Reference Sheet Created by Shayne Lurya, Summer 2015.

**Virtual Environment**

**mkvirtualenv your\_venv**

*Make new virtual environment*

**workon (your\_venv)**

*Displays list of virtual environments (or specify which one)*

**deactivate your\_venv**

*Leave virtual environment*

**Scrapy**

Creating a project: You need to have both an **items.py** and a **spider.py** file for the program to work.

**scrapy startproject newproject**

Creates:

**newproject/**

**newproject/** #Python module, where your stuff goes

**items.py** #items file, automatically created

**spiders/** #Where you will store your spiders

**cd newproject/newproject/spiders**

* + Navigate to the project folder to create your spider

**echo “” > spider.py**

Quick way to create an spider.py file

Remember to name the spider so that you can run it using **scrapy** **crawl** command

**name = npr\_scrape**

**Running the spider**

**cd newproject/newproject/**

**scrapy crawl npr\_scrape -o <nameoffile.XXX>**

Collect the data from Items file

Allowed extensions:

* + **xml, jl, json, jsonlines, csv, pickle, marshal**

**scrapy crawl npr\_scrape > textfile.txt**

Output text from print statements in your code

**If you are only printing text and not using items file, you don’t need to create a project:**

**scrapy runspider yourspider.py > textfile.txt**

**DEBUGGING:**

Open a python shell to debug scrapy

**scrapy shell ‘**[**http://www.npr.org**](http://www.npr.org)**’**

>>> response

<200 <http://www.npr.org>>

>>>for url in response.css("a::attr('href')").extract():url

u'/sections/news/'

u'/sections/us/'

u'/sections/world/'

u'/sections/politics/'

…

>>>quit()