BS vs Selenium vs API

Beautiful Soup

It is a tool used for web scraping. It helps to extract data from a web page.

Advantages:

1. Helps to pull data from XML and HTML files.
2. Easy to learn and master.
3. Good community support to figure out issues which arise while working with its library.

Beautiful Soup

Disadvantages;

1. It cannot do the entire job on its own requires specific modules to do it.
2. The library needs to make a request to website, to help with this it needs requests or urllib2 modules.
3. Also, after downloading HTML,XML data in our local machine it needs external parser to parse the entire data. Ex - html5lib
4. Not helpful with AJAX requests.

Selenium

Advantages:

1. It can easily work with core Javascript.
2. It can easily handle AJAX and PJAX requests.
3. Doesn’t need external modules to do the job.
4. Easy to visualize process i.e. we use python to access browser to scrape.

Selenium

Disadvantages:

1. Comparatively slower than BeautifulSoup method as it uses browser to extract data rather than using any requests module like BeautifulSoup.
2. With website which load items while scrolling in such cases PageSource function is used with Beautiful Soup in such cases we need other modules to scrape data.

API

Another way that you can send requests to a website rather than simply visiting the page, is through an API or Application Programming Interface.The most common file types are JSON and XML.

Advantages:

1. Easier to use use, acces and extract data.
2. Usually provide more information than basic web scraping.
3. Generally, Uses JSON which is quite compatible with python programming language and makes our scraping much more effective.

API

Disadvantages:

1. It follows strict rules about what we can send and what it will return and those rules can’t be changed unless someone changes the API itself.
2. We can get some specific data fields from the API.
3. Mostly, these APIs are not free or have a free data limit after which we will be charged.

What should we use?

- Usually we use APIs for accessing web data. But if the costs are too high or the website doesn’t provide any API then we should use selenium or beautifulsoup.
- If the data we are extracting is not too big then it is good to use BeautifulSoup as it gets the job done in much less time than Selenium method.
- If the data is big then BS method shows timeout error and the website can block our IP from scraping in future.In such cases, using selenium is good as it loads the data in browser like a user.
- If the website has AJAX or PJAX requests then it is good to use selenium approach in such cases.

What should we use?

- Sometimes we came across pages in which we need to use both the methods (BS and Selenium) to scrape the data. Like when we are trying to scrape data of site which loads data while scrolling in this case we need to use PageSource function and BeautifulSoup with selenium to extract the data.
- Before using Selenium always use console to know exactly what data we will get from a particular selector.

Comparison

API Method	Selenium Method	BS Method
Widely used	Not Widely used	Not Widely used
Not Free	Free	Free
Not required	Not required	Uses external modules
Less Time	More Time	Medium Time
Different method	PJAX/AJAX is processed	Not processed
Easy code	Normal code	Little complexed