UC Riverside

CS 105 : Data Analysis Methods

Lab 4

1 Instructions

For this week's lab, we will use lab time to start working on the final project. Be sure to:

- Join your team's repository on github. You should accept the invitation to join the team.
- Read through the Scrapy web-crawler tutorial: `https://docs.scrapy.org/en/latest/intro/tutorial.html`
- Get started with your project

2 Important notes

When using Scrapy, you don't have to use the python Jupyter Notebook. Instead, just open a terminal window on the lab computers (or ssh to the provided machine cluster) to get started.

You will have to install Scrapy before getting started.
`https://docs.scrapy.org/en/latest/intro/install.html#intro-install`
Scapy Tutorial

- pip install –user Scrapy
- Create a project in Scrapy: **scrapy startproject tutorial**
- This will create a tutorial directory structure.
- Create a file called quotes_spider.py and copy the code found on the provided link

```python
import scrapy


class QuotesSpider(scrapy.Spider):
    name = "quotes"

    def start_requests(self):
        urls = [
            'http://quotes.toscrape.com/page/1/',
            'http://quotes.toscrape.com/page/2/',
        ]
        for url in urls:
            yield scrapy.Request(url=url, callback=self.parse)

    def parse(self, response):
        page = response.url.split("/")[-2]
        filename = 'quotes-%s.html' % page
        with open(filename, 'wb') as f:
            f.write(response.body)
        self.log('Saved file %s' % filename)
```

- To run, in the terminal type **scrapy crawl quotes**. Notice 'quotes' is the name we gave to the crawler.
- This will download 2 html files into the directory. Now, all you have to do is pull-out the specific data of interest.

Use this model to build a scraper for your project pages.