In [ ]:
```python
import requests
```

In [ ]:
```python
from bs4 import BeautifulSoup
```

In [ ]:
```python
import nltk
```

In [ ]:
```python
page = requests.get("https://www.rjcollege.edu.in/about-us/")
page1 = requests.get("https://www.rjcollege.edu.in/pgadmission2022-23/")
soup = BeautifulSoup(page.content,'html.parser')
soup1 = BeautifulSoup(page1.content, 'html.parser')
str3 = soup.find_all('p')[0].get_text()
#strall = [x.get_text() for x in soup.find_all('p')]
```

In [ ]:
```python
str3
```

Out[ ]:
```
'On the auspicious day of Shri Krishna Janmashtami, 15th August 1938, the people of Ghatkopar and the surrounding suburbs witnessed the birth of Hindi Vidya P
rachar Samiti, a brain child of a visionary Late Shri Nandkishore Singh Jairamji. The Samiti was established with the objectives of catering to the educationa
l needs of the Hindi speaking community. It made a humble beginning by starting a primary school, which gradually expanded into a full-fledged secondary schoo
l.\nThe Hindi High School with its high academic standards has carved for itself a place not only among leading secondary schools in Mumbai but also education
al institutions imparting instructions in Hindi throughout Maharashtra. With its primary objectives achieved the Samiti decided to extend its frontiers and br
oaden its horizons. As a result, Ramniranjan Jhunjhunwala College came into existence in 1963, enabling a larger section of the society to take advantage of t
he facilities provided for higher education.\nIn 1976 the Junior College section was introduced and in 1981 the Commerce faculty commenced both at the Junior
and Degree College level.\nFrom 1999-2000 the College has added a number of self-financing courses like BMS, B.B.I, B.Sc. in C.S., I.T., Biotechnology, M.Sc.
in Computer Science and Biotechnology as well as add on courses, which further hone the special skills of the students.\nIn 2014 saw a change in education sys
tem with greater emphasis being given to employability of youth. As an effort to realize the dream of Make in India, Digital India, Clean and Green India, we
have started skill based program supported by University Grants commission known as Bachelor in Vocation.\nThe college has been reaccredited with ‘A’ Grade by
NAAC in 2014 with a CGPA 3.50 and received the Best College Award (2007-2008) of the University of Mumbai. The College has been bestowed with IMC RAMKRISHNA B
AJAJ PERFORMANCE EXCELLENCE TROPHY, 2010. The Principal of the college was awarded “Best Teacher” by Government of Maharashtra in 2011. Government of Maharash
tra conferred the college with “JAAGAR JAANIVANCHA” (First in Mumbai Suburban- in 2013 and Second in Mumbai Suburban- in 2014) for safety of girls.'
```

In [ ]:
```python
nltk.download('punkt')
from nltk.tokenize import sent_tokenize, word_tokenize
```

```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
```

In [ ]:
```python
sents = sent_tokenize(str3)
sents
```

Out[ ]:
```
['On the auspicious day of Shri Krishna Janmashtami, 15th August 1938, the people of Ghatkopar and the surrounding suburbs witnessed the birth of Hindi Vidya
Prachar Samiti, a brain child of a visionary Late Shri Nandkishore Singh Jairamji.',
 'The Samiti was established with the objectives of catering to the educational needs of the Hindi speaking community.',
 'It made a humble beginning by starting a primary school, which gradually expanded into a full-fledged secondary school.',
 'The Hindi High School with its high academic standards has carved for itself a place not only among leading secondary schools in Mumbai but also educational
institutions imparting instructions in Hindi throughout Maharashtra.',
 'With its primary objectives achieved the Samiti decided to extend its frontiers and broaden its horizons.',
 'As a result, Ramniranjan Jhunjhunwala College came into existence in 1963, enabling a larger section of the society to take advantage of the facilities prov
ided for higher education.',
 'In 1976 the Junior College section was introduced and in 1981 the Commerce faculty commenced both at the Junior and Degree College level.',
 'From 1999-2000 the College has added a number of self-financing courses like BMS, B.B.I, B.Sc.',
 'in C.S., I.T., Biotechnology, M.Sc.',
 'in Computer Science and Biotechnology as well as add on courses, which further hone the special skills of the students.',
 'In 2014 saw a change in education system with greater emphasis being given to employability of youth.',
 'As an effort to realize the dream of Make in India, Digital India, Clean and Green India, we have started skill based program supported by University Grants
commission known as Bachelor in Vocation.',
 'The college has been reaccredited with ‘A’ Grade by NAAC in 2014 with a CGPA 3.50 and received the Best College Award (2007-2008) of the University of Mumba
i.',
 'The College has been bestowed with IMC RAMKRISHNA BAJAJ PERFORMANCE EXCELLENCE TROPHY, 2010.',
 'The Principal of the college was awarded “Best Teacher” by Government of Maharashtra in 2011.',
 'Government of Maharashtra conferred the college with “JAAGAR JAANIVANCHA” (First in Mumbai Suburban- in 2013 and Second in Mumbai Suburban- in 2014) for saf
ety of girls.']
```

In [ ]:
```python
words = word_tokenize(str3)
words
```

Out[ ]:
```
['On',
 'the',
 'auspicious',
 'day',
 'of',
 'Shri',
 'Krishna',
 'Janmashtami',
 ',',
 '15th',
 'August',
 '1938',
 ',',
 'the',
 'people',
 'of',
 'Ghatkopar',
 'and',
 'the',
 'surrounding',
 'suburbs',
 'witnessed',
 'the',
 'birth',
 'of',
 'Hindi',
 'Vidya',
 'Prachar',
 'Samiti',
 ',',
 'a',
 'brain',
 'child',
 'of',
 'a',
 'visionary',
 'Late',
 'Shri',
 'Nandkishore',
 'Singh',
 'Jairamji',
 '.',
 'The',
 'Samiti',
 'was',
 'established',
 'with',
 'the',
 'objectives',
 'of',
 'catering',
 'to',
 'the',
 'educational',
 'needs',
 'of',
 'the',
 'Hindi',
 'speaking',
 'community',
 '.',
 'It',
 'made',
 'a',
 'humble',
 'beginning',
 'by',
 'starting',
 'a',
 'primary',
 'school',
 ',',
 'which',
 'gradually',
 'expanded',
 'into',]
```

In [ ]:
```python
# https://www.geeksforgeeks.org/python-web-scraping-tutorial/
```

In [ ]:
```python
# <table id="tablepress-57" class="tablepress tablepress-id-57">
```

In [ ]:
```python
page1 = requests.get("https://www.rjcollege.edu.in/pgadmission2022-23/")
soup1 = BeautifulSoup(page1.content, 'html.parser')
table1 = soup1.find('table', id='tablepress-57')
```

In [ ]:
```python
table1
```

Out[ ]:
```
<table class="tablepress tablepress-id-57" id="tablepress-57">
<thead>
<tr class="row-1 odd">
<th class="column-1">Program</th><th class="column-2">FULL</th><th class="column-3">I ST INSTALLMENT</th><th class="column-4">II ND INSTALLMENT</th><th class
="column-5">SC</th><th class="column-6">ST</th>
</tr>
</thead>
<tbody class="row-hover">
<tr class="row-2 even">
<td class="column-1">MSC I  (Botany, Zoology, Chemistry-Physical, Organic and Inorganic)</td><td class="column-2">15465</td><td class="column-3">10100</td><td
class="column-4">5365</td><td class="column-5">590</td><td class="column-6">3740</td>
</tr>
<tr class="row-3 odd">
<td class="column-1">MA I (HINDI / ENGLISH)</td><td class="column-2">11615</td><td class="column-3">7500</td><td class="column-4">4115</td><td class="column-
5">590</td><td class="column-6">3740</td>
</tr>
<tr class="row-4 even">
<td class="column-1">MCOM I  </td><td class="column-2">15415</td><td class="column-3">10000</td><td class="column-4">5415</td><td class="column-5">590</td><td
class="column-6">3740</td>
</tr>
<tr class="row-5 odd">
<td class="column-1">MSC PHYSICS I</td><td class="column-2">30615</td><td class="column-3">19800</td><td class="column-4">10815</td><td class="column-5">590</
td><td class="column-6">3740</td>
</tr>
<tr class="row-6 even">
<td class="column-1">MSC I BT</td><td class="column-2">45015</td><td class="column-3">29100</td><td class="column-4">15915</td><td class="column-5">590</td><t
d class="column-6">3740</td>
</tr>
<tr class="row-7 odd">
<td class="column-1">MSC I CHEM ANALYTICAL</td><td class="column-2">45015</td><td class="column-3">29200</td><td class="column-4">15815</td><td class="column-
5">590</td><td class="column-6">3740</td>
</tr>
<tr class="row-8 even">
<td class="column-1">MSC I COMP SCI</td><td class="column-2">44615</td><td class="column-3">29000</td><td class="column-4">15615</td><td class="column-5">590
</td><td class="column-6">3740</td>
</tr>
<tr class="row-9 odd">
<td class="column-1">MSC I INFORMATION TECH</td><td class="column-2">44615</td><td class="column-3">29000</td><td class="column-4">15615</td><td class="column
-5">590</td><td class="column-6">3740</td>
</tr>
<tr class="row-10 even">
<td class="column-1">MA EMA I  </td><td class="column-2">52375</td><td class="column-3">34000</td><td class="column-4">18375</td><td class="column-5">2200</td
><td class="column-6">5350</td>
</tr>
<tr class="row-11 odd">
<td class="column-1">MSC I ENVIRONMENTAL SCIENCE &amp; DISASTER MANAGEMENT  </td><td class="column-2">45015</td><td class="column-3">29200</td><td class="colu
mn-4">15815</td><td class="column-5">-</td><td class="column-6">-</td>
</tr>
<tr class="row-12 even">
<td class="column-1">MSC I Data Science &amp; Artificial Intelligence </td><td class="column-2">61515</td><td class="column-3">40000</td><td class="column-4">
21515</td><td class="column-5">-</td><td class="column-6">-</td>
</tr>
<tr class="row-13 odd">
<td class="column-1">MSC I STATISTICS</td><td class="column-2">61515</td><td class="column-3">40000</td><td class="column-4">21515</td><td class="column-5">-
</td><td class="column-6">-</td>
</tr>
<tr class="row-14 even">
<td class="column-1">CLINICAL RESERCH</td><td class="column-2">36200</td><td class="column-3">23500</td><td class="column-4">12700</td><td class="column-5">-
</td><td class="column-6">-</td>
</tr>
<tr class="row-15 odd">
<td class="column-1">PG RA</td><td class="column-2">26200</td><td class="column-3">17000</td><td class="column-4">9200</td><td class="column-5">-</td><td clas
s="column-6">-</td>
</tr>
<tr class="row-16 even">
<td class="column-1">PG DAN</td><td class="column-2">26200</td><td class="column-3">17000</td><td class="column-4">9200</td><td class="column-5">-</td><td cla
ss="column-6">-</td>
</tr>
<tr class="row-17 odd">
<td class="column-1">HORTICULTURE &amp; GARDENING</td><td class="column-2">16200</td><td class="column-3">10500</td><td class="column-4">5700</td><td class="c
olumn-5">-</td><td class="column-6">-</td>
</tr>
</tbody>
</table>
```

In [ ]:
```python
page2 = requests.get("https://www.google.com/search?q=sensex&rlz=1C1CHBD_enIN981IN981&oq=sensex&aqs=chrome..69i57j0i67i131i433j0i131i433i512l2j0i20i131i263i433
soup2 = BeautifulSoup(page2.content, 'html.parser')
table2 = soup2.find('div', class_='zz63rd')
```

In [ ]:
```python
table2
```

In [ ]:
```python
page3 = requests.get('https://www.rjcollege.edu.in/gallery/')
soup3 = BeautifulSoup(page3.content, 'html.parser')
images_list = []

images = soup3.select('img')
for image in images:
    src = image.get('src')
    alt = image.get('alt')
    images_list.append({"src": src, "alt": alt})

for image in images_list:
    print(image)
```

```
{'src': '
BSRzC0wAAAABJRU5ErkJggg==', 'alt': 'R J COLLEGE'}
{'src': 'https://www.rjcollege.edu.in/wp-content/uploads/2021/12/website-logo-1.png', 'alt': 'R J COLLEGE'}
{'src': 'data:image/svg+xml,%3Csvg%20xmlns%3D%22http%3A%2F%2Fwww.w3.org%2F2000%2Fsvg%22%20width%3D%22150%22%20height%3D%22150%22%20viewBox%3D%220%200%20150%20
150%22%3E%3C%2Fsvg%3E', 'alt': None}
{'src': 'data:image/svg+xml,%3Csvg%20xmlns%3D%22http%3A%2F%2Fwww.w3.org%2F2000%2Fsvg%22%20width%3D%22150%22%20height%3D%22150%22%20viewBox%3D%220%200%20150%20
150%22%3E%3C%2Fsvg%3E', 'alt': None}
{'src': 'data:image/svg+xml,%3Csvg%20xmlns%3D%22http%3A%2F%2Fwww.w3.org%2F2000%2Fsvg%22%20width%3D%22150%22%20height%3D%22150%22%20viewBox%3D%220%200%20150%20
150%22%3E%3C%2Fsvg%3E', 'alt': None}
{'src': 'data:image/svg+xml,%3Csvg%20xmlns%3D%22http%3A%2F%2Fwww.w3.org%2F2000%2Fsvg%22%20width%3D%22150%22%20height%3D%22150%22%20viewBox%3D%220%200%20150%20
150%22%3E%3C%2Fsvg%3E', 'alt': None}
{'src': 'data:image/svg+xml,%3Csvg%20xmlns%3D%22http%3A%2F%2Fwww.w3.org%2F2000%2Fsvg%22%20width%3D%22150%22%20height%3D%22150%22%20viewBox%3D%220%200%20150%20
150%22%3E%3C%2Fsvg%3E', 'alt': None}
{'src': 'data:image/svg+xml,%3Csvg%20xmlns%3D%22http%3A%2F%2Fwww.w3.org%2F2000%2Fsvg%22%20width%3D%22150%22%20height%3D%22150%22%20viewBox%3D%220%200%20150%20
150%22%3E%3C%2Fsvg%3E', 'alt': None}
{'src': 'data:image/svg+xml,%3Csvg%20xmlns%3D%22http%3A%2F%2Fwww.w3.org%2F2000%2Fsvg%22%20width%3D%22150%22%20height%3D%22150%22%20viewBox%3D%220%200%20150%20
150%22%3E%3C%2Fsvg%3E', 'alt': None}
```

In [ ]:
```python
page4 = requests.get("https://www.w3schools.com/xml/note.xml")
soup4 = BeautifulSoup(page4.content, 'xml')
to = soup4.find_all('to')
to
```

Out[ ]:
```
[<to>Tove</to>]
```

In [ ]:
```python
# https://medium.com/@cmukesh8688/web-scraping-json-dictionary-and-pandas-part-2-4d34432281d9
# https://www.w3schools.com/xml/note.xml
```

In [ ]:
```python
page4 = requests.get("https://maps2.dcgis.dc.gov/dcgis/rest/services/FEEDS/MPD/MapServer/2/query?where=1%3D1&outFields=*&outSR=4326&f=json")
data_json = page4.json()
```

In [ ]:
```python
print(type(data_json))
```

```
<class 'dict'>
```

In [ ]:
```python
data_json
```

Out[ ]:
```
{'displayFieldName': 'CCN',
 'exceededTransferLimit': True,
 'features': [{'attributes': {'ANC': '8A',
    'BID': 'ANACOSTIA',
    'BLOCK': '1200 - 1299 BLOCK OF GOOD HOPE ROAD SE',
    'BLOCK_GROUP': '007503 1',
    'CCN': '20040945',
    'CENSUS_TRACT': '007503',
    'DISTRICT': '7',
    'END_DATE': 1583602253000,
    'LATITUDE': 38.8672656498,
    'LONGITUDE': -76.9878419796,
    'METHOD': 'OTHERS',
    'NEIGHBORHOOD_CLUSTER': 'Cluster 28',
    'OBJECTID': 175253942,
    'OCTO_RECORD_ID': None,
    'OFFENSE': 'THEFT/OTHER',
    'PSA': '701',
    'REPORT_DAT': 1583609902000,
    'SHIFT': 'DAY',
    'START_DATE': 1583596843000,
    'VOTING_PRECINCT': 'Precinct 114',
    'WARD': '8',
    'XBLOCK': 401055.1176309321,
    'YBLOCK': 133271.6069371067},
   'geometry': {'x': -76.98784425767002, 'y': 38.867273431546366}},
  {'attributes': {'ANC': '5E',
    'BID': None,
    'BLOCK': '1 - 99 BLOCK OF Q STREET NW',
    'BLOCK_GROUP': '004600 1',
    'CCN': '20040957',
    'CENSUS_TRACT': '004600',
    'DISTRICT': '3',
    'END_DATE': 1583604937000,
    'LATITUDE': 38.9111209077,
    'LONGITUDE': -77.0109900409,
    'METHOD': 'OTHERS',
    'NEIGHBORHOOD_CLUSTER': 'Cluster 21',
    'OBJECTID': 175253943,
    'OCTO_RECORD_ID': None,
    'OFFENSE': 'THEFT/OTHER',
    'PSA': '308',
    'REPORT_DAT': 1583616006000,
    'SHIFT': 'EVENING',
    'START_DATE': 1583497820000,
    'VOTING_PRECINCT': 'Precinct 19',
    'WARD': '5',
    'XBLOCK': 399046.83,
    'YBLOCK': 138139.87},
   'geometry': {'x': -77.01099232855701, 'y': 38.911128697745966}},
  {'alias': 'PSA', 'length': 3, 'name': 'PSA', 'type': 'esriFieldTypeString'},
  {'alias': 'NEIGHBORHOOD_CLUSTER',
   'length': 200,
   'name': 'NEIGHBORHOOD_CLUSTER',
   'type': 'esriFieldTypeString'},
  {'alias': 'BLOCK_GROUP',
   'length': 20,
   'name': 'BLOCK_GROUP',
   'type': 'esriFieldTypeString'},
  {'alias': 'CENSUS_TRACT',
   'length': 20,
   'name': 'CENSUS_TRACT',
   'type': 'esriFieldTypeString'},
  {'alias': 'VOTING_PRECINCT',
   'length': 25,
   'name': 'VOTING_PRECINCT',
   'type': 'esriFieldTypeString'},
  {'alias': 'LATITUDE', 'name': 'LATITUDE', 'type': 'esriFieldTypeDouble'},
  {'alias': 'LONGITUDE', 'name': 'LONGITUDE', 'type': 'esriFieldTypeDouble'},
  {'alias': 'BID',
   'length': 100,
   'name': 'BID',
   'type': 'esriFieldTypeString'},
  {'alias': 'START_DATE',
   'length': 8,
   'name': 'START_DATE',
   'type': 'esriFieldTypeDate'},
  {'alias': 'END_DATE',
   'length': 8,
   'name': 'END_DATE',
   'type': 'esriFieldTypeDate'},
  {'alias': 'OBJECTID', 'name': 'OBJECTID', 'type': 'esriFieldTypeOID'},
  {'alias': 'OCTO_RECORD_ID',
   'length': 25,
   'name': 'OCTO_RECORD_ID',
   'type': 'esriFieldTypeString'}],
 'geometryType': 'esriGeometryPoint',
 'spatialReference': {'latestWkid': 4326, 'wkid': 4326}}
```

In [ ]:
```python
for x in data_json:
    print(x)
```

```
displayFieldName
fieldAliases
geometryType
spatialReference
fields
features
exceededTransferLimit
```

In [ ]:
```python
print(data_json["features"])
```

```
[{'attributes': {'CCN': '20040945', 'REPORT_DAT': 1583609902000, 'SHIFT': 'DAY', 'METHOD': 'OTHERS', 'OFFENSE': 'THEFT/OTHER', 'BLOCK': '1200 - 1299 BLOCK OF
GOOD HOPE ROAD SE', 'XBLOCK': 401055.1176309321, 'YBLOCK': 133271.6069371067, 'WARD': '8', 'ANC': '8A', 'DISTRICT': '7', 'PSA': '701', 'NEIGHBORHOOD_CLUSTER':
'Cluster 28', 'BLOCK_GROUP': '007503 1', 'CENSUS_TRACT': '007503', 'VOTING_PRECINCT': 'Precinct 114', 'LATITUDE': 38.8672656498, 'LONGITUDE': -76.9878419796,
'BID': 'ANACOSTIA', 'START_DATE': 1583596843000, 'END_DATE': 1583602253000, 'OBJECTID': 175253942, 'OCTO_RECORD_ID': None}, 'geometry': {'x': -76.987844257670
02, 'y': 38.867273431546366}}, {'attributes': {'CCN': '20040957', 'REPORT_DAT': 1583616006000, 'SHIFT': 'EVENING', 'METHOD': 'OTHERS', 'OFFENSE': 'THEFT/OTHE
R', 'BLOCK': '1 - 99 BLOCK OF Q STREET NW', 'XBLOCK': 399046.83, 'YBLOCK': 138139.87, 'WARD': '5', 'ANC': '5E', 'DISTRICT': '3', 'PSA': '308', 'NEIGHBORHOOD_C
LUSTER': 'Cluster 21', 'BLOCK_GROUP': '004600 1', 'CENSUS_TRACT': '004600', 'VOTING_PRECINCT': 'Precinct 19', 'LATITUDE': 38.9111209077, 'LONGITUDE': -77.0109
900409, 'BID': None, 'START_DATE': 1583497820000, 'END_DATE': 1583604937000, 'OBJECTID': 175253943, 'OCTO_RECORD_ID': None}, 'geometry': {'x': -77.01099232855
701, 'y': 38.911128697745966}}, {'attributes': {'CCN': '20015748', 'REPORT_DAT': 1580089308000, 'SHIFT': 'EVENING', 'METHOD': 'OTHERS', 'OFFENSE': 'THEFT/OTHE
R', 'BLOCK': '600 - 699 BLOCK OF F STREET NW', 'XBLOCK': 398185.53, 'YBLOCK': 136610.53, 'WARD': '2', 'ANC': '2C', 'DISTRICT': '1', 'PSA': '101', 'NEIGHBORHOO
D_CLUSTER': 'Cluster 8', 'BLOCK_GROUP': '005801 2', 'CENSUS_TRACT': '005801', 'VOTING_PRECINCT': 'Precinct 129', 'LATITUDE': 38.8973427196, 'LONGITUDE': -77.0
209167811, 'BID': 'DOWNTOWN', 'START_DATE': 1579980618000, 'END_DATE': 1580075542000, 'OBJECTID': 175022918, 'OCTO_RECORD_ID': None}, 'geometry': {'x': -77.02
091907096168, 'y': 38.897350506437604}}]
```

In [ ]:
```python
df = pd.DataFrame(data_json['features'])
df.head()
```

Out[ ]:

	attributes	geometry
0	{'CCN': '20040945', 'REPORT_DAT': 158360990200...	{'x': -76.98784425767002, 'y': 38.867273431546...
1	{'CCN': '20040957', 'REPORT_DAT': 158361600600...	{'x': -77.01099232855701, 'y': 38.911128697745...
2	{'CCN': '20040978', 'REPORT_DAT': 158361453900...	{'x': -77.00588677394491, 'y': 38.90605178244354}
3	{'CCN': '20040993', 'REPORT_DAT': 158363898100...	{'x': -76.9984525859422, 'y': 38.83972359334092}
4	{'CCN': '20041028', 'REPORT_DAT': 158363841300...	{'x': -76.95607344162872, 'y': 38.888191980794...