Web Scraping Report

Group 1: Gavin Stone, Isaac Adams, Kaden Hicklin, Owen Miller, Samuel Shevlin, Sullivan Gleason

BeautifulSoup and MechanicalSoup

MechanicalSoup

A Python library for automating website interaction.

BeautifulSoup Architecture

Find Functions

Functions:

- find_all()
- find()
- find_parents()
- find_parent()
- find_all_next()
- find_next()
- find_all_previous()
- find_previous()
- etc

DOM Tree Functions

Functions:

- append()
- extend()
- NavigableString()
- insert()
- clear()
- extract()
- decompose()
- etc

MechanicalSoup Architecture

Superset of BeautifulSoup

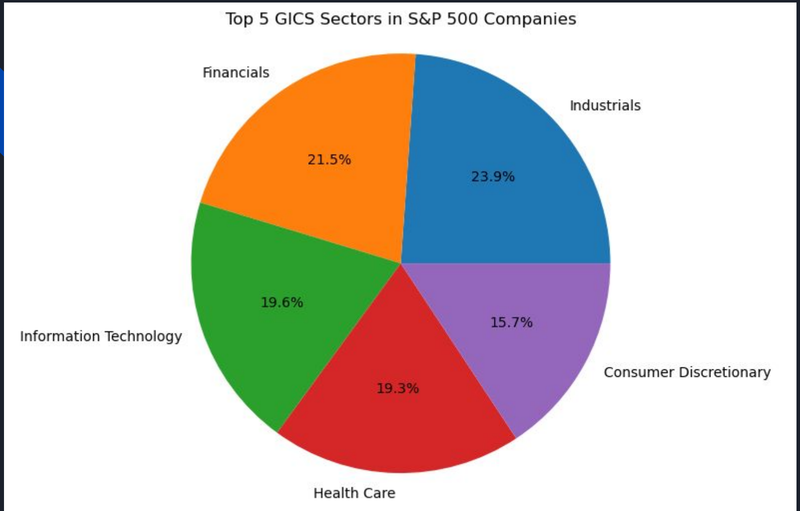
High-level

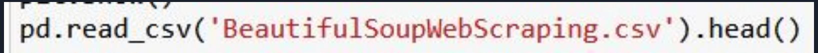
Classes:

- Form
- InvalidFormMethod
- LinkNotFoundError
- _BrowserState
- StatefulBrowser

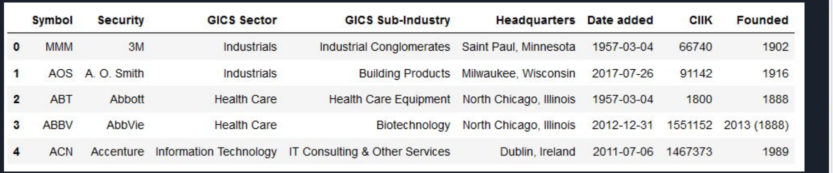
BeautifulSoup

- Serves as a tool for pulling data out of HTML and XML files
- Beautiful Soup 4 is the most current version
- Works with both Python 2.7 and Python 3.2
- Created by Leonard Richardson in 2004

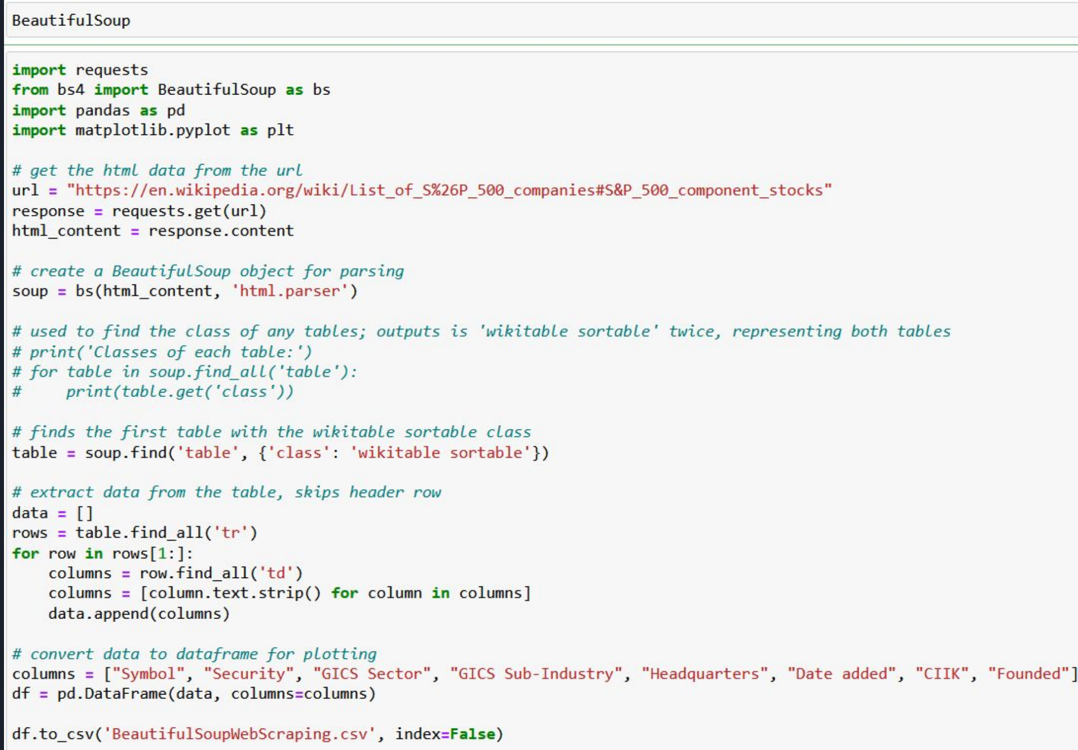




```
pd.read_csv('BeautifulSoupWebScraping.csv').head()
```



	Symbol	Security	GICS Sector	GICS Sub-Industry	Headquarters	Date added	CIIK	Founded
0	MMM	3M	Industrials	Industrial Conglomerates	Saint Paul, Minnesota	1957-03-04	66740	1902
1	AOS	A. O. Smith	Industrials	Building Products	Milwaukee, Wisconsin	2017-07-26	91142	1916
2	ABT	Abbott	Health Care	Health Care Equipment	North Chicago, Illinois	1957-03-04	1800	1888
3	ABBV	AbbVie	Health Care	Biotechnology	North Chicago, Illinois	2012-12-31	1551152	2013 (1888)
4	ACN	Accenture	Information Technology	IT Consulting & Other Services	Dublin, Ireland	2011-07-06	1467373	1989

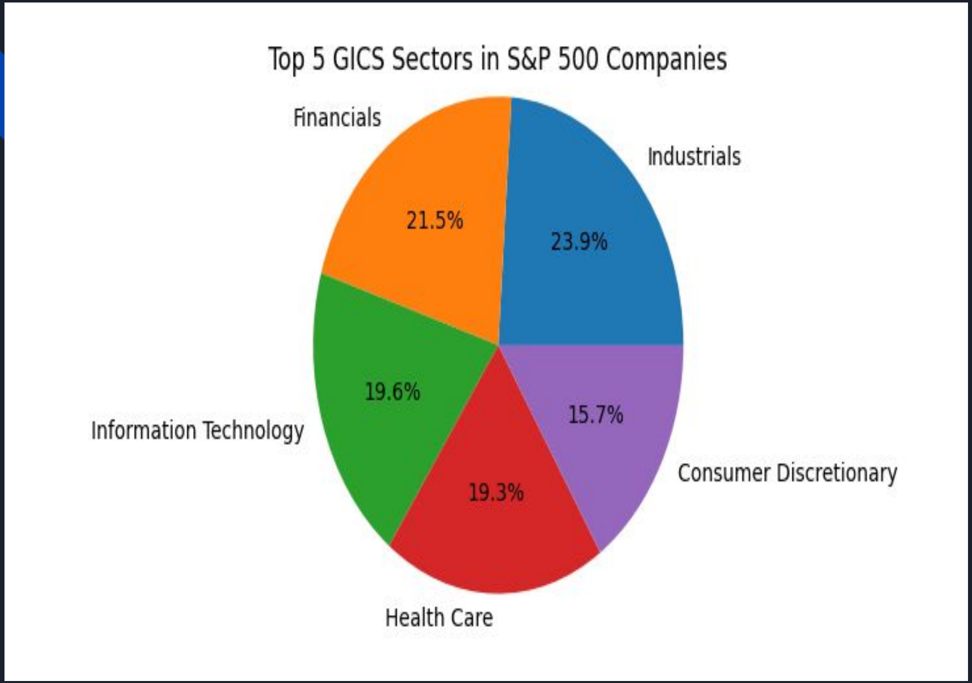


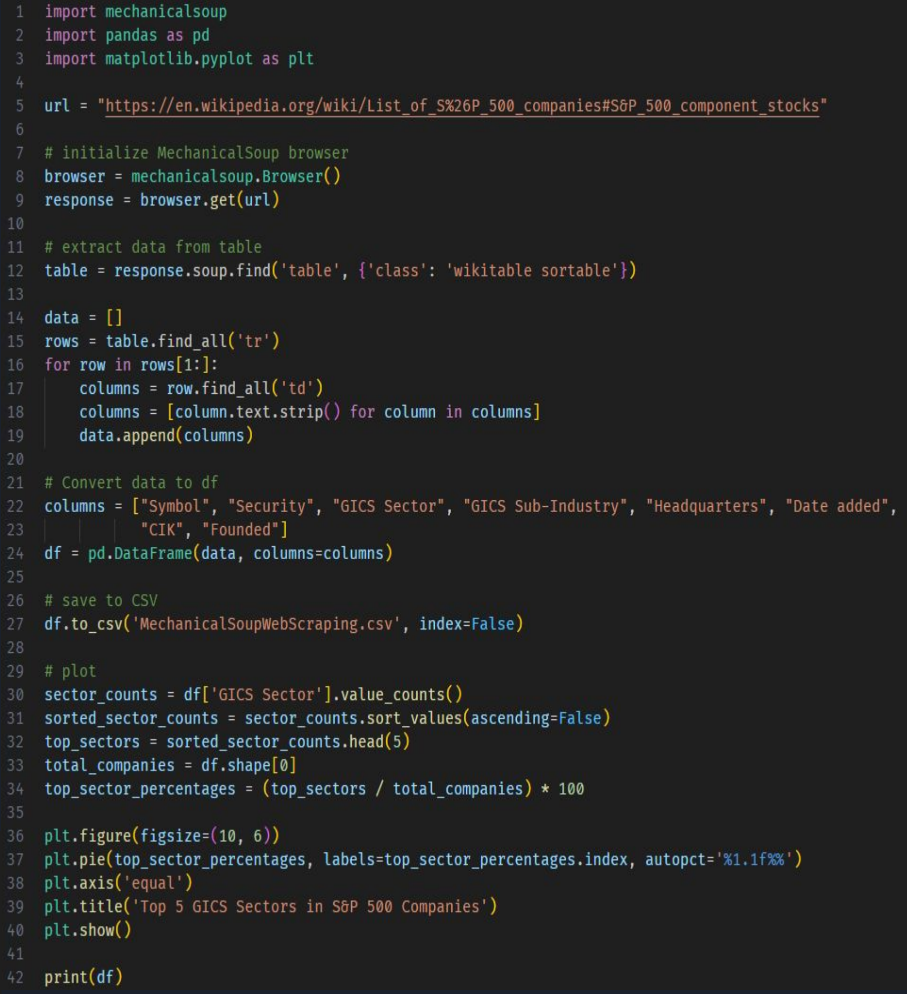
BeautifulSoup

```python
import requests
from bs4 import BeautifulSoup as bs
import pandas as pd
import matplotlib.pyplot as plt

# get the html data from the url
url = "https://en.wikipedia.org/wiki/List_of_S%26P_500_companies#S&P_500_component_stocks"
response = requests.get(url)
html_content = response.content

# create a BeautifulSoup object for parsing
soup = bs(html_content, 'html.parser')

# used to find the class of any tables; outputs is 'wikitable sortable' twice, representing both tables
# print('Classes of each table:')
# for table in soup.find_all('table'):
#     print(table.get('class'))

# finds the first table with the wikitable sortable class
table = soup.find('table', {'class': 'wikitable sortable'})

# extract data from the table, skips header row
data = []
rows = table.find_all('tr')
for row in rows[1:]:
    columns = row.find_all('td')
    columns = [column.text.strip() for column in columns]
    data.append(columns)

# convert data to dataframe for plotting
columns = ["Symbol", "Security", "GICS Sector", "GICS Sub-Industry", "Headquarters", "Date added", "CIIK", "Founded"]
df = pd.DataFrame(data, columns=columns)

df.to_csv('BeautifulSoupWebScraping.csv', index=False)
```

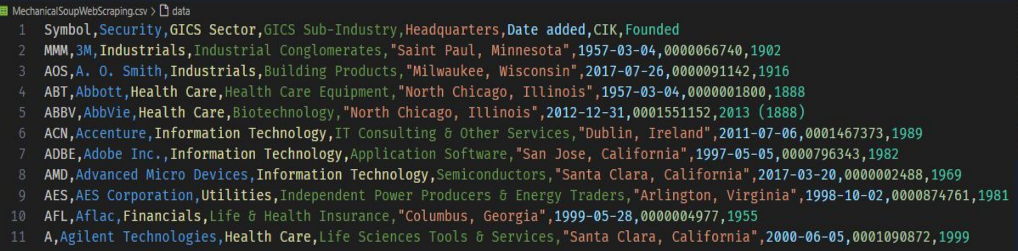
MechanicalSoup

- Based off of Request and MechanicalSoup
- Maintained By Open Source Community
- Great at automated web browsing and form submission automation
- No support for JavaScript





```python
import mechanicalsoup
import pandas as pd
import matplotlib.pyplot as plt

url = "https://en.wikipedia.org/wiki/List_of_S%26P_500_companies#S&P_500_component_stocks"

# initialize MechanicalSoup browser
browser = mechanicalsoup.Browser()
response = browser.get(url)

# extract data from table
table = response.soup.find('table', {'class': 'wikitable sortable'})

data = []
rows = table.find_all('tr')
for row in rows[1:]:
    columns = row.find_all('td')
    columns = [column.text.strip() for column in columns]
    data.append(columns)

# Convert data to df
columns = ["Symbol", "Security", "GICS Sector", "GICS Sub-Industry", "Headquarters", "Date added",
           "CIK", "Founded"]
df = pd.DataFrame(data, columns=columns)

# save to CSV
df.to_csv('MechanicalSoupWebScraping.csv', index=False)

# plot
sector_counts = df['GICS Sector'].value_counts()
sorted_sector_counts = sector_counts.sort_values(ascending=False)
top_sectors = sorted_sector_counts.head(5)
total_companies = df.shape[0]
top_sector_percentages = (top_sectors / total_companies) * 100

plt.figure(figsize=(10, 6))
plt.pie(top_sector_percentages, labels=top_sector_percentages.index, autopct='%1.1f%%')
plt.axis('equal')
plt.title('Top 5 GICS Sectors in S&P 500 Companies')
plt.show()

print(df)
```



```
Symbol,Security,GICS Sector,GICS Sub-Industry,Headquarters,Date added,CIK,Founded
MMM,3M,Industrials,Industrial Conglomerates,"Saint Paul, Minnesota",1957-03-04,0000066740,1902
AOS,A. O. Smith,Industrials,Building Products,"Milwaukee, Wisconsin",2017-07-26,0000091142,1916
ABT,Abbott,Health Care,Health Care Equipment,"North Chicago, Illinois",1957-03-04,0000001800,1888
ABBV,AbbVie,Health Care,Biotechnology,"North Chicago, Illinois",2012-12-31,0001551152,2013 (1888)
ACN,Accenture,Information Technology,IT Consulting & Other Services,"Dublin, Ireland",2011-07-06,0001467373,1989
ADBE,Adobe Inc.,Information Technology,Application Software,"San Jose, California",1997-05-05,0000796343,1982
AMD,Advanced Micro Devices,Information Technology,Semiconductors,"Santa Clara, California",2017-03-20,0000002488,1969
AES,AES Corporation,Utilities,Independent Power Producers & Energy Traders,"Arlington, Virginia",1998-10-02,0000874761,1981
AFL,Aflac,Financials,Life & Health Insurance,"Columbus, Georgia",1999-05-28,0000004977,1955
A,Agilent Technologies,Health Care,Life Sciences Tools & Services,"Santa Clara, California",2000-06-05,0001090872,1999
```