure; Domain=.wikipedia.org
accept-ch: Sec-CH-UA-Arch,Sec-CH-UA-Bitness,Sec-CH-UA-Full-Version-List,Sec-CH-UA-Model,Sec-CH-UA-Platform-Version
permissions-policy: interest-cohort=(),ch-ua-arch=(self "intake-analytics.wikimedia.org"),ch-ua-bitness=(self "intake-analyti
cs.wikimedia.org"),ch-ua-full-version-list=(self "intake-analytics.wikimedia.org"),ch-ua-model=(self "intake-analytics.wikime
```



```
permissions-policy: interest-cohort=(),ch-ua-arch=(self "intake-analytics.wikimedia.org"),ch-ua-bitness=(self "intake-analyti
cs.wikimedia.org"),ch-ua-full-version-list=(self "intake-analytics.wikimedia.org"),ch-ua-model=(self "intake-analytics.wikime
dia.org"),ch-ua-platform-version=(self "intake-analytics.wikimedia.org")
x-client-ip: 203.192.230.166
cache-control: private, s-maxage=0, max-age=0, must-revalidate
accept-ranges: bytes
content-length: 65676
Status code: 200
Status code:                  200
Status code: -----------------200
<!DOCTYPE html>
<html class="client-nojs" dir="ltr" lang="en">
 <head>
  <meta charset="utf-8"/>
  <title>
   Stock market - Wikipedia
  </title>
  <script>
   document.documentElement.className="client-js";RLCONF={"wgBreakFrames":false,"wgSeparatorTransformTable":["",""],"wgDigitT
```



```
   <a id="top"></a>
   <div id="siteNotice"><!-- CentralNotice --></div>
   <div class="mw-indicators">
   </div>
   <h1 class="firstHeading mw-first-heading" id="firstHeading">Stock market</h1>
   <div class="vector-body" id="bodyContent">
   <div class="noprint" id="siteSub">From Wikipedia, the free encyclopedia</div>
   <div id="contentSub"></div>
   <div id="contentSub2"></div>
   <div id="jump-to-nav"></div>
   <a class="mw-jump-link" href="#mw-head">Jump to navigation</a>
   <a class="mw-jump-link" href="#searchInput">Jump to search</a>
   <div class="mw-body-content mw-content-ltr" dir="ltr" id="mw-content-text" lang="en"><div class="mw-parser-output"><div class
   ="shortdescription nomobile noexcerpt noprint searchaux" style="display:none">Place where stocks are traded</div>
   <p class="mw-empty-elt">
   </p>
   <style data-mw-deduplicate="TemplateStyles:r1045330069">.mw-parser-output .sidebar{width:22em;float:right;clear:right;margin:
   0.5em 0 1em 1em;background:#f8f9fa;border:1px solid #aaa;padding:0.2em;text-align:center;line-height:1.4em;font-size:88%;bord
   er-collapse:collapse;display:table}body.skin-minerva .mw-parser-output .sidebar{display:table!important;float:right!importan
```



```
<li><a href="/wiki/Ayn_Rand" title="Ayn Rand">Ayn Rand</a></li>
<li><a href="/wiki/Murray_Rothbard" title="Murray Rothbard">Murray Rothbard</a></li>
<li><a href="/wiki/Joseph_Schumpeter" title="Joseph Schumpeter">Joseph Schumpeter</a></li>
<li><a href="/wiki/Thorstein_Veblen" title="Thorstein Veblen">Thorstein Veblen</a></li>
<li><a href="/wiki/Max_Weber" title="Max Weber">Max Weber</a></li>
<li><a href="/wiki/Ronald_Coase" title="Ronald Coase">Ronald Coase</a></li></ul>,
<li><a href="/wiki/Adam_Smith" title="Adam Smith">Adam Smith</a></li>,
<a href="/wiki/Adam_Smith" title="Adam Smith">Adam Smith</a>,
<li><a href="/wiki/John_Stuart_Mill" title="John Stuart Mill">John Stuart Mill</a></li>,
<a href="/wiki/John_Stuart_Mill" title="John Stuart Mill">John Stuart Mill</a>,
<li><a href="/wiki/David_Ricardo" title="David Ricardo">David Ricardo</a></li>,
<a href="/wiki/David_Ricardo" title="David Ricardo">David Ricardo</a>,
<li><a href="/wiki/Thomas_Robert_Malthus" title="Thomas Robert Malthus">Thomas Robert Malthus</a></li>,
<a href="/wiki/Thomas_Robert_Malthus" title="Thomas Robert Malthus">Thomas Robert Malthus</a>,
<li><a href="/wiki/Jean-Baptiste_Say" title="Jean-Baptiste Say">Jean-Baptiste Say</a></li>,
<a href="/wiki/Jean-Baptiste_Say" title="Jean-Baptiste Say">Jean-Baptiste Say</a>,
<li><a href="/wiki/Karl_Marx" title="Karl Marx">Karl Marx</a></li>,
<a href="/wiki/Karl_Marx" title="Karl Marx">Karl Marx</a>,
...]
```

Practical 7

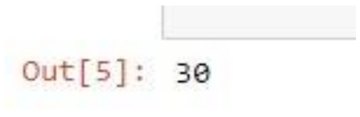
AIM :- Develop a focused crawler for local search .

THEORY : -

A focused crawler is a web crawler that collects Web pages that satisfy some specific property, by carefully prioritizing the crawl frontier and managing the hyperlink exploration process. Some predicates may be based on simple, deterministic and surface properties. For example, a crawler's mission may be to crawl pages from only the .jp domain. Other predicates may be softer or comparative, e.g., "crawl pages about baseball", or "crawl pages with large PageRank". An important page property pertains to topics, leading to 'topical crawlers'. For example, a topical crawler may be deployed to collect pages about solar power, swine flu, or even more abstract concepts like controversy while minimizing resources spent fetching pages on other topics.

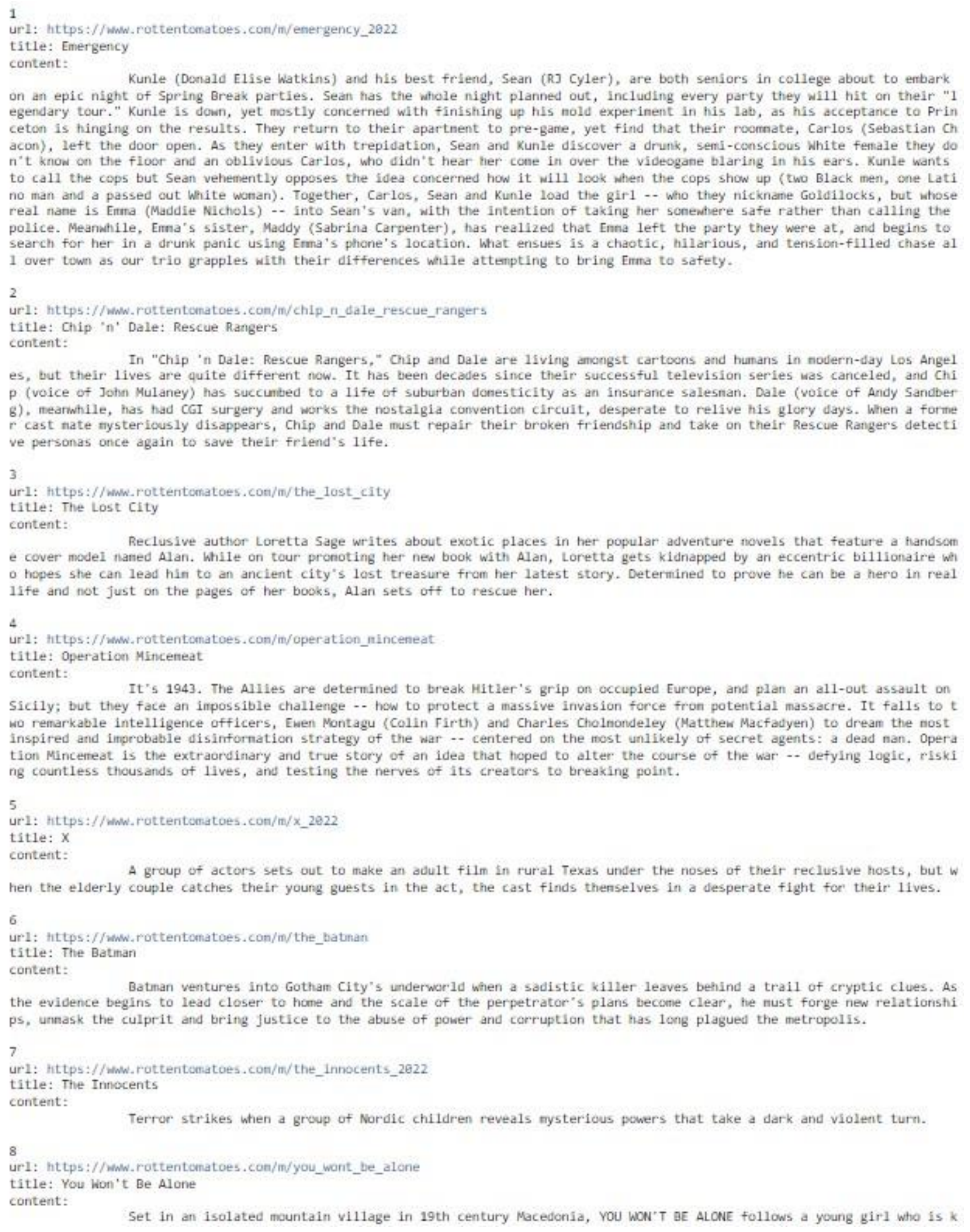
CODE / OUTPUT : -

```
import requests

import lxml

from bs4 import BeautifulSoup


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

f = requests.get(url)



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

#parse in normal html


movies = soup.find_all('div',{'class':'discovery-
tiles__wrap'})[1].find_all_next('a')


movies_list = []

num = 0

for movie in movies:
  try:
    url = 'https://www.rottentomatoes.com' + movie["href"]
    movie_f = requests.get(url)
    movie_page = BeautifulSoup(movie_f.content,'html')
    num += 1
    title = movie_page.find('h1', {'class': 'scoreboard__title'})
    movie_content = movie_page.find('div', {'class': 'movie_synopsis'})
    movies_list.append({
        "#": num,
        "url": url,
        "title": title.text,
        "content": movie_content.getText()
    })
  except:
    continue

len(movies_list)
```

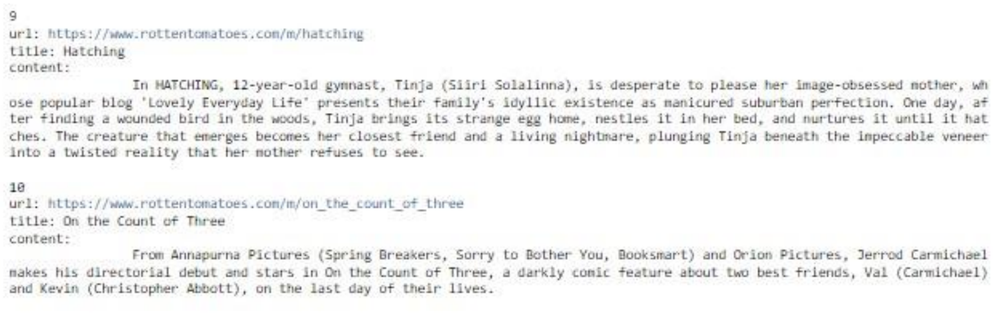


```
Out[5]: 30
```

```
for movie in movies_list[:10]:
  print(f"{movie['#']} \nurl: {movie['url']} \ntitle: {movie['title']}
\ncontent: {movie['content']}")
```



```
1
url: https://www.rottentomatoes.com/m/emergency_2022
title: Emergency
content:
                Kunle (Donald Elise Watkins) and his best friend, Sean (RJ Cyler), are both seniors in college about to embark
on an epic night of Spring Break parties. Sean has the whole night planned out, including every party they will hit on their "l
egendary tour." Kunle is down, yet mostly concerned with finishing up his mold experiment in his lab, as his acceptance to Prin
ceton is hinging on the results. They return to their apartment to pre-game, yet find that their roommate, Carlos (Sebastian Ch
acon), left the door open. As they enter with trepidation, Sean and Kunle discover a drunk, semi-conscious White female they do
n't know on the floor and an oblivious Carlos, who didn't hear her come in over the videogame blaring in his ears. Kunle wants
to call the cops but Sean vehemently opposes the idea concerned how it will look when the cops show up (two Black men, one Lati
no man and a passed out White woman). Together, Carlos, Sean and Kunle load the girl -- who they nickname Goldilocks, but whose
real name is Emma (Maddie Nichols) -- into Sean's van, with the intention of taking her somewhere safe rather than calling the
police. Meanwhile, Emma's sister, Maddy (Sabrina Carpenter), has realized that Emma left the party they were at, and begins to
search for her in a drunk panic using Emma's phone's location. What ensues is a chaotic, hilarious, and tension-filled chase al
l over town as our trio grapples with their differences while attempting to bring Emma to safety.

2
url: https://www.rottentomatoes.com/m/chip_n_dale_rescue_rangers
title: Chip 'n' Dale: Rescue Rangers
content:
                In "Chip 'n Dale: Rescue Rangers," Chip and Dale are living amongst cartoons and humans in modern-day Los Angel
es, but their lives are quite different now. It has been decades since their successful television series was canceled, and Chi
p (voice of John Mulaney) has succumbed to a life of suburban domesticity as an insurance salesman. Dale (voice of Andy Sandber
g), meanwhile, has had CGI surgery and works the nostalgia convention circuit, desperate to relive his glory days. When a forme
r cast mate mysteriously disappears, Chip and Dale must repair their broken friendship and take on their Rescue Rangers detecti
ve personas once again to save their friend's life.

3
url: https://www.rottentomatoes.com/m/the_lost_city
title: The Lost City
content:
                Reclusive author Loretta Sage writes about exotic places in her popular adventure novels that feature a handsom
e cover model named Alan. While on tour promoting her new book with Alan, Loretta gets kidnapped by an eccentric billionaire wh
o hopes she can lead him to an ancient city's lost treasure from her latest story. Determined to prove he can be a hero in real
life and not just on the pages of her books, Alan sets off to rescue her.

4
url: https://www.rottentomatoes.com/m/operation_mincemeat
title: Operation Mincemeat
content:
                It's 1943. The Allies are determined to break Hitler's grip on occupied Europe, and plan an all-out assault on
Sicily; but they face an impossible challenge -- how to protect a massive invasion force from potential massacre. It falls to t
wo remarkable intelligence officers, Ewen Montagu (Colin Firth) and Charles Cholmondeley (Matthew Macfadyen) to dream the most
inspired and improbable disinformation strategy of the war -- centered on the most unlikely of secret agents: a dead man. Opera
tion Mincemeat is the extraordinary and true story of an idea that hoped to alter the course of the war -- defying logic, riski
ng countless thousands of lives, and testing the nerves of its creators to breaking point.

5
url: https://www.rottentomatoes.com/m/x_2022
title: X
content:
                A group of actors sets out to make an adult film in rural Texas under the noses of their reclusive hosts, but w
hen the elderly couple catches their young guests in the act, the cast finds themselves in a desperate fight for their lives.

6
url: https://www.rottentomatoes.com/m/the_batman
title: The Batman
content:
                Batman ventures into Gotham City's underworld when a sadistic killer leaves behind a trail of cryptic clues. As
the evidence begins to lead closer to home and the scale of the perpetrator's plans become clear, he must forge new relationshi
ps, unmask the culprit and bring justice to the abuse of power and corruption that has long plagued the metropolis.

7
url: https://www.rottentomatoes.com/m/the_innocents_2022
title: The Innocents
content:
                Terror strikes when a group of Nordic children reveals mysterious powers that take a dark and violent turn.

8
url: https://www.rottentomatoes.com/m/you_wont_be_alone
title: You Won't Be Alone
content:
                Set in an isolated mountain village in 19th century Macedonia, YOU WON'T BE ALONE follows a young girl who is k
```



```
9
url: https://www.rottentomatoes.com/m/hatching
title: Hatching
content:
                In HATCHING, 12-year-old gymnast, Tinja (Siiri Solalinna), is desperate to please her image-obsessed mother, wh
ose popular blog 'Lovely Everyday Life' presents their family's idyllic existence as manicured suburban perfection. One day, af
ter finding a wounded bird in the woods, Tinja brings its strange egg home, nestles it in her bed, and nurtures it until it hat
ches. The creature that emerges becomes her closest friend and a living nightmare, plunging Tinja beneath the impeccable veneer
into a twisted reality that her mother refuses to see.

10
url: https://www.rottentomatoes.com/m/on_the_count_of_three
title: On the Count of Three
content:
                From Annapurna Pictures (Spring Breakers, Sorry to Bother You, Booksmart) and Orion Pictures, Jerrod Carmichael
makes his directorial debut and stars in On the Count of Three, a darkly comic feature about two best friends, Val (Carmichael)
and Kevin (Christopher Abbott), on the last day of their lives.
```

Practical 8

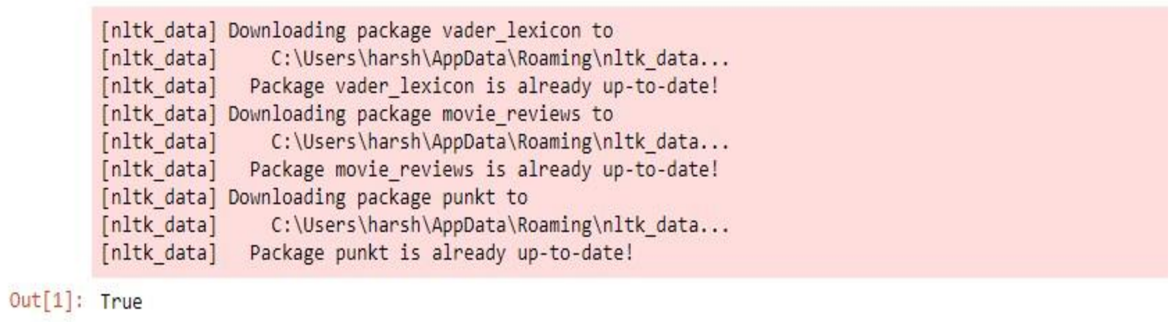
AIM :- Sentiment analysis for reviews by customers and visualize the same.

THEORY : -

Sentiment analysis (also known as opinion mining or emotion AI) is the use of natural language processing, text analysis, computational linguistics, and biometrics to systematically identify, extract, quantify, and study affective states and subjective information. Sentiment analysis is widely applied to voice of the customer materials such as reviews and survey responses, online and social media, and healthcare materials for applications that range from marketing to customer service to clinical medicine. With the rise of deep language models, such as RoBERTa, also more difficult data domains can be analyzed, e.g., news texts where authors typically express their opinion/sentiment less explicitly.

CODE / OUTPUT : -

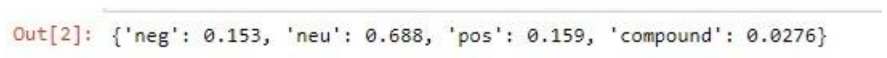
```
"""Pract10_WM.ipynb"""

!pip install matplotlib pandas nltk textblob


import nltk

nltk.download('vader_lexicon')

nltk.download('movie_reviews')

nltk.download('punkt')
```



```
[nltk_data] Downloading package vader_lexicon to
[nltk_data]     C:\Users\harsh\AppData\Roaming\nltk_data...
[nltk_data]   Package vader_lexicon is already up-to-date!
[nltk_data] Downloading package movie_reviews to
[nltk_data]     C:\Users\harsh\AppData\Roaming\nltk_data...
[nltk_data]   Package movie_reviews is already up-to-date!
[nltk_data] Downloading package punkt to
[nltk_data]     C:\Users\harsh\AppData\Roaming\nltk_data...
[nltk_data]   Package punkt is already up-to-date!
Out[1]: True
```

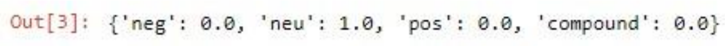
```
from nltk.sentiment.vader import SentimentIntensityAnalyzer as SIA
```

```
sia = SIA()

sia.polarity_scores("This restaurant was great, but I'm not sure if I'll
go there again.")
```



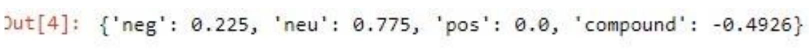
```
Out[2]: {'neg': 0.153, 'neu': 0.688, 'pos': 0.159, 'compound': 0.0276}
```

```
text = "I just got a call from my boss - does he realise it's Saturday?"

sia.polarity_scores(text)
```



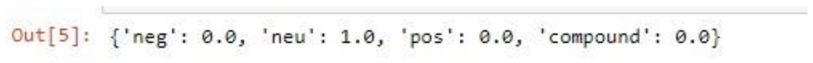
```
Out[3]: {'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0}
```

```
text = "I just got a call from my boss - does he realise it's Saturday?
:)"

sia.polarity_scores(text)
```



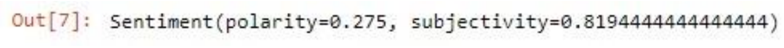
```
Out[4]: {'neg': 0.225, 'neu': 0.775, 'pos': 0.0, 'compound': -0.4926}
```

```
text = "I just got a call from my boss - does he realise it's Saturday?
☺"

sia.polarity_scores(text)
```



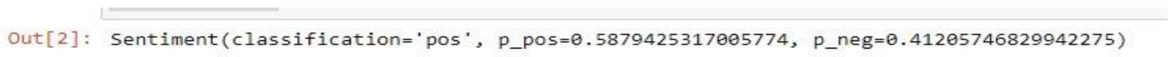
```
Out[5]: {'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0}
```

```
from textblob import TextBlob

from textblob import Blobber

from textblob.sentiments import NaiveBayesAnalyzer


blob = TextBlob("This restaurant was great, but I'm not sure if I'll go
there again.")

blob.sentiment
```



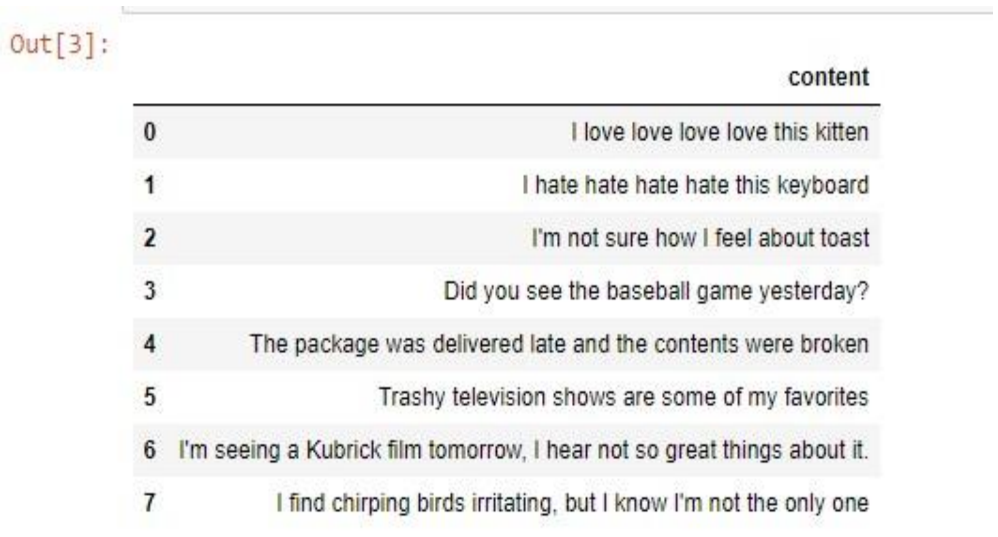
```
Out[7]: Sentiment(polarity=0.275, subjectivity=0.8194444444444444)
```

```
blobber = Blobber(analyzer=NaiveBayesAnalyzer())


blob = blobber("This restaurant was great, but I'm not sure if I'll go
there again.")

blob.sentiment
```



```
Out[2]: Sentiment(classification='pos', p_pos=0.5879425317005774, p_neg=0.41205746829942275)
```

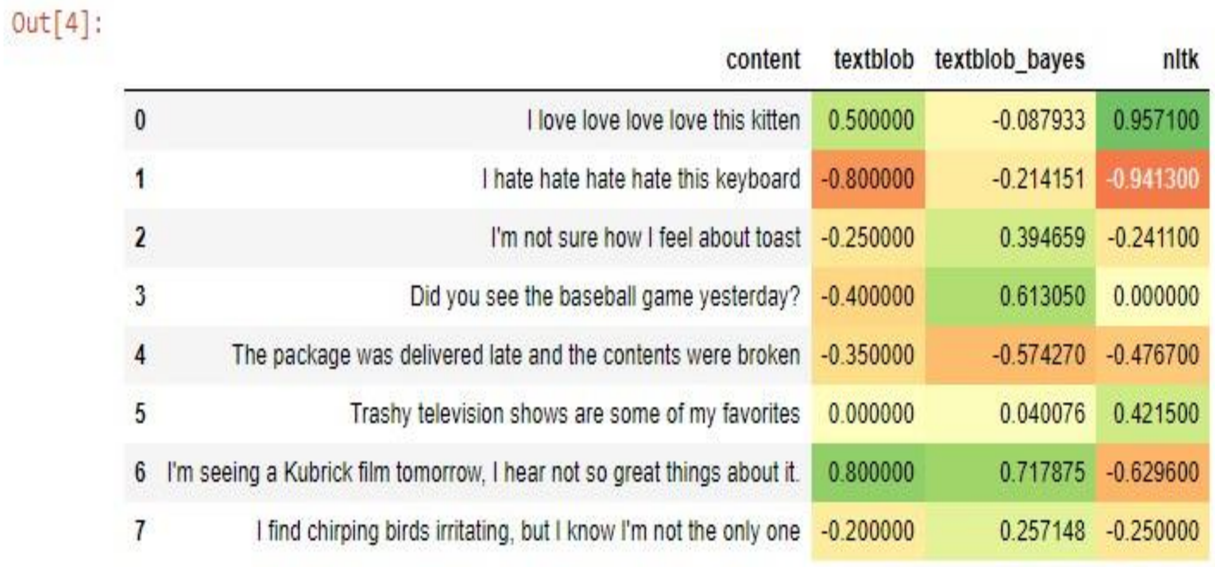
```
import pandas as pd
df = pd.DataFrame({'content': [
    "I love love love love this kitten",
    "I hate hate hate hate this keyboard",
    "I'm not sure how I feel about toast",
    "Did you see the baseball game yesterday?",
    "The package was delivered late and the contents were broken",
    "Trashy television shows are some of my favorites",
    "I'm seeing a Kubrick film tomorrow, I hear not so great things about
it.",
    "I find chirping birds irritating, but I know I'm not the only one",
]})
df
```



Out[3]:

	content
0	I love love love love this kitten
1	I hate hate hate hate this keyboard
2	I'm not sure how I feel about toast
3	Did you see the baseball game yesterday?
4	The package was delivered late and the contents were broken
5	Trashy television shows are some of my favorites
6	I'm seeing a Kubrick film tomorrow, I hear not so great things about it.
7	I find chirping birds irritating, but I know I'm not the only one

```
def get_scores(content):
    blob = TextBlob(content)
    nb_blob = blobber(content)
    sia_scores = sia.polarity_scores(content)

    return pd.Series({
        'content': content,
        'textblob': blob.sentiment.polarity,
        'textblob_bayes': nb_blob.sentiment.p_pos -
nb_blob.sentiment.p_neg,
        'nltk': sia_scores['compound'],
    })

scores = df.content.apply(get_scores)
scores.style.background_gradient(cmap='RdYlGn', axis=None, low=0.4,
high=0.4)
```



Out[4]:

	content	textblob	textblob_bayes	nltk
0	I love love love love this kitten	0.500000	-0.087933	0.957100
1	I hate hate hate hate this keyboard	-0.800000	-0.214151	-0.941300
2	I'm not sure how I feel about toast	-0.250000	0.394659	-0.241100
3	Did you see the baseball game yesterday?	-0.400000	0.613050	0.000000
4	The package was delivered late and the contents were broken	-0.350000	-0.574270	-0.476700
5	Trashy television shows are some of my favorites	0.000000	0.040076	0.421500
6	I'm seeing a Kubrick film tomorrow, I hear not so great things about it.	0.800000	0.717875	-0.629600
7	I find chirping birds irritating, but I know I'm not the only one	-0.200000	0.257148	-0.250000