BEE2041 Empirical Project Blog

In [1]:
```
pip install requests beautifulsoup4
```

```
Requirement already satisfied: requests in c:\users\socor\anaconda3\li
b\site-packages (2.24.0)
Requirement already satisfied: beautifulsoup4 in c:\users\socor\anacon
da3\lib\site-packages (4.9.1)
Requirement already satisfied: chardet<4,>=3.0.2 in c:\users\socor\ana
conda3\lib\site-packages (from requests) (3.0.4)
Requirement already satisfied: idna<3,>=2.5 in c:\users\socor\anaconda
3\lib\site-packages (from requests) (2.10)
Requirement already satisfied: certifi>=2017.4.17 in c:\users\socor\an
aconda3\lib\site-packages (from requests) (2020.6.20)
Requirement already satisfied: urllib3!=1.25.0,!=1.25.1,<1.26,>=1.21.1
in c:\users\socor\anaconda3\lib\site-packages (from requests) (1.25.9)
Requirement already satisfied: soupsieve>1.2 in c:\users\socor\anacond
a3\lib\site-packages (from beautifulsoup4) (2.0.1)
Note: you may need to restart the kernel to use updated packages.
```

In [37]:
```
import requests

url = 'https://news.ycombinator.com/'
response = requests.get(url)
html_content = response.text
html_headers = response.headers
html_headers
```

Out[37]:
```
<html lang="en" op="news"><head><meta content="origin" name="referr
er"/><meta content="width=device-width, initial-scale=1.0" name="vi
ewport"/><link href="news.css?i3TCm9mQaQeIHjM4t2Io" rel="styleshee
t" type="text/css"/>
<link href="y18.svg" rel="icon"/>
<link href="rss" rel="alternate" title="RSS" type="application/rss+
xml"/>
<title>Hacker News</title></head><body><center><table bgcolor="#f6f
6ef" border="0" cellpadding="0" cellspacing="0" id="hnmain" width
="85%">
<tr><td bgcolor="#ff6600"><table border="0" cellpadding="0" cellspa
cing="0" style="padding:2px" width="100%"><tr><td style="width:18p
x;padding-right:4px"><a href="https://news.ycombinator.com"><img he
ight="18" src="y18.svg" style="border:1px white solid; display:bloc
k" 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="newc
omments">comments</a> | <a href="ask">ask</a> | <a href="show">show
```

In [20]:
```
soup = BeautifulSoup(html_content, 'html.parser')
soup
```

Out[20]:
```
<html lang="en" op="news"><head><meta content="origin" name="referr
er"/><meta content="width=device-width, initial-scale=1.0" name="vi
ewport"/><link href="news.css?i3TCm9mQaQeIHjM4t2Io" rel="styleshee
t" type="text/css"/>
<link href="y18.svg" rel="icon"/>
<link href="rss" rel="alternate" title="RSS" type="application/rss+
xml"/>
<title>Hacker News</title></head><body><center><table bgcolor="#f6f
6ef" border="0" cellpadding="0" cellspacing="0" id="hnmain" width
="85%">
<tr><td bgcolor="#ff6600"><table border="0" cellpadding="0" cellspa
cing="0" style="padding:2px" width="100%"><tr><td style="width:18p
x;padding-right:4px"><a href="https://news.ycombinator.com"><img he
ight="18" src="y18.svg" style="border:1px white solid; display:bloc
k" 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="newc
omments">comments</a> | <a href="ask">ask</a> | <a href="show">show
```

In [35]:

```
headline_rows = soup.find_all('span', class_='titleline')
headlines = []
for row in headline_rows:
    headline_link = row.find('a')
    if headline_link:
    # Extract the text (headline) and the 'href' attribute (URL)
        headline_text = headline_link.text
        headline_url = headline_link['href']

    # Append a tuple of the headline text and URL to the headlines list
        headlines.append((headline_text, headline_url))
headlines
```

Out[35]:
```
[('Structuralism as a Philosophy of Mathematics',
  'https://www.infinitelymore.xyz/p/structuralism'),
 ('Did any processor implement an integer square root instruction?',
  'https://retrocomputing.stackexchange.com/questions/29787/did-any-pr
ocessor-implement-an-integer-square-root-instruction'),
 ('ElephantSQL Is Shutting Down',
  'https://www.elephantsql.com/blog/end-of-life-announcement.html'),
 ('Is the frequency domain a real place?',
  'https://lcamtuf.substack.com/p/is-the-frequency-domain-a-real-plac
e'),
 ('WinBtrfs – an open-source btrfs driver for Windows',
  'https://github.com/maharmstone/btrfs'),
 ('Sophia: Scalable Stochastic 2nd-Order Optimizer for Language Model
Pre-Training',
  'https://arxiv.org/abs/2305.14342'),
 ('PM2: Production Process Manager with a Built-In Load Balancer',
  'https://github.com/Unitech/pm2'),
 ('Show HN: Online database diagram editor',
  'https://github.com/drawdb-io/drawdb'),
 ('Cache is King: A guide for Docker layer caching in GitHub Actions',
  'https://blacksmith.sh/blog/cache-is-king-a-guide-for-docker-layer-c
aching-in-github-actions'),
 ('Dot: use of local LLMs and RAG in particular to interact with docum
ents',
  'https://github.com/alexpinel/Dot'),
 ('Faces.js, a JavaScript library for generating vector-based cartoon
faces',
  'https://zengm.com/facesjs/'),
 ('Gakken Ex-System', 'https://en.wikipedia.org/wiki/Gakken_EX-Syste
m'),
 ('A memory model for Rust code in the kernel',
  'https://lwn.net/SubscriberLink/967049/0ffb9b9ed8940013/'),
 ('A canonical Hamiltonian formulation of the Navier–Stokes problem',
  'https://www.cambridge.org/core/journals/journal-of-fluid-mechanics/
article/canonical-hamiltonian-formulation-of-the-navierstokes-problem/
B3EB9389AE700867A6A3EA63A45E69C6'),
 ('Lago, Open-Source Stripe Alternative, banks $22M in funding',
  'https://techcrunch.com/2024/03/14/lago-a-paris-based-open-source-bi
lling-platform-banks-22m/'),
 ('Anti-crime humps in medieval Venice',
  'https://www.visitvenezia.eu/en/venetianity/discover-venice/the-vene
tian-antibandito-humps-or-pissotte-what-exactly-are-they'),
 ('A Theory of Composing Protocols (2023)',
  'https://programming-journal.org/2023/7/6/'),
 ('More Agents Is All You Need: LLMs performance scales with the numbe
r of agents',
  'https://arxiv.org/abs/2402.05120'),
 ('Chisel: A fast TCP/UDP tunnel over HTTP',
  'https://github.com/jpillora/chisel'),
 ('The xz sshd backdoor rabbithole goes quite a bit deeper',
  'https://twitter.com/bl4sty/status/1776691497506623562'),
 ('Exposure therapy for arachnophobia can benefit unrelated fears, stu
dy finds',
  'https://www.psypost.org/exposure-therapy-for-arachnophobia-can-bene
fit-unrelated-fears-study-finds/'),
 ('Show HN: Brutalist Hacker News – A HN reader inspired by brutalist
web design',
  'https://brutalisthackernews.com'),
 ('ChrysaLisp GUI Demo [video]',
  'https://www.youtube.com/watch?v=ADvyZOxlBu4'),
 ('Zep AI (YC W24) is hiring a founding Go engineer',
```

```
  'https://jobs.gem.com/zep/am9icG9zdDre4RbzEeB4wYY7s9TjXwhp'),
 ('Tokens, n-grams, and bag-of-words models (2023)',
  'https://zilliz.com/learn/introduction-to-natural-language-processin
g-tokens-ngrams-bag-of-words-models'),
 ('Home insurers are dropping customers based on aerial images',
  'https://www.wsj.com/real-estate/home-insurance-aerial-images-37a18b
16'),
 ('System/360 - CHM Revolution',
  'https://www.computerhistory.org/revolution/mainframe-computers/7/16
4'),
 ('Language models are Super Mario: Absorbing abilities from homologou
s models',
  'https://arxiv.org/abs/2311.03099'),
 ('What I think about when I edit (2019)',
  'https://evaparish.com/blog/how-i-edit'),
 ('New sunflower family tree reveals multiple origins of flower symmet
ry',
  'https://phys.org/news/2024-04-sunflower-family-tree-reveals-multipl
e.html')]
```

In [ ]:

```
# Find all 'a' tags within 'td' tags with the class 'title'
headline_tags = soup.find_all('a')

# Extract headlines and URLs
headlines = [(tag.text, tag['href']) for tag in headline_tags]
type(headline_rows)
```

In [9]:

```
for i, (headline, url) in enumerate(headlines, 1):
    print(f"{i}. {headline}\n   {url}\n")
```

In [38]:

```
url = 'https://www.rbst.org.uk/Handlers/Download.ashx?IDMF=91a9de2c-bbdb
response = requests.get(url)
html_content = response.text
html_headers = response.headers
html_headers
soup = BeautifulSoup(html_content, 'html.parser')
soup
```

Out[38]:

```
%PDF-1.4
%����
4 0 obj
&lt;&gt;
endobj

xref
4 92
0000000016 00000 n
0000002486 00000 n
0000002611 00000 n
0000002651 00000 n
0000003645 00000 n
0000003704 00000 n
0000007564 00000 n
0000011204 00000 n
0000014819 00000 n
0000018388 00000 n
```

https://github.com/SOCStudentUoE/BEE2041-Empirical-Assignment (https://github.com/SOCStudentUoE/BEE2041-Empirical-Assignment)