In [1]:

```
#importing the important libraries for the web scraping
import requests
from bs4 import BeautifulSoup
```

In [2]:

```
#Here we are using request.get to take all information in our url
res = requests.get('https://news.ycombinator.com/news')
res2 = requests.get('https://news.ycombinator.com/news?p=2')
```

In [3]:

```
#To know the response we would get from our server
print(res)
print(res2)
```

```
<Response [200]>
<Response [200]>
```

In [4]:

```
#using beautifulsoup to parse the html
soup = BeautifulSoup(res.text,'html.parser')
soup2 = BeautifulSoup(res2.text,'html.parser')
```

In [5]:

```
# using Beautifulsoup method to show the content of the page
soup.body.contents
```

Out[5]:

```
[<center><table bgcolor="#f6f6ef" border="0" cellpadding="0" cellspacing
="0" id="hnmain" width="85%">
 <tr><td bgcolor="#ff6600"><table border="0" cellpadding="0" cellspacing
="0" style="padding:2px" width="100%"><tr><td style="width:18px;padding-ri
ght:4px"><a href="https://news.ycombinator.com"><img height="18" src="y18.
gif" style="border:1px white solid;" width="18"/></a></td>
 <td style="line-height:12pt; height:10px;"><span class="pagetop"><b class
="hnname"><a href="news">Hacker News</a></b>
 <a href="newest">new</a> | <a href="front">past</a> | <a href="newcomment
s">comments</a> | <a href="ask">ask</a> | <a href="show">show</a> | <a hre
f="jobs">jobs</a> | <a href="submit">submit</a> </span></td><td style="tex
t-align:right;padding-right:4px;"><span class="pagetop">
 <a href="login?goto=news">login</a>
 </span></td>
 </tr></table></td></tr>
 <tr id="pagespace" style="height:10px" title=""></tr><tr><td><table borde
r="0" cellpadding="0" cellspacing="0" class="itemlist">
```

In [6]:

```
#using Beautifulsoup findall method to find the div in our page
soup2.find_all('div')
```

Out[6]:

```
[<div class="votearrow" title="upvote"></div>,
 <div class="votearrow" title="upvote"></div>,
 <div class="votearrow" title="upvote"></div>,
 <div class="votearrow" title="upvote"></div>,
 <div class="votearrow" title="upvote"></div>,
 <div class="votearrow" title="upvote"></div>,
 <div class="votearrow" title="upvote"></div>,
 <div class="votearrow" title="upvote"></div>,
 <div class="votearrow" title="upvote"></div>,
 <div class="votearrow" title="upvote"></div>,
 <div class="votearrow" title="upvote"></div>,
 <div class="votearrow" title="upvote"></div>,
 <div class="votearrow" title="upvote"></div>,
 <div class="votearrow" title="upvote"></div>,
 <div class="votearrow" title="upvote"></div>,
 <div class="votearrow" title="upvote"></div>,
 <div class="votearrow" title="upvote"></div>,
```

In [7]:

```
#using the title method to know the title of our page
soup2.title
```

Out[7]:

```
<title>Hacker News</title>
```

In [8]:

```
#Getting all the score of each news posted in our page
soup.select('.score')
```

```
 <span class="score" id="score_31770348">47 points</span>,
 <span class="score" id="score_31769059">126 points</span>,
 <span class="score" id="score_31767830">159 points</span>,
 <span class="score" id="score_31769294">231 points</span>,
 <span class="score" id="score_31768705">85 points</span>,
 <span class="score" id="score_31768936">78 points</span>,
 <span class="score" id="score_31769484">277 points</span>,
 <span class="score" id="score_31766179">187 points</span>,
 <span class="score" id="score_31748884">16 points</span>,
 <span class="score" id="score_31767858">87 points</span>,
 <span class="score" id="score_31762670">7 points</span>,
 <span class="score" id="score_31756451">46 points</span>,

 <span class="score" id="score_31766285">121 points</span>,
 <span class="score" id="score_31742755">77 points</span>,
 <span class="score" id="score_31767575">97 points</span>,
 <span class="score" id="score_31770618">17 points</span>,
 <span class="score" id="score_31769520">112 points</span>,
 <span class="score" id="score_31764722">161 points</span>,
 <span class="score" id="score_31746967">17 points</span>,
```

In [9]:

```
# using the select method to get the title links
soup.select('.titlelink')
```

```
 <a class="titlelink" href="https://github.com/dspinellis/unix-history-rep
o">Continuous Unix commit history from 1970 until today</a>,
 <a class="titlelink" href="https://www.nytimes.com/2022/06/14/science/cat
s-catnip-insects.html">Chewed and Rolled: How Cats Make the Most of Their
Catnip High</a>,
 <a class="titlelink" href="https://www.ycombinator.com/companies/keeper-t
ax/jobs/RsXu0Yl-product-designer-design-systems-contract" rel="nofollow">K
eeper Tax (YC W19) Is Hiring a freelance designer (design systems)</a>,
 <a class="titlelink" href="https://telegra.ph/No-QuestDB-is-not-Faster-th
an-ClickHouse-06-15">No, QuestDB is not Faster than ClickHouse</a>,
 <a class="titlelink" href="https://hazlitt.net/longreads/other-french-che
f">The (Other) French Chef</a>,
 <a class="titlelink" href="https://www.atlasobscura.com/articles/the-beau
tiful-network-of-ancient-roman-roads">The Beautiful Network of Ancient Rom
an Roads (2015)</a>,
 <a class="titlelink" href="https://trevorklee.substack.com/p/why-im-now-m
aking-drugs-for-cats">I'm making drugs for cats</a>,
 <a class="titlelink" href="https://neurosciencenews.com/chemical-fetal-br
ain-20810/">Endocrine-disrupting chemical exposure in womb impact fear, an
xiety behavior</a>,
```

In [10]:

```
#using the beautifulsoup select method to grab the title of the first news in our website p
soup2.select('.titlelink')[0]
```

Out[10]:

```
<a class="titlelink" href="https://readyset.io/blog/readyset-core">ReadySet
Core: next-generation SQL caching, freely available</a>
```

In [11]:

```
# Getting the 1st page and 2nd page news title with their scores
links = soup.select('.titlelink')
links2 = soup2.select('.titlelink')
votes = soup.select('.score')
votes2 = soup2.select('.score')
print(links[0])
print(links2[0])
print(votes[0])
print(votes2[0])
```

```
<a class="titlelink" href="https://justine.lol/redbean2/">Redbean 2.0 turned
into more than a hobby project</a>
<a class="titlelink" href="https://readyset.io/blog/readyset-core">ReadySet
Core: next-generation SQL caching, freely available</a>
<span class="score" id="score_31764521">382 points</span>
<span class="score" id="score_31767827">85 points</span>
```

In [12]:

```
# Getting the titles of all the news in our website page
def create_custom_hn(links, votes):
    hn =[]
    for idx, item in enumerate(links):
        title = links[idx].getText()
        href =  links[idx].get('href', None)
        hn.append(title)
    return hn
create_custom_hn(links, votes)
```

Out[12]:

```
['Redbean 2.0 turned into more than a hobby project',
 'Tell HN: Triplebyte is, yet again, making user profiles public without con
sent',
 'Cool Desktops Don’t Change',
 'Apple Reneged on OCSP Privacy',
 'Calling for Antitrust Reform',
 'The silent majority of experts (2012)',
 'macOS Screenshot Tricks to Impress Your Co-Workers',
 'Citus: Make enterprise features open source',
 'The Animated Elliptic Curve',
 "Show HN: Recut automatically removes silence from videos. It's built with
Tauri",
 "The computers are fast, but you don't know it",
 'CVE-2022-23088: Exploiting a Heap Overflow in the FreeBSD Wi-Fi Stack',
 'Wasmer – The Universal WebAssembly Runtime',
 'The Cult Inside Google',
 'Continuous Unix commit history from 1970 until today',
 'Chewed and Rolled: How Cats Make the Most of Their Catnip High',
 'Keeper Tax (YC W19) Is Hiring a freelance designer (design systems)',
 'No, QuestDB is not Faster than ClickHouse',
 'The (Other) French Chef',
 'The Beautiful Network of Ancient Roman Roads (2015)',
 "I'm making drugs for cats",
 'Endocrine-disrupting chemical exposure in womb impact fear, anxiety behavi
or',
 'Safari on iOS can overlap multiple full-screen videos',
 "CockroachDB's Consistency Model",
 'GitHub waited 3 months to notify about potential compromise',
 'How fast can a 6502 transfer memory?',
 'The Phone Booth of the Mind',
 'Tesla sends untrained employees to work on cars as service becomes problem
atic',
 'NextDNS API',
 'Nanoparticle sensor can distinguish between viral and bacterial pneumoni
a']
```

In [13]:

```
#Getting the links of each news in our website page
def create_custom_hn(links, votes):
    hn =[]
    for idx, item in enumerate(links):
        title = links[idx].getText()
        href =  links[idx].get('href', None)
        hn.append(links)
    return hn
create_custom_hn(links, votes)
```

Out[13]:

```
[[<a class="titlelink" href="https://justine.lol/redbean2/">Redbean 2.0 tu
rned into more than a hobby project</a>,
  <a class="titlelink" href="item?id=31769601">Tell HN: Triplebyte is, yet
again, making user profiles public without consent</a>,
  <a class="titlelink" href="https://tylercipriani.com/blog/2022/06/15/cho
ose-boring-desktop-technology/">Cool Desktops Don’t Change</a>,
  <a class="titlelink" href="https://mjtsai.com/blog/2022/06/16/apple-rene
ged-on-ocsp-privacy/">Apple Reneged on OCSP Privacy</a>,
  <a class="titlelink" href="https://blog.mozilla.org/en/mozilla/calling-f
or-antitrust-reform/">Calling for Antitrust Reform</a>,
  <a class="titlelink" href="https://prog21.dadgum.com/143.html">The silen
t majority of experts (2012)</a>,
  <a class="titlelink" href="https://sal.dev/macos/macos-screenshotting-ti
ps-and-tricks/">macOS Screenshot Tricks to Impress Your Co-Workers</a>,
  <a class="titlelink" href="https://github.com/citusdata/citus/commit/184
c7c0bce6b7bca61d25b828855fac5fba64816">Citus: Make enterprise features ope
n source</a>,
```

In [14]:

```
#Getting the titles and the links of the news in our website page
def create_custom_hn(links, votes2):
    hn = []
    for idx, item in enumerate(links):
        title = links[idx].getText()
        href = links[idx].get('href', None)
        hn.append({'title':title, 'links':href })
    return hn
create_custom_hn(links, votes2)
```

Out[14]:

```
[{'title': 'Redbean 2.0 turned into more than a hobby project',
  'links': 'https://justine.lol/redbean2/'},
 {'title': 'Tell HN: Triplebyte is, yet again, making user profiles public
without consent',
  'links': 'item?id=31769601'},
 {'title': 'Cool Desktops Don’t Change',
  'links': 'https://tylercipriani.com/blog/2022/06/15/choose-boring-deskto
p-technology/'},
 {'title': 'Apple Reneged on OCSP Privacy',
  'links': 'https://mjtsai.com/blog/2022/06/16/apple-reneged-on-ocsp-priva
cy/'},
 {'title': 'Calling for Antitrust Reform',
  'links': 'https://blog.mozilla.org/en/mozilla/calling-for-antitrust-refo
rm/'},
 {'title': 'The silent majority of experts (2012)',
  'links': 'https://prog21.dadgum.com/143.html'},
 {'title': 'macOS Screenshot Tricks to Impress Your Co-Workers',
```

In [15]:

```
#Getting the links, titles and scores of the news with  scores greater than 99 in our websi
#1st/2nd page
subtext = soup.select('.subtext')
subtext2 = soup2.select('.subtext')
mega_links = links + links2
mega_subtext = subtext + subtext2
def sort_stories_by_votes(hnlist):
    return sorted(hnlist, key = lambda k:k['votes'], reverse = True)

def create_custom_hn(mega_links, mega_subtext):
    hn = []
    for idx, item in enumerate(links):
        title = item.getText()
        href = item.get('href', None)
        vote = subtext[idx].select('.score')
        if len(vote):
            points = int(vote[0].getText().replace('points', ' '))
            if points > 99:
                hn.append({'title':title, 'links':href, 'votes' : points})
    return sort_stories_by_votes(hn)
create_custom_hn(mega_links, mega_subtext)
```

Out[15]:

```
[{'title': 'Tell HN: Triplebyte is, yet again, making user profiles public w
ithout consent',
  'links': 'item?id=31769601',
  'votes': 480},
 {'title': 'Redbean 2.0 turned into more than a hobby project',
  'links': 'https://justine.lol/redbean2/',
  'votes': 382},
 {'title': 'The Cult Inside Google',
  'links': 'https://medium.com/@kwilliamlloyd/the-cult-in-google-3c1a910214d
1',
  'votes': 277},
 {'title': "The computers are fast, but you don't know it",
  'links': 'http://shvbsle.in/computers-are-fast-but-you-dont-know-it-p1/',
  'votes': 231},
 {'title': 'The silent majority of experts (2012)',
  'links': 'https://prog21.dadgum.com/143.html',
  'votes': 191},
 {'title': 'Continuous Unix commit history from 1970 until today',
  'links': 'https://github.com/dspinellis/unix-history-repo',
  'votes': 187},
 {'title': 'How fast can a 6502 transfer memory?',
  'links': 'https://imapenguin.com/how-fast-can-a-6502-transfer-memory/',
  'votes': 161},
 {'title': "Show HN: Recut automatically removes silence from videos. It's b
uilt with Tauri",
  'links': 'https://getrecut.com/',
  'votes': 159},
 {'title': 'Cool Desktops Don’t Change',
  'links': 'https://tylercipriani.com/blog/2022/06/15/choose-boring-desktop-
technology/',
  'votes': 156},
 {'title': 'macOS Screenshot Tricks to Impress Your Co-Workers',
  'links': 'https://sal.dev/macos/macos-screenshotting-tips-and-tricks/',
  'votes': 137},
 {'title': 'The Animated Elliptic Curve',
```

```
    'links': 'https://curves.ulfheim.net/',
    'votes': 126},
   {'title': "I'm making drugs for cats",
    'links': 'https://trevorklee.substack.com/p/why-im-now-making-drugs-for-ca
  ts',
    'votes': 121},
   {'title': 'NextDNS API',
    'links': 'https://nextdns.github.io/api/',
    'votes': 114},
   {'title': 'GitHub waited 3 months to notify about potential compromise',
    'links': 'item?id=31769520',
    'votes': 112}]
```