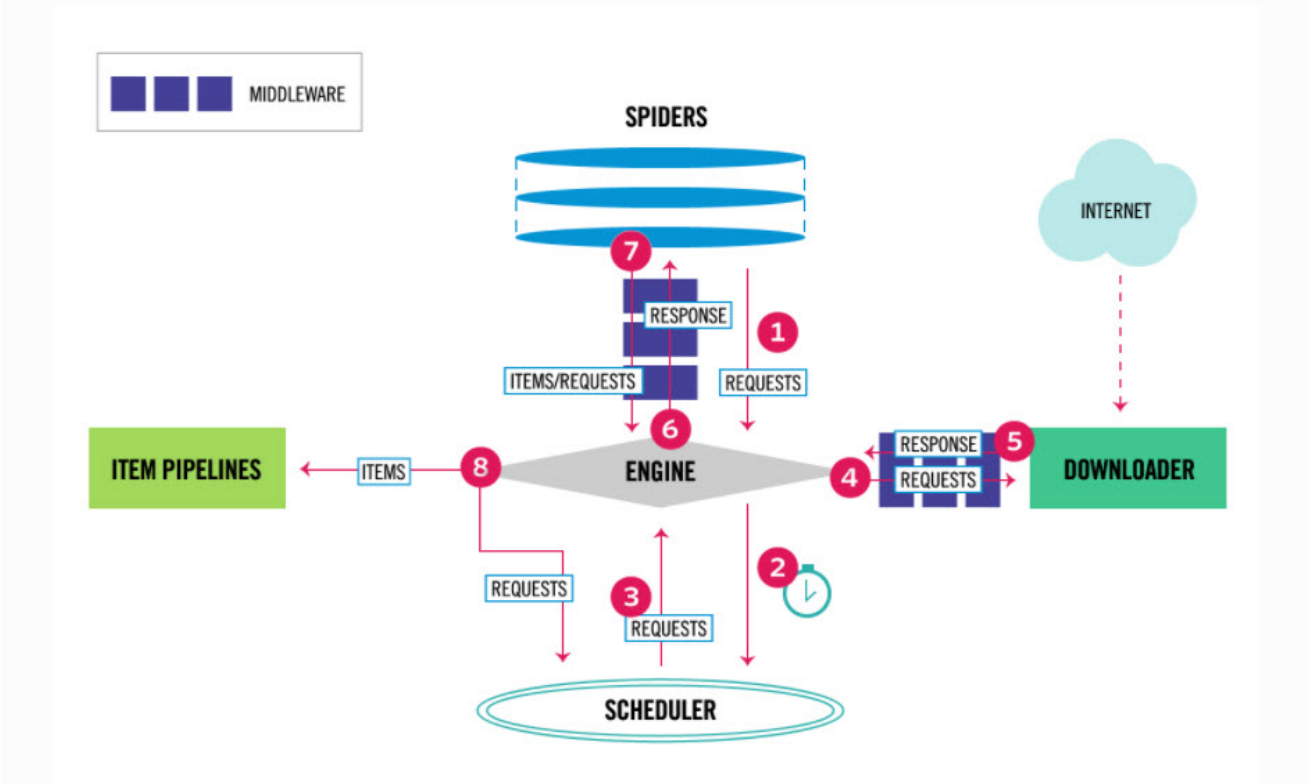
https://tanpham.org

Explain core skill need to prepare before using Scrapy : **regular expression, css selector**.

The Big Picture



The data flow in **Scrapy** is controlled by the execution engine, and goes like this:

1. The Engine gets the initial Requests to crawl from the Spider.
2. The Engine schedules the Requests in the Scheduler and asks for the next Requests to crawl.
3. The Scheduler returns the next Requests to the Engine.
4. The Engine sends the Requests to the Downloader, passing through the Downloader Middlewares (see `process_request()`).
5. Once the page finishes downloading the Downloader generates a Response (with that page) and sends it to the Engine, passing through the Downloader Middlewares (see `process_response()`).
6. The Engine receives the Response from the Downloader and sends it to the Spider for processing, passing through the Spider Middleware (see `process_spider_input()`).
7. The Spider processes the Response and returns scraped items and new Requests (to follow) to the Engine, passing through the Spider Middleware (see `process_spider_output()`).
8. The Engine sends processed items to Item Pipelines, then send processed Requests to the Scheduler and asks for possible next Requests to crawl.
9. The process repeats (from step 1) until there are no more requests from the Scheduler.

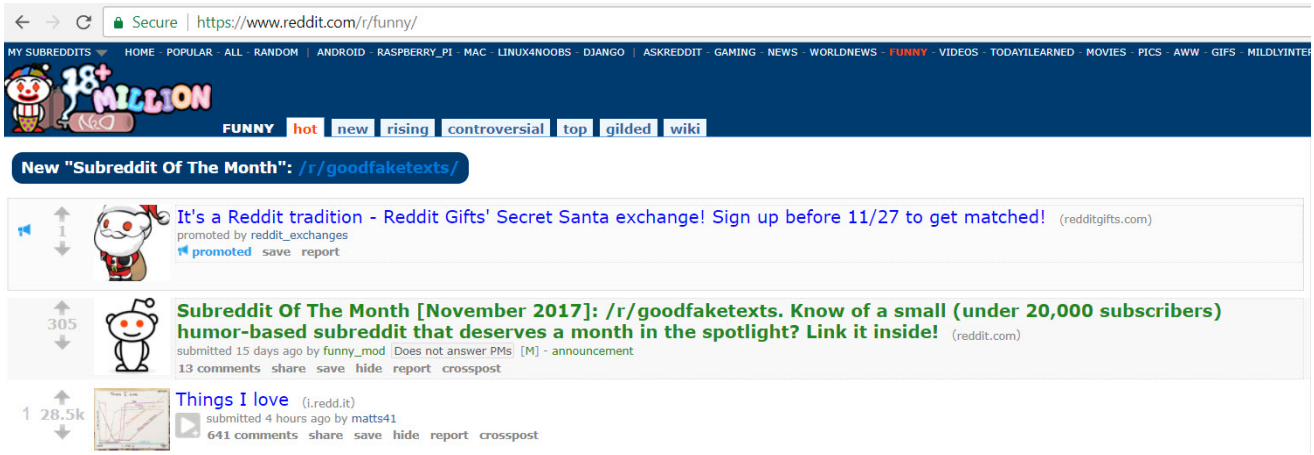
So What is Your Remain Job ?

When using framework for scrape data, it do a lot for you in a systematic way, from schedule, download , extracting and saving data. So your most important job while using Scrapy will be :

- Specify **where** you want to scraping data ?. Basically is a set of url, so Scrapy will crawl data from.
- Specify **what** you want in each data page ?. Name, title, image ...

For Example

For example, We want to get all funny title from Reddit which you could access from link https://www.reddit.com/r/funny/



Where to scrape?. It is collection of pages which you could access by click to `Next` button at bottom of page.

In detail, it will be following Urls.

```
https://www.reddit.com/r/funny/
https://www.reddit.com/r/funny/?count=25&after=t3_7e16ie
https://www.reddit.com/r/funny/?count=50&after=t3_7e42mi
https://www.reddit.com/r/funny/?count=75&after=t3_7e1n14
https://www.reddit.com/r/funny/?count=100&after=t3_7dw9e5
https://www.reddit.com/r/funny/?count=125&after=t3_7e4u1p
...
```

What to scrape ? With each funny story, I care about **title**, **image**, and **score**. Important thing : these information are keep inside HTML tags. So our job is select these tags using `css selector` or `xpath` .

Select Urls with Regular Expression

The first important question is how to feed Scrapy with right collection of URL ?. So Scrapy will help you crawl HTML from that pages.

Scrapy using Regular Expression to filter out urls (You will see this in detail next parts). For examples, we want Scrapy crawl following urls

```
https://www.reddit.com/r/funny/
https://www.reddit.com/r/funny/?count=25&after=t3_7e16ie
https://www.reddit.com/r/funny/?count=50&after=t3_7e42mi
https://www.reddit.com/r/funny/?count=75&after=t3_7e1n14
https://www.reddit.com/r/funny/?count=100&after=t3_7dw9e5
https://www.reddit.com/r/funny/?count=125&after=t3_7e4u1p
...
```

What regular expression could filter out theses urls. Let try to find out this in real time with https://regexr.com/

Following regular expression will match required urls

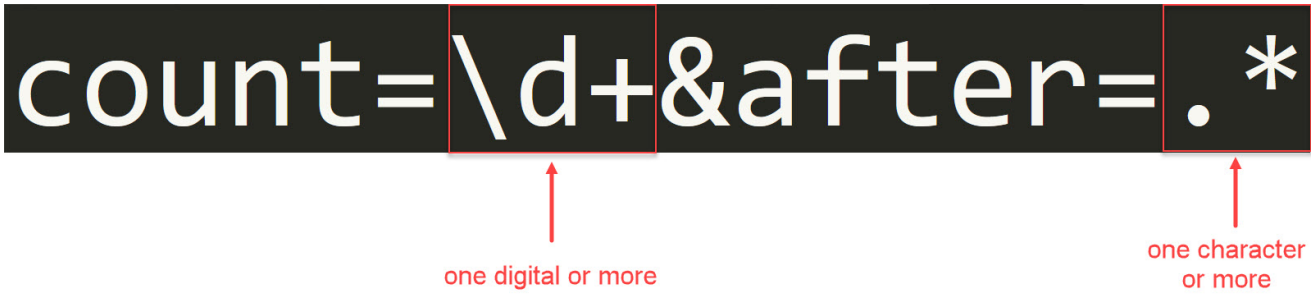
Expression

```
/count=\d+&after=.*/g
```

Text

```
https://www.reddit.com/r/funny/?count=25&after=t3_7e16ie
https://www.reddit.com/r/funny/?count=50&after=t3_7e42mi
https://www.reddit.com/r/funny/?count=75&after=t3_7e1n14
https://www.reddit.com/r/funny/?count=100&after=t3_7dw9e5
https://www.reddit.com/r/funny/?count=125&after=t3_7e4u1p
```

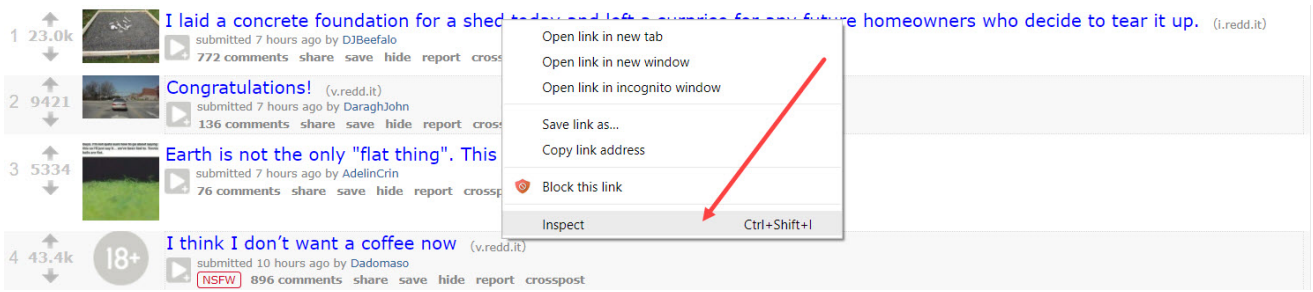
Let explain some thing about this regular expression and you will understand how regular expression work.



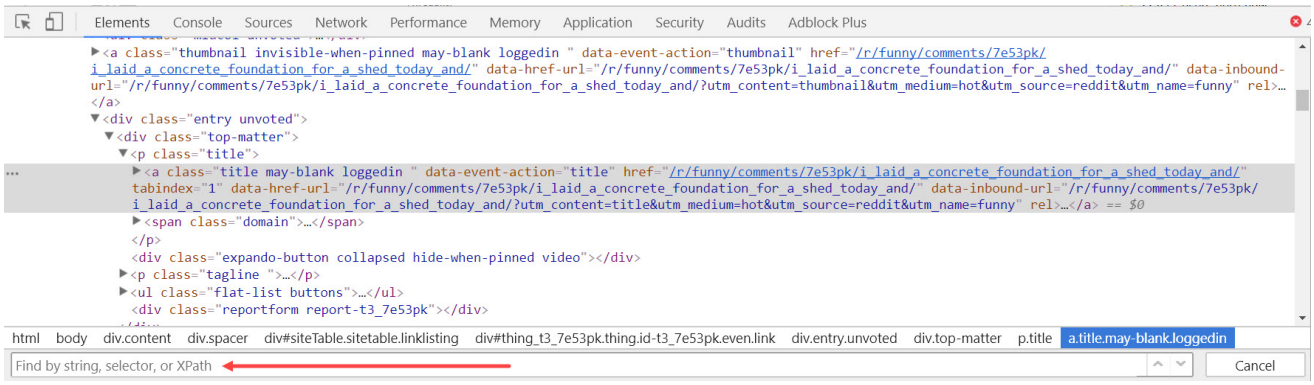
For more detail and practice on regular expression, please access this site https://regexone.com/

Select HTML Tags with CSS Selector

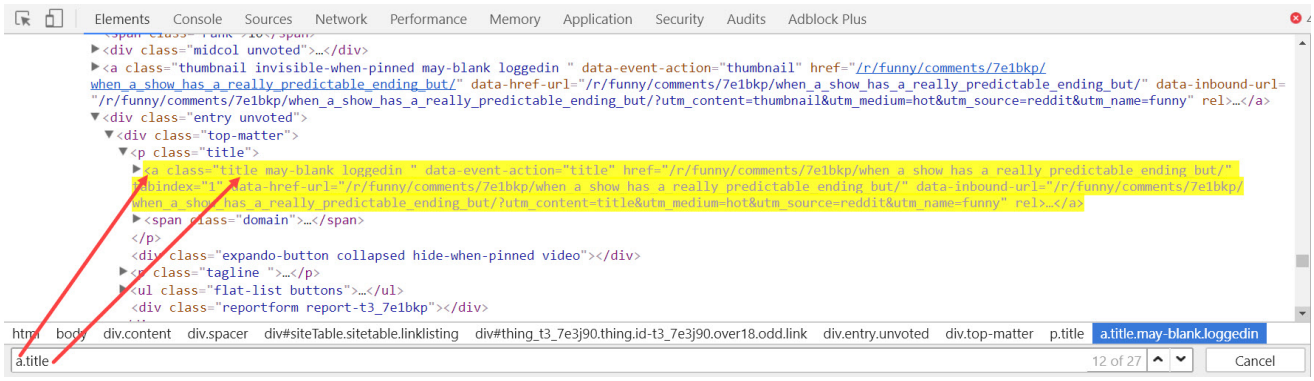
The second important thing is define what data you want when HTML already crawled. For example you open this page from Chrome browser https://www.reddit.com/r/funny/ . Move mouse above a title and right click then choose ***"Inspect"***.



Chrome inspection tool will show up with all HTML tags from current page. Type in ***"Ctrl + F"*** search tool appear, allow us try to use css selector to select HTML tags.



For example, to search for `a` tag with class `title` , we put in following css selector `a.title` then click **Enter** . The result will show up tag by tag.



That is how css selector work. To make more clear and detail about css selector, please refer to link

https://www.w3schools.com/cssref/css_selectors.asp