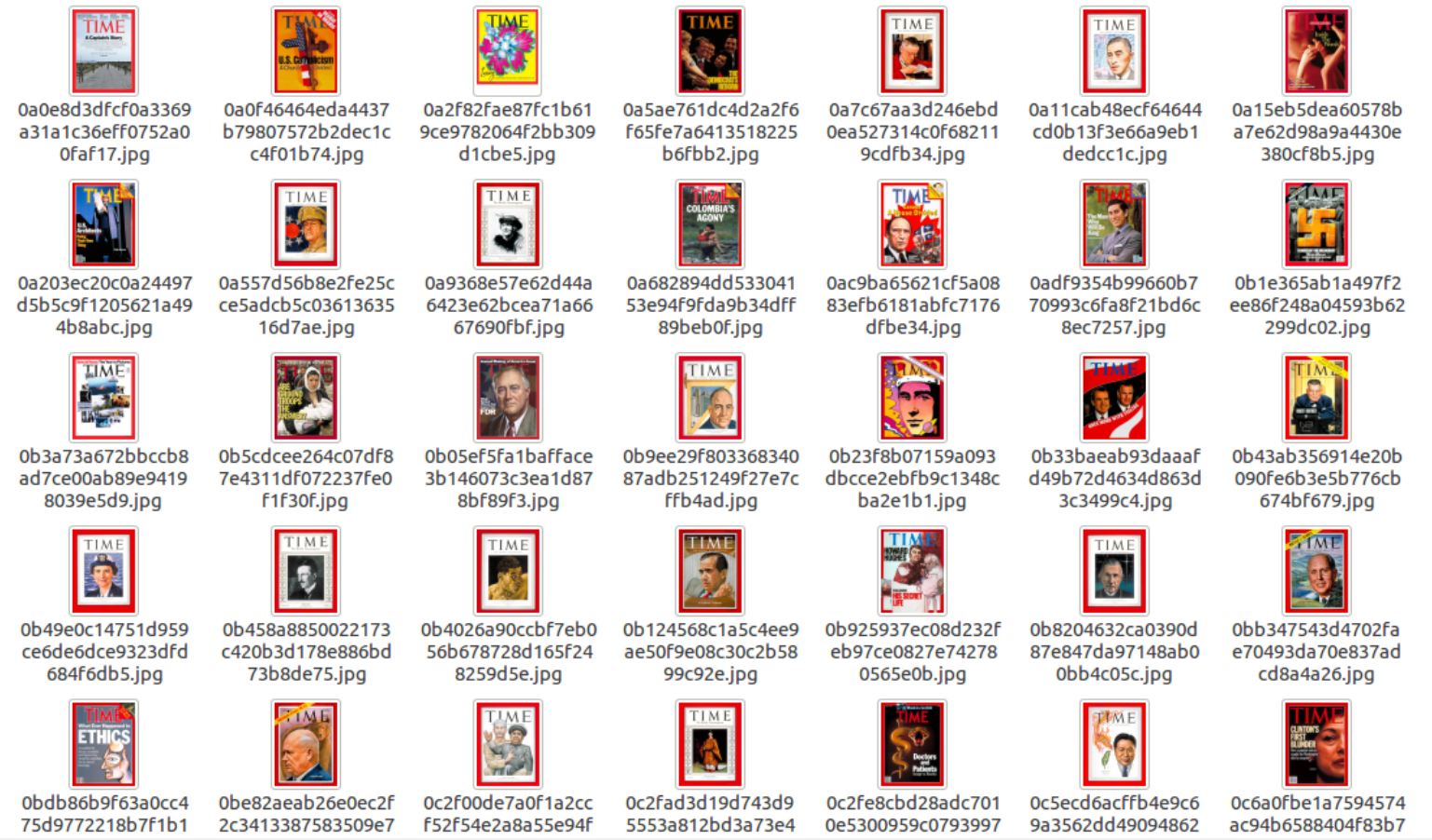
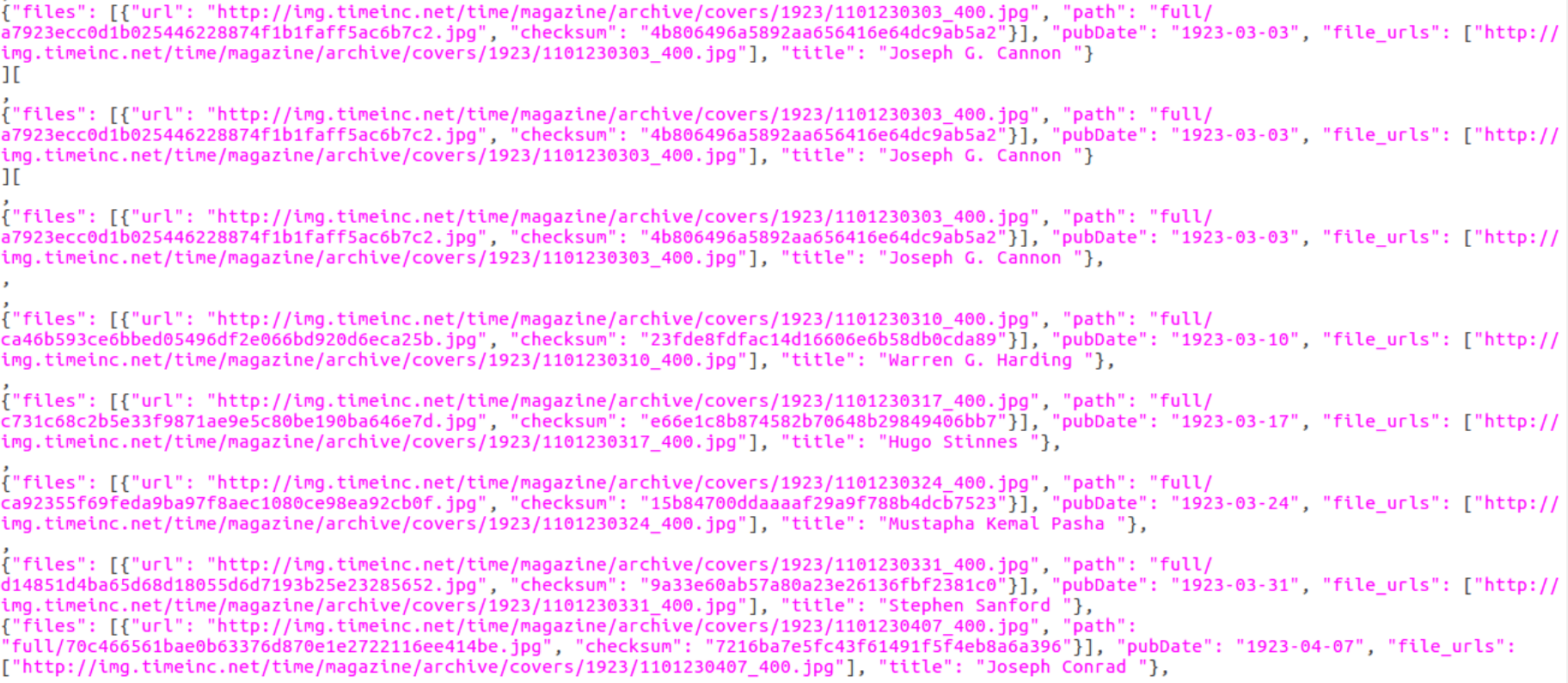
Scraping images with Python and Scrapy

**4762** Images Downloaded



## JSON File



## Installing Scrapy

I am using my **Ubuntu 16** System and **Python 2.7**

After that I am installing few dependencies for Scrapy parse documents

|  |  |
| --- | --- |
| 1 | $ sudo apt-get install libffi-dev |
| 2 | $ sudo apt-get install libssl-dev |
| 3 | $ sudo apt-get install libxml2-dev libxslt1-dev |

In order to scrap data in actual binary files We need the Pillow libraries.

|  |  |
| --- | --- |
| 1 | $ pip install pillow |
| 2 | $ pip install scrapy |

## Creating the Scrapy project

|  |  |
| --- | --- |
| 1 | $ scrapy startproject timecoverspider |

**Project Structure**

|--- scrapy.cfg

|    |--- timecoverspider

|    |    |--- \_\_init\_\_.py

|    |    |--- **items.py**

|    |    |--- pipelines.py

|    |    |--- **settings.py**

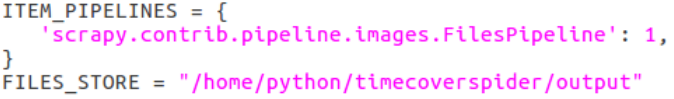
|    |    |--- spiders

|    |    |    |---- \_\_init\_\_.py

|    |    |    |---- **coverspider.py** # (we need to manually create this file)

**Changes Required in File**

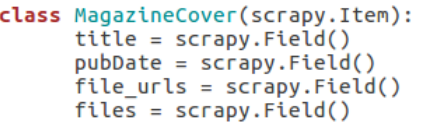
setting.py



1 First setting will activate default capability of file scraping capability.

2 Output Directory where our output will be stored.

Item.py



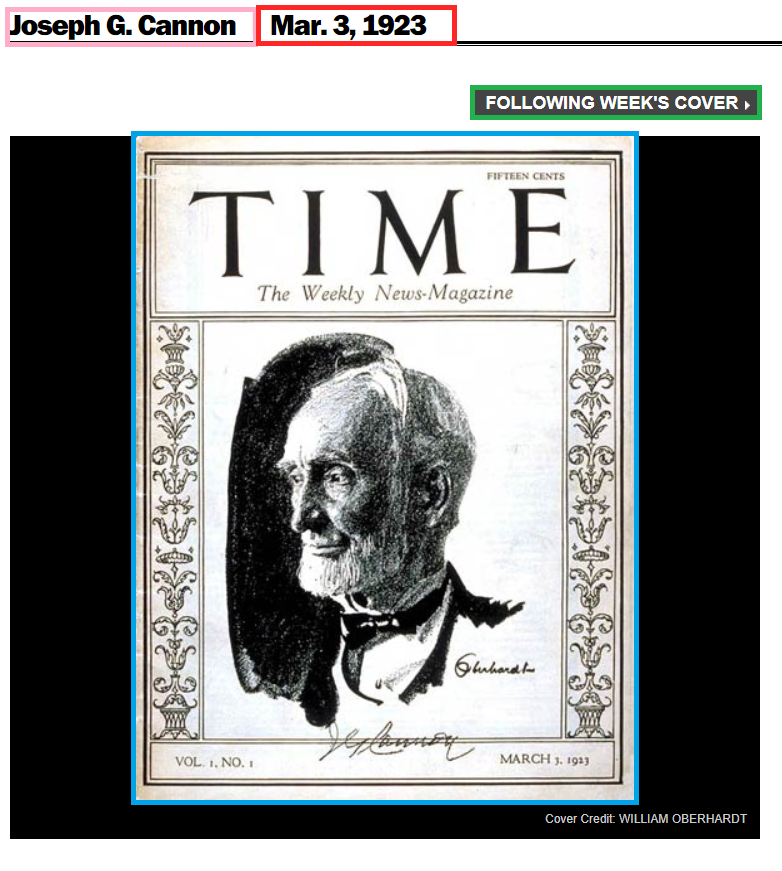
Self-explanatory!!

coverspider.py

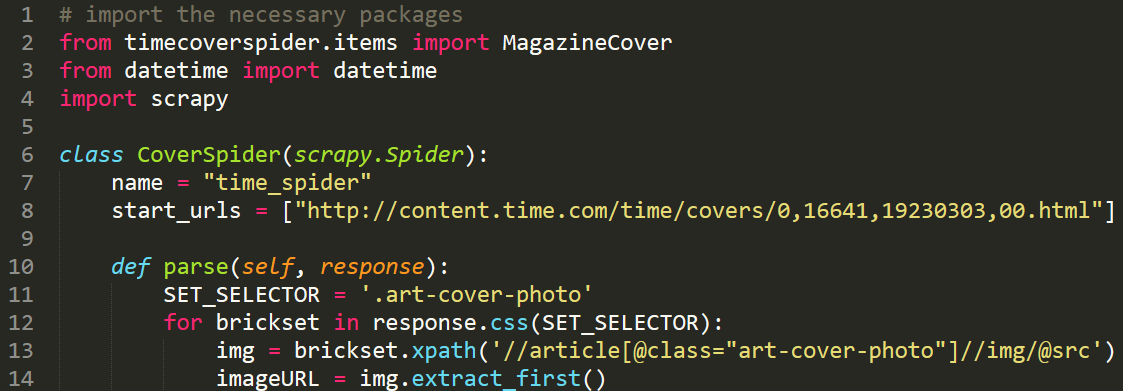
Goal 1: Extract Image

Goal 2: Extract Name and Date and time

Goal 3: Click on Next page (Following Weeks cover)

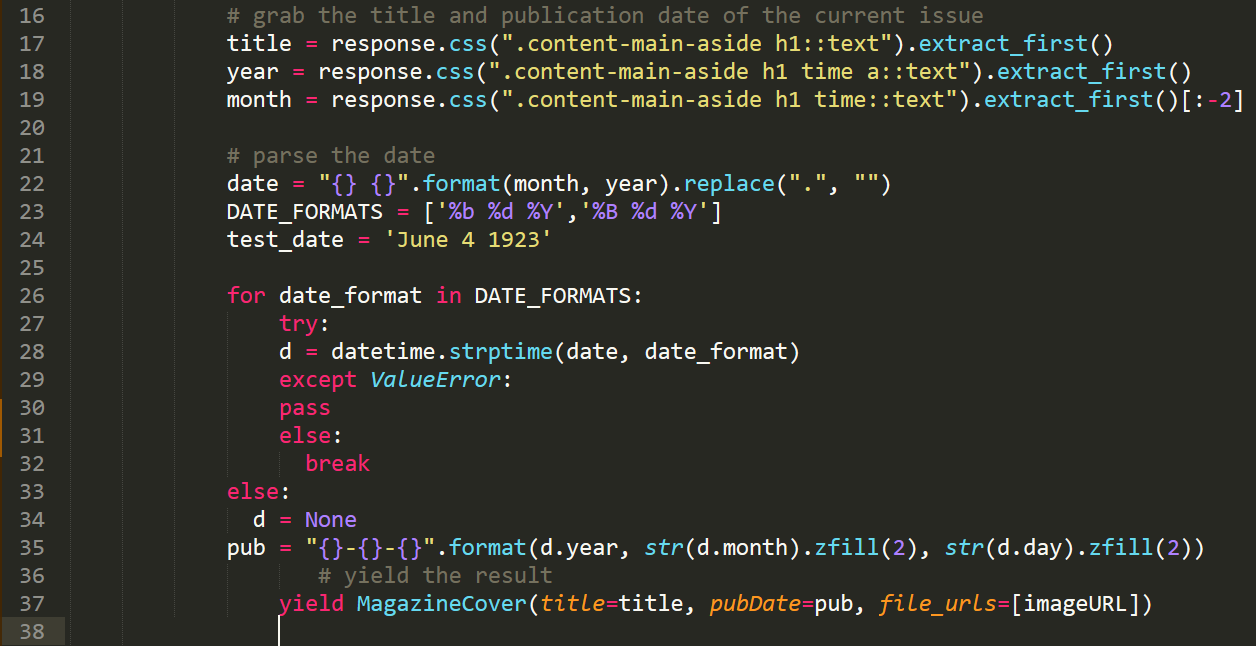


Extract Images



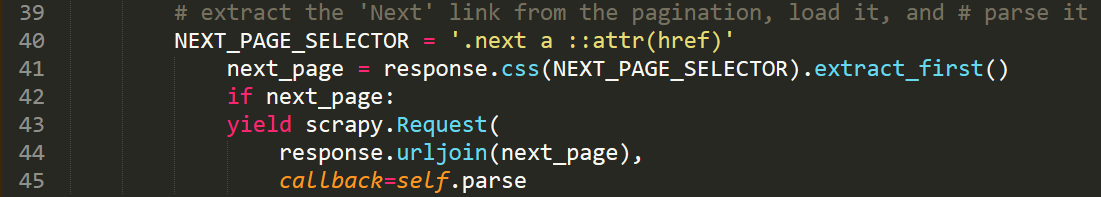


Extract Date and Time



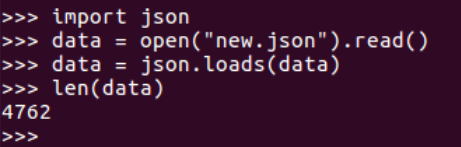


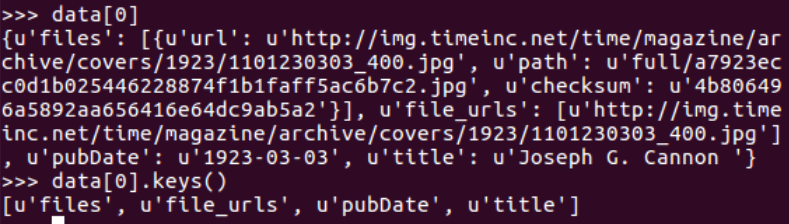
Click on next page utile last page

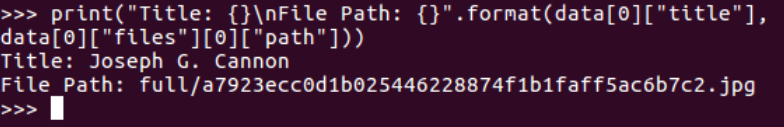
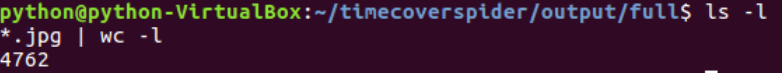




Some







1. What are the important Libraries you have used?

* Pillow
* Scrapy
* date time

1. What were the issues with scraping data?

* First of all was the url was I looking was out of date
* Scrap the data parsing method I have redefine
* Date time format
* Json file format
* Data storage
* I had few leaking in my pipeline.

1. What were the issues you faced with Libraries?

* Not really any issue I had with libraries expect data time one.

1. What were the data conversion issues?

Pipeline for data conversion was broken I have to fix that one

Other then I have data in Img file format and JSON I am happy with it.

This is a really cool thing I’ve done and I’m super excited about it!!!

References

1. R. (2015, October 12). Scraping images with Python and Scrapy. Retrieved March 21, 2018, from <https://www.pyimagesearch.com/2015/10/12/scraping-images-with-python-and-scrapy/>

P. T. (2013, September 09). Retrieved March 21, 2018, from https://groups.google.com/forum/#!msg/scrapy-users/kzGHFjXywuY/O6PIhoT3thsJ

D. (n.d.). Dhawal1306/Scrapy. Retrieved March 21, 2018, from <https://github.com/Dhawal1306/Scrapy>

How to download image using Scrapy? (n.d.). Retrieved March 21, 2018, from <https://stackoverflow.com/questions/39745685/how-to-download-image-using-scrapy>

Scrapy - Select specific link based on text. (n.d.). Retrieved March 21, 2018, from <https://stackoverflow.com/questions/12145067/scrapy-select-specific-link-based-on-text>

Scrapy: Images Pipeline, download images. (n.d.). Retrieved March 21, 2018, from <https://stackoverflow.com/questions/38589188/scrapy-images-pipeline-download-images>

15.3. time - Time access and conversions. (n.d.). Retrieved March 21, 2018, from <https://docs.python.org/2/library/time.html>