An Assignment from Stirring Minds to find all the mode of Contacts by Web Scraping

For extracting the mode of contact we will perform the web scrapping

1 (a). importing the Important Libraries

```
In [2]: 1 import BeautifulSoup
        2 # By importing BeautifulSoup Soup we can parse and scrap the HTML and XML document
        3 equests
        4 # By Importing requests we can request to API and we can get the raw HTML contents from website i.e to scraping
        5 s
        6 # This library helps to intrect directly from Operating System
        7 ltk
        8 # By this we can import Natural Language toolkit
        9 k import corpus
        10 # By this we can import coprora reader which helps to handle specific corpus format
        11 e
        12 # By this we can import Regular Expression which will help to find  some specific things from scrap
```

1(b). First taking a URL and store it in 'url1'

```
In [45]: 1 url1='https://www.purpledocs.com/'
```

1(c). Making a request to access the URL

```
In [46]: 1 html_content1=requests.get(url1).text
```

In [47]: 1 html_content1

Out[47]: '<!DOCTYPE html>\n<html lang="en-US">\n<head>\n\t<meta charset="UTF-8">\n\t<meta name="viewport" content="width=device-width, initial-scale=1.0, viewport-fit=cover" />\t\t<link rel=\'preconnect\' href=\'https://fonts.googleapis.com\' />\n<link rel=\'preconnect\' href=\'https://fonts.gstatic.com\' />\n<meta http-equiv=\'x-dns-prefetch-control\' content=\'on\'>\n<link rel=\'dns-prefetch\' href=\'//fonts.googleapis.com\' />\n<link rel=\'dns-prefetch\' href=\'//fonts.gstatic.com\' />\n<link rel=\'dns-prefetch\' href=\'//s.gravatar.com\' />\n<link rel=\'dns-prefetch\' href=\'//www.google-analytics.com\' />\n<link rel=\'preload\' as=\'image\' href=\'https://www.purpledocs.com/wp-content/uploads/2021/10/PurpleDocs-Fav-Icon-P.png\' type=\'image/png\'>\n<link rel=\'preload\' as=\'font\' href=\'https://www.purpledocs.com/wp-content/themes/soledad/fonts/fontawesome-webfont.woff2?v=4.7.0\' type=\'font/woff2\' crossorigin=\'anonymous\' />\n<link rel=\'preload\' as=\'font\' href=\'https://www.purpledocs.com/wp-content/themes/soledad/fonts/weathericons.woff2\' type=\'font/woff2\' crossorigin=\'anonymous\' />\n<link rel=\'preload\' as=\'font\' href=\'https://www.purpledocs.com/wp-content/plugins/penci-recipe/fonts/ficon.woff2\' type=\'font/woff2\' crossorigin=\'anonymous\' />\n<meta name=\'robots\' content=\'index, follow, max-image-preview:large, max-snippet:-1, max-video-preview:-1\' />\n\n\t<!-- This site is optimized with the Yoast SEO plugin v17.6 - https://yoast.com/wordpress/plugins/seo/ (https://yoast.com/wordpress/plugins/seo/) -->\n\t<title>PurpleDocs- EMR, Patient Engagement, Medical Record Digitization</title>\n\t<meta name="description" content="PurpleDocs helps hospitals and doctors achieve their patient engagement goals through technology solutions & interactive services." />\n\t<link rel="canonical" href="https://www.purpledocs.com/" />\n\t<meta property="og:locale" content="en_US" />\n\t<meta property="og:type" content="website" />\n\t<meta property="og:title" content="PurpleDocs- EMR, Patient Engagement, M

here we get <Response [200]> that means we have succesfully accessed the URI

```
1  #html_content1=requests.get(url1).text
2  soup1=BeautifulSoup(html_content1,'html')  # here html is the type of document
3  print(soup1.prettify())   #prettify() is used to  Beautiful Soup parse tree into a nicely formatted Unicode
```

```
eicons/fonts/eicons.eot?5.10.0);src:url(https://www.purpledocs.com/wp-content/plugins/elementor/assets/lib/e
icons/fonts/eicons.eot?5.10.0#iefix) format("embedded-opentype"),url(https://www.purpledocs.com/wp-content/p
lugins/elementor/assets/lib/eicons/fonts/eicons.woff2?5.10.0) format("woff2"),url(https://www.purpledocs.co
m/wp-content/plugins/elementor/assets/lib/eicons/fonts/eicons.woff?5.10.0) format("woff"),url(https://www.pu
rpledocs.com/wp-content/plugins/elementor/assets/lib/eicons/fonts/eicons.ttf?5.10.0) format("truetype"),url
(https://www.purpledocs.com/wp-content/plugins/elementor/assets/lib/eicons/fonts/eicons.svg?5.10.0#eicon) fo
rmat("svg");font-weight:400;font-style:normal}
  </style>
  <link href="https://www.purpledocs.com/wp-content/uploads/elementor/css/post-312.css?ver=1637852600" id="e
lementor-post-312-css" media="all" rel="stylesheet" type="text/css"/>
  <link href="https://www.purpledocs.com/wp-content/plugins/elementor-pro/assets/css/frontend.min.css?ver=3.
0.10" id="elementor-pro-css" media="all" rel="stylesheet" type="text/css"/>
  <link href="https://www.purpledocs.com/wp-content/uploads/elementor/css/post-1540.css?ver=1637852614" id
="elementor-post-1540-css" media="all" rel="stylesheet" type="text/css"/>

  <link href="https://www.purpledocs.com/wp-content/uploads/elementor/css/post-4336.css?ver=1637852601" id
="elementor-post-4336-css" media="all" rel="stylesheet" type="text/css"/>
  <link href="https://www.purpledocs.com/wp-content/uploads/elementor/css/post-3704.css?ver=1637852601" id
="elementor-post-3704-css" media="all" rel="stylesheet" type="text/css"/>
  <link href="https://www.purpledocs.com/wp-content/uploads/elementor/css/post-375.css?ver=1637852601" id="e
lementor post 375 css" media "all" rel "stylesheet" type "text/css"/>
```

```
In [52]:  1  for link in soup1.find_all('a'):    # Here we can find all the contents stored in directory 'a'
          2      print('Inner Text:{}'.format(link.text))   # Here all text of link is stored in '{}' and assigned as 'Inr
          3      print('Title:{}'.format(link.get('title')))   # Here all title of link is stored in '{}' and assigned as
          4      print('href:{}'.format(link.get('href')))   #  # Here all hypertexts  of link is stored in '{}' and assig
          5          # and printing all this we will get Output as:-
```

```
Title:None
href:https://www.purpledocs.com/medical-record-digitization/
Inner Text:Chronic Care Management Program
Title:None
href:https://www.purpledocs.com/chronic-care-management-program/
Inner Text:Post-Discharge Engagement Program
Title:None
href:https://www.purpledocs.com/post-discharge-engagement-program/
Inner Text:Healthcare Solutions for Corporates
Title:None
href:https://www.purpledocs.com/health-solutions-for-corporate/
Inner Text:About
Title:None
href:https://www.purpledocs.com/about-us/
Inner Text:Contact
Title:None
href:https://www.purpledocs.com/contact-us/
Inner Text:Home
Title:None
href:https://www.purpledocs.com/
```

Note: By Doing the same we can Extract the data of 'n' number of pages but here we have done only for 5 URL

2. Taking a second URL and store it in 'url2'

```
In [10]:  1  url2='https://www.wimerasys.com/'
```

```python
1  html_content2=requests.get(url2).text
2  soup2=BeautifulSoup(html_content2,'lxml')
3  print(soup2.prettify())
```

```
<script src="https://www.wimerasys.com/wp-includes/js/jquery/jquery.js?ver=1.12.4-wp" type="text/javascrip
t">
</script>
<script src="https://www.wimerasys.com/wp-includes/js/jquery/jquery-migrate.min.js?ver=1.4.1" type="text/j
avascript">
</script>
<script src="https://www.wimerasys.com/wp-includes/js/jquery/ui/core.min.js?ver=1.11.4" type="text/javascr
ipt">
</script>
<script src="https://www.wimerasys.com/wp-includes/js/jquery/ui/widget.min.js?ver=1.11.4" type="text/javas
cript">
</script>
<script src="https://www.wimerasys.com/wp-includes/js/jquery/ui/mouse.min.js?ver=1.11.4" type="text/javasc
ript">
</script>
<script src="https://www.wimerasys.com/wp-includes/js/jquery/ui/draggable.min.js?ver=1.11.4" type="text/ja
vascript">
</script>
<script src="https://www.wimerasys.com/wp-includes/js/jquery/ui/slider.min.js?ver=1.11.4" type="text/javas
cript">
```

```
In [12]:   1  for link in soup2.find_all('a'):
           2      print('Inner Text:{}'.format(link.text))
           3      print('Title:{}'.format(link.get('title')))
           4      print('href:{}'.format(link.get('href')))
```

href:https://www.wimerasys.com/about-us/
Inner Text:
Facebook

Title:None
href:https://www.facebook.com/WimeraSys
Inner Text:
Twitter

Title:None
href:https://twitter.com/systemswimera?lang=en
Inner Text:
Linkedin

Title:None
href:https://www.linkedin.com/company/wimera
Inner Text:Top
Title:Scroll to Top
href:#

3. Taking a second URL and store it in 'url3'

```
In [13]:   1  url3='https://signzy.com/'
           2  html_content3=requests.get(url3).text
           3  soup3=BeautifulSoup(html_content3,'lxml')
           4  print(soup3.prettify())
```

<!DOCTYPE html>
<html class="no-js no-svg" lang="en-US" prefix="og: https://ogp.me/ns#"> (https://ogp.me/ns#">)
 <head>
  <meta charset="utf-8"/>
  <meta content="width=device-width, initial-scale=1" name="viewport"/>
  <link href="http://gmpg.org/xfn/11" rel="profile"/>
  <!--[if lt IE 9]>
              <script src="https://signzy.com/wp-content/themes/signzy/assets/js/html5.js"></script>
              <![endif]-->
  <link href="https://signzy.com/wp-content/themes/signzy/assets/css/custom.css" rel="stylesheet" type="tex
t/css"/>
  <title>
   AI | Digital Onboarding | Video KYC | ID checks
  </title>
  <!-- All in One SEO 4.1.2.2 -->
  <link href="https://signzy.com/" rel="canonical"/>
  <meta content="Signzy" property="og:site_name"/>
  <meta content="article" property="og:type"/>
  <meta content="AI | Digital Onboarding | Video KYC | ID checks" property="og:title"/>
```

```
In [14]:   1  for link in soup3.find_all('a'):
           2      print('Inner Text:{}'.format(link.text))
           3      print('Title:{}'.format(link.get('title')))
           4      print('href:{}'.format(link.get('href')))
```

```
href:https://www.facebook.com/TeamSignzy/
Inner Text:


Twitter

Title:Twitter
href:https://twitter.com/TeamSignzy
Inner Text:

AngelCo

Title:AngelCo
href:https://angel.co/signzy/
Inner Text:

Instagram

Title:Instagram
href:https://www.instagram.com/teamsignzy/
```

4. Taking a second URL and store it in 'url4'

```
1  url4='https://www.smartjoules.co.in/'
2  html_content4=requests.get(url4).text
3  soup4=BeautifulSoup(html_content4,'lxml')
4  print(soup4.prettify())
```

t}.search_submit input{background:#fb8c00!important;color:#000!important;font-weight:bold!important;border-r
adius:5px!important;border:none!important}.single_job_listing .meta .job-type{background:transparent!importa
nt;font-weight:bold}.job_filters .job_types li{list-style-type:none!important}.single_job_listing .job-listi
ng-meta li{list-style-type:none!important}.post-content .post-meta{display:none}.meta .location a{color:#00
0!important}.job-listing-meta .date-posted{color:#000!important}.job_application .application_button{backgro
und:#fb8c00!important;color:#000!important;font-weight:bold!important;border-radius:5px!important;border:non
e!important}.share-options a{color:#fe7e17!important}.page-title-subtitle{margin-bottom:0!important}.job_lis
tings a .date{display:none!important}.single_job_listing .date-posted{display:none!important}#header{padding
-top:16px;padding-bottom:5px}
  </style>
  <link href="https://www.smartjoules.co.in/wp-content/themes/patti/css/prettyPhoto.css?ver=5.4.8" id="dt-pr
ettyphoto-lightbox-css" media="all" rel="stylesheet" type="text/css"/>
  <link href="https://www.smartjoules.co.in/wp-content/themes/patti/css/superfish.css?ver=5.4.8" id="dt-supe
rfish-css" media="all" rel="stylesheet" type="text/css"/>
  <link href="https://www.smartjoules.co.in/wp-content/themes/patti/framework/fonts/font-awesome/css/font-aw
esome.css?ver=5.4.8" id="dt-font-awesome-css" media="all" rel="stylesheet" type="text/css"/>
  <link href="https://www.smartjoules.co.in/wp-content/themes/patti/css/audioplayer.css?ver=5.4.8" id="dt-au
dioplayer-css" media="all" rel="stylesheet" type="text/css"/>
  <link href="https://www.smartjoules.co.in/wp-content/themes/patti/css/owl.carousel.css?ver=5.4.8" id="dt-o
wl-carousel-css" media="all" rel="stylesheet" type="text/css"/>

```
In [16]:    1  for link in soup4.find_all('a'):
            2      print('Inner Text:{}'.format(link.text))
            3      print('Title:{}'.format(link.get('title')))
            4      print('href:{}'.format(link.get('href')))
```

Title:None
href:http://www.berkeley.edu/
Inner Text:Tweets by @SmartJoules
Title:None
href:https://twitter.com/SmartJoules
Inner Text:info@smartjoules.in
Title:None
href:mailto:info@smartjoules.in
Inner Text:https://www.facebook.com/SmartJoulesIndia/
Title:None
href:https://www.facebook.com/SmartJoulesIndia/
Inner Text:
Title:None
href:https://unreasonablegroup.com/c/?id=5c2de5f4-9476-4f4f-b92e-331af50f08f3
Inner Text:Locate us on Map
Title:None
href:None
Inner Text:
Title:None
href:#

5. Taking a second URL and store it in 'url5'

```python
In [17]:   1  url5='https://www.indusos.com//'
           2  html_content5=requests.get(url5).text
           3  soup5=BeautifulSoup(html_content5,'lxml')
           4  print(soup5.prettify())
```

```
...1 /7
  <link href="https://www.indusos.com/comments/feed/" rel="alternate" title="Indus OS » Comments Feed" type
="application/rss+xml"/>
  <script type="text/javascript">
   window._wpemojiSettings = {"baseUrl":"https:\/\/s.w.org\/images\/core\/emoji\/13.0.1\/72x72\/","ext":".pn
g","svgUrl":"https:\/\/s.w.org\/images\/core\/emoji\/13.0.1\/svg\/","svgExt":".svg","source":{"concatemoj
i":"https:\/\/www.indusos.com\/wp-includes\/js\/wp-emoji-release.min.js?ver=5.6.4"}};
                        !function(e,a,t){var n,r,o,i=a.createElement("canvas"),p=i.getContext&&i.getContext
("2d");function s(e,t){var a=String.fromCharCode;p.clearRect(0,0,i.width,i.height),p.fillText(a.apply(this,
e),0,0);e=i.toDataURL();return p.clearRect(0,0,i.width,i.height),p.fillText(a.apply(this,t),0,0),e===i.toDat
aURL()}function c(e){var t=a.createElement("script");t.src=e,t.defer=t.type="text/javascript",a.getElementsB
yTagName("head")[0].appendChild(t)}for(o=Array("flag","emoji"),t.supports={everything:!0,everythingExceptFla
g:!0},r=0;r<o.length;r++)t.supports[o[r]]=function(e){if(!p||!p.fillText)return!1;switch(p.textBaseline="to
p",p.font="600 32px Arial",e){case"flag":return s([127987,65039,8205,9895,65039],[127987,65039,8203,9895,650
39])?!1:!s([55356,56826,55356,56819],[55356,56826,8203,55356,56819])&&!s([55356,57332,56128,56423,56128,5641
8,56128,56421,56128,56430,56128,56423,56128,56447],[55356,57332,8203,56128,56423,8203,56128,56418,8203,5612
8,56421,8203,56128,56430,8203,56128,56423,8203,56128,56447]);case"emoji":return!s([55357,56424,8205,55356,57
212],[55357,56424,8203,55356,57212])}return!1}(o[r]),t.supports.everything=t.supports.everything&&t.supports
[o[r]],"flag"!==o[r]&&(t.supports.everythingExceptFlag=t.supports.everythingExceptFlag&&t.supports[o[r]]);t.
supports.everythingExceptFlag=t.supports.everythingExceptFlag&&!t.supports.flag,t.DOMReady=!1,t.readyCallbac
k function()[t DOMReady !0] t supports everything||(p function()[t readyCallback()] a addEventListener](a ad
```

```
1  for link in soup5.find_all('a'):
2      print('Inner Text:{}'.format(link.text))
3      print('Title:{}'.format(link.get('title')))
4      print('href:{}'.format(link.get('href')))
```

```
Inner Text:We are hiring!
Title:None
href:/careers
Inner Text:×
Title:None
href:#
Inner Text:
                        Skip to content
Title:None
href:#content
Inner Text:
Title:Indus OS
href:https://www.indusos.com/
Inner Text:Home
Title:None
href:https://www.indusos.com/
Inner Text:About us
Title:None
href:https://www.indusos.com/about-us/
Inner Text:Indus App Bazaar
Title:None
href:https://www.indusos.com/indus-app-bazaar/
Inner Text:Blog
Title:None
href:https://www.indusos.com/blog/
Inner Text:Developer Login
Title:None
href:https://developer.appbazaar.com/
Inner Text:
Title:Indus OS
href:https://www.indusos.com/
Inner Text:Home
Title:None
href:https://www.indusos.com/
Inner Text:About us
Title:None
href:https://www.indusos.com/about-us/
```

Inner Text:Indus App Bazaar
Title:None
href:https://www.indusos.com/indus-app-bazaar/
Inner Text:Blog
Title:None
href:https://www.indusos.com/blog/
Inner Text:Developer Login
Title:None
href:https://developer.appbazaar.com/
Inner Text:Learn More
Title:None
href:/indus-app-bazaar/
Inner Text:Learn More
Title:None
href:/about-us/
Inner Text:Indus App Bazaar
Title:None
href:#
Inner Text:Indus One
Title:None
href:#
Inner Text:Prisma
Title:None
href:#
Inner Text:Learn More
Title:None
href:/indus-app-bazaar/
Inner Text:Learn More
Title:None
href:/indus-one/
Inner Text:Learn More
Title:None
href:/prisma/
Inner Text:
Title:Need For App Fairness, Choice and Innovation In the App Store Ecosystem
href:https://www.indusos.com/blog/need-for-app-fairness-choice-and-innovation-in-the-app-store-ecosystem/
Inner Text:Need For App Fairness, Choice and Innovation In the App Store Ecosystem
Title:None
href:https://www.indusos.com/blog/need-for-app-fairness-choice-and-innovation-in-the-app-store-ecosystem/
Inner Text:October 7, 2021November 17, 2021
Title:5:46 pm
href:https://www.indusos.com/blog/need-for-app-fairness-choice-and-innovation-in-the-app-store-ecosystem/
Inner Text:Read more →

Title:Need For App Fairness, Choice and Innovation In the App Store Ecosystem
href:https://www.indusos.com/blog/need-for-app-fairness-choice-and-innovation-in-the-app-store-ecosystem/
Inner Text:
Title:Indus OS Celebrates 200 Million Users
href:https://www.indusos.com/blog/indus-os-celebrates-200-million-users/
Inner Text:Indus OS Celebrates 200 Million Users
Title:None
href:https://www.indusos.com/blog/indus-os-celebrates-200-million-users/
Inner Text:September 24, 2021November 17, 2021
Title:6:40 pm
href:https://www.indusos.com/blog/indus-os-celebrates-200-million-users/
Inner Text:Read more →
Title:Indus OS Celebrates 200 Million Users
href:https://www.indusos.com/blog/indus-os-celebrates-200-million-users/
Inner Text:
Title:Indus App Bazaar encourages developers to build local solutions for their audience
href:https://www.indusos.com/media-story/indus-app-bazaar-encourages-developers-to-build-local-solutions-for-their-audience/
Inner Text:Indus App Bazaar encourages developers to build local solutions for their audience
Title:None
href:https://www.indusos.com/media-story/indus-app-bazaar-encourages-developers-to-build-local-solutions-for-their-audience/
Inner Text:August 21, 2021August 24, 2021
Title:7:56 pm
href:https://www.indusos.com/media-story/indus-app-bazaar-encourages-developers-to-build-local-solutions-for-their-audience/
Inner Text:Read more →
Title:Indus App Bazaar encourages developers to build local solutions for their audience
href:https://www.indusos.com/media-story/indus-app-bazaar-encourages-developers-to-build-local-solutions-for-their-audience/
Inner Text:+9122 45101451
Title:None
href:tel:02245101451
Inner Text:About Us
Title:None
href:/about-us/
Inner Text:Media
Title:None
href:/media/
Inner Text:Corporate
Title:None
href:/corporate-information
Inner Text:Indus App Bazaar

```
Title:None
href:/indus-app-bazaar/
Inner Text:Indus One
Title:None
href:/indus-one/
Inner Text:Prisma
Title:None
href:/prisma/
Inner Text:Careers
Title:None
href:/careers/
Inner Text:FAQs
Title:None
href:/faq/
Inner Text:Contact Us
Title:None
href:/contact-us/
Inner Text:Privacy Policy
Title:None
href:/privacy-policy-and-terms-of-use/
Inner Text:
Title:None
href:https://www.facebook.com/indusos/
Inner Text:
Title:None
href:https://mobile.twitter.com/indusos
Inner Text:
Title:None
href:https://www.linkedin.com/company/indusos/
Inner Text:
Title:None
href:https://www.youtube.com/channel/UCPJqw2f-WCc0SmLuZl2Rhew
Inner Text:
Title:None
href:https://www.instagram.com/indusappbazaar/
Inner Text:Terms of Use
Title:None
href:/privacy-policy-and-terms-of-use/
```

6. Now what ever the information of mode of contact we got from above scraps we can populate in a saperate tablualar chart.

```python
In [53]:  1  # For making a Table We have to import pandas library
          2  import pandas as pd
          3
```

```python
In [54]:  1  table=pd.DataFrame({'websites':[url1,url2,url3,url4,url5],
          2                      'Twitter':['twitter.com/purple_docs',
          3                                 'twitter.com/systemswimera',
          4                                 'twitter.com/TeamSignzy',
          5                                 'twitter.com/SmartJoules',
          6                                 'mobile.twitter.com/indusos'],
          7                      'Contact_no':['919909903539',
          8                                    '',
          9                                    '',
         10                                    '+91- 11- 46046513',
         11                                    '+9122 45101451'],
         12                      'Email':['',
         13                               '',
         14                               '',
         15                               'info@smartjoules.in',
         16                               'grievances@indusos.com'],
         17                      "Instagram":['https://www.linkedin.com/company/purpledocshealthtech/',
         18                                   'https://www.linkedin.com/company/wimera',
         19                                   'https://www.linkedin.com/company/teamsignzy/',
         20                                   '',
         21                                   'https://www.instagram.com/indusappbazaar/'],
         22                      "Facebook":['https://www.facebook.com/PurpleDocs/',
         23                                  'https://www.facebook.com/WimeraSys',
         24                                  'https://www.facebook.com/TeamSignzy/',
         25                                  'https://www.facebook.com/SmartJoulesIndia/',
         26                                  'https://www.facebook.com/indusos/']})
```

In [55]:

```
1 table
```

Out[55]:

	websites	Twitter	Contact_no	Email	Instagra
0	https://www.purpledocs.com/	twitter.com/purple_docs	919909903539		https://www.linkedin.com/company/purpledocshea
1	https://www.wimerasys.com/	twitter.com/systemswimera			https://www.linkedin.com/company/wime
2	https://signzy.com/	twitter.com/TeamSignzy			https://www.linkedin.com/company/teamsignz
3	https://www.smartjoules.co.in/	twitter.com/SmartJoules	+91- 11-46046513	info@smartjoules.in	
4	https://www.indusos.com//	mobile.twitter.com/indusos	+9122 45101451	grievances@indusos.com	https://www.instagram.com/indusappbaza

Submitted by: Ashwin Gorakhnath Dubey

Submitted to: Stirring Minds

Date: 27-nov-2021