**Web Scraping**

**Commands:**- View Available Spider templates: scrapy genspider -l  
- Start a new project: scrapy startproject **project\_name**- Generate a new “basic” spider: scrapy genspider **spider\_name page\_URL**- Generate a new “crawl” spider: scrapy genspider -t crawl **spider\_name page\_URL**- Try an Xpath response:  
 - shell scrapy **site\_url**  
- **list\_name** = response.xpath(“**xpath\_keyword**”)  
- **list\_name**  
- Export as a json file: scrapy crawl **spider\_name** -o **output\_file\_name**.json  
- Export as csv file: scrapy crawl **spider\_name** -o **output\_file\_name** csv  
- Export as a xml file: scrapy crawl **spider\_name** -o **output\_file\_name**.xml  
- Try an Xpath response:   
 - shell scrapy **site\_url**  
 - request.headers

**Shortcuts:**

- Show/Hide Bottom panel: ctrl + j  
- Show/Hide Side panel: ctrl + b  
- Duplicate a complete line: alt + shift + down arrow  
- Comment a complete line: ctrl + k + c  
- Activate developer tools in browser: ctrl + shift + i  
- Show HTML of a webpage: ctrl + u

**Notes**

- Scrapy does not use the utf-8 encoding for json files so you will usually see that some of the collected data is encode. To fix this issue; open the settings file inside the project & add this line @ the end: FEED\_EXPORT\_ENCODING = ‘utf-8’

- To see the request headers sent by the browser:  
 - Open Developer Tools.  
 - Go to Network tab.  
 - Refresh webpage.  
 - Select any of the sent requests; you’ll find the headers @ the bottom (User-Agent).

- To change request headers sent by the spider into one similar to the browser’s:  
 - Copy browser headers as shown in the previous note.  
 - 1st way:  
 - Open the settings file inside the project folder.  
 - Uncomment the **DEFAULT\_REQUEST\_HEADERS** lines.  
 - Delete whatever’s already inside the dictionary.  
 - Paste the browser’s copied request with key **User-Agent**.  
 -2nd way:  
 - Remove the **start\_urls** line.  
 - Define a new function as follows:

 def start\_requests(self):

        yield scrapy.Request(url='http://worldpopulationreview.com/countries/countries-by-national-debt/', callback=self.parse, headers={'User-Agent': })

- If there is a pagination yield; add the same headers dictionary to the arguments of the scrapy.Request (or response.follow) function.