

Name: - L Prathyusha

WEB CRAWLING

Link chosen on INDIA: -<https://en.wikipedia.org/wiki/India> (<https://en.wikipedia.org/wiki/India>).

In [3]:

```
from urllib.request import urlopen
from bs4 import BeautifulSoup as soup
import re
```

In [4]:

```
url = urlopen("https://en.wikipedia.org/wiki/India")
bsobj=soup(url.read())
```

In [5]:

```
url_process_stack = []
url_process_stack.append(bsobj)
```

EXTRA LINKS IN THE WEBPAGE

In [6]:

```
for link in bsobj.findAll('a'):
    if 'href' in link.attrs:
        print(link.attrs['href'])
```

```
/wiki/Wikipedia:Featured_articles
/wiki/Wikipedia:Protection_policy#extended
#mw-head
#searchInput
/wiki/India_(disambiguation)
/wiki/Bharat_(disambiguation)
/wiki/Names_of_India_in_its_official_languages
/wiki/File:Flag_of_India.svg
/wiki/Flag_of_India
/wiki/File:Emblem_of_India.svg
/wiki/State_Emblem_of_India
/wiki/Satyameva_Jayate
/wiki/Sanskrit_language
#cite_note-FOOTNOTENational_Informatics_Centre2005-1
/wiki/Jana_Gana_Manā
#cite_note-india.gov.in-2
#cite_note-tatsama-3
#cite_note-FOOTNOTEWolpert20031-4
#cite_note-india.gov.in-2
/wiki/India_Museum
```

In [7]:

```
bsobj.findAll('a')
```

Out[7]:

```
[<a id="top"></a>,
 <a href="/wiki/Wikipedia:Featured_articles" title="This is a featured article. Click here for more information."></a>,
 <a href="/wiki/Wikipedia:Protection_policy#extended" title="This article is extended-confirmed protected"></a>.
```

In [9]:

```
bsobj.find('div',{'id':'bodyContent'})
```

Out[9]:

```
<div class="mw-body-content" 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-content-ltr" dir="ltr" id="mw-content-text" lang="en"><div class="mw-parser-output"><div class="hatnote navigation-not-searchable" role="note">This article is about the Republic of India. For other uses, see <a class="mw-disambig" href="/wiki/India_(disambiguation)" title="India (disambiguation)">India (disambiguation)</a>.</div>
<div class="hatnote navigation-not-searchable" role="note">"Bharat" redirects here. For other uses, see <a class="mw-disambig" href="/wiki/Bharat_(disambiguation)" title="Bharat (disambiguation)">Bharat (disambiguation)</a>.</div>
<div class="mw-emntv-elt">
```

In [8]:

```
for link in bsobj.find('div',{'id':'bodyContent'}).findAll('a',href = re.compile('^(/wiki/))
    if 'href' in link.attrs:
        print(link.attrs['href'])
```

```
/wiki/India_(disambiguation)
/wiki/Bharat_(disambiguation)
/wiki/Names_of_India_in_its_official_languages
/wiki/File:Flag_of_India.svg
/wiki/Flag_of_India
/wiki/File:Emblem_of_India.svg
/wiki/State_Emblem_of_India
/wiki/Satyameva_Jayate
/wiki/Sanskrit_language
/wiki/Jana_Gana_Manā
/wiki/Vande_Mataram
/wiki/Sanskrit_language
/wiki/File:India_(orthographic_projection).svg
/wiki/New_Delhi
/wiki/Mumbai
/wiki/Delhi
/wiki/Hindi
/wiki/Indian_English
/wiki/Languages_with_official_status_in_India#State_level
/wiki/List_of_languages_spoken_in_India
```

In []:

In []:

```
from urllib.request import urlopen
from bs4 import BeautifulSoup as soup
import re
```

In [1]:

```
import requests
```

```
res = requests.get('https://forms.gle/UZginaXbxYviF3mw9')
```

```
print(res.text)
```

```
print(res.status_code)
```

```
<!DOCTYPE html>
<html lang="en">
  <head>
    <meta charset="utf-8">
    <meta content="width=300, initial-scale=1" name="viewport">
    <meta name="description" content="Create a new survey on your own or with others at the same time. Choose from a variety of survey types and analyse results in Google Forms. Free from Google.">
    <meta name="google-site-verification" content="LrdTUW9psUAMbh4Ia074-BPEVmcBxF6Gwf0MSgQXZs">
    <title>Google Forms - create and analyse surveys, for free.</title>
    <style>
      @font-face {
        font-family: 'Open Sans';
        font-style: normal;
        font-weight: 300;
        src: url(//fonts.gstatic.com/s/opensans/v15/mem5YaGs126MiZpBA-UN_r8OUuh
s.ttf) format('truetype');
      ,
```

In []: