SCRAPY

CS 010

Design and Implementation of Solutions to Computational Problems

Prof. Donald J. Patterson

WHAT IS WEBCRAWLING?

https://scrapy.org/

EXTENDING PYTHON

MODULES EXTEND PYTHON

- Numpy, Scipy, Tweepy
- Scrapy
  - software to read and crawl through web sites

Scrapy

An open source and collaborative framework for extracting the data you need from websites. In a fast, simple, yet extensible way.

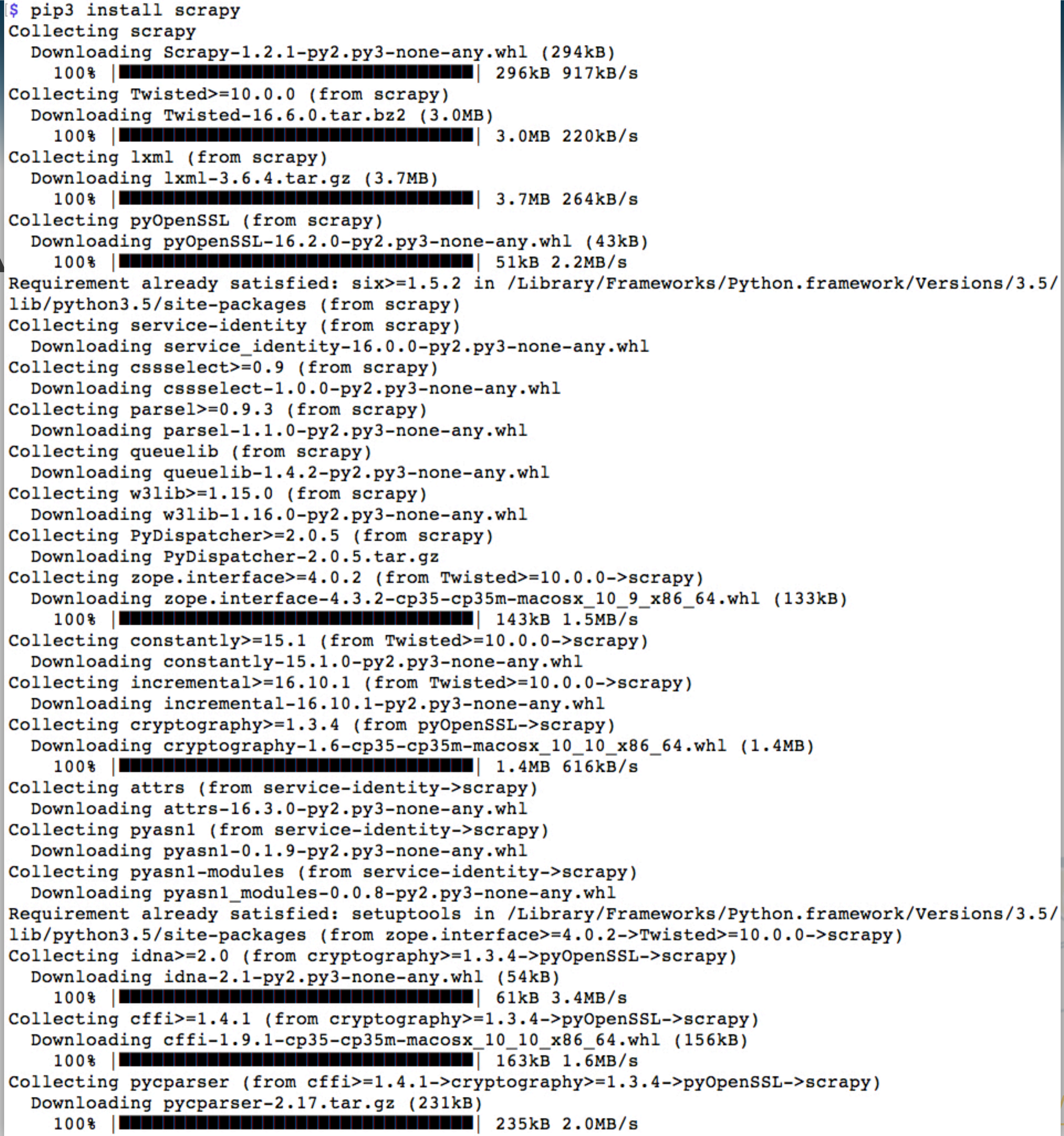
https://scrapy.org/

SCRAPY

INSTALLATION

SCRA

INSTA



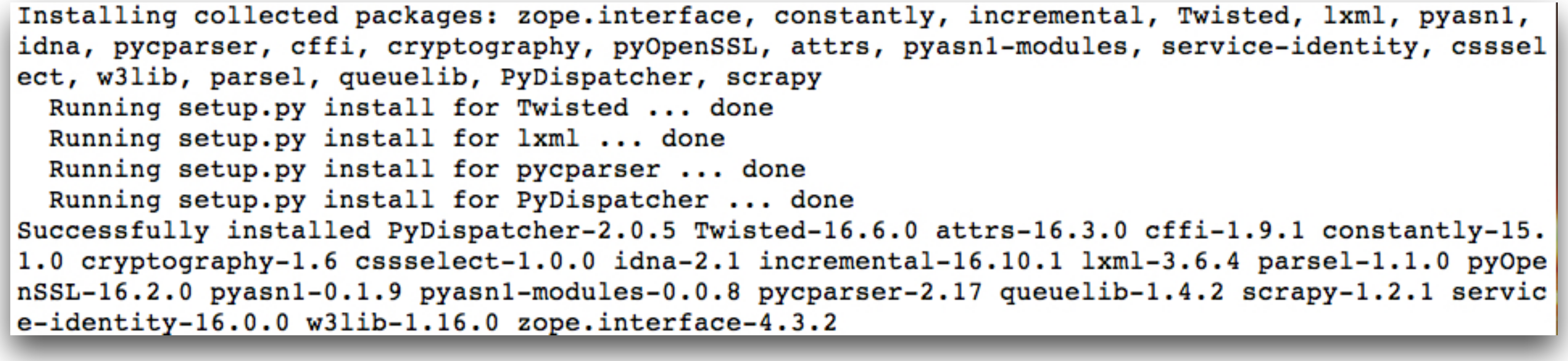
```
$ pip3 install scrapy
Collecting scrapy
  Downloading Scrapy-1.2.1-py2.py3-none-any.whl (294kB)
    100% |████████████████████████████████| 296kB 917kB/s
Collecting Twisted>=10.0.0 (from scrapy)
  Downloading Twisted-16.6.0.tar.bz2 (3.0MB)
    100% |████████████████████████████████| 3.0MB 220kB/s
Collecting lxml (from scrapy)
  Downloading lxml-3.6.4.tar.gz (3.7MB)
    100% |████████████████████████████████| 3.7MB 264kB/s
Collecting pyOpenSSL (from scrapy)
  Downloading pyOpenSSL-16.2.0-py2.py3-none-any.whl (43kB)
    100% |████████████████████████████████| 51kB 2.2MB/s
Requirement already satisfied: six>=1.5.2 in /Library/Frameworks/Python.framework/Versions/3.5/
lib/python3.5/site-packages (from scrapy)
Collecting service-identity (from scrapy)
  Downloading service_identity-16.0.0-py2.py3-none-any.whl
Collecting cssselect>=0.9 (from scrapy)
  Downloading cssselect-1.0.0-py2.py3-none-any.whl
Collecting parsel>=0.9.3 (from scrapy)
  Downloading parsel-1.1.0-py2.py3-none-any.whl
Collecting queuelib (from scrapy)
  Downloading queuelib-1.4.2-py2.py3-none-any.whl
Collecting w3lib>=1.15.0 (from scrapy)
  Downloading w3lib-1.16.0-py2.py3-none-any.whl
Collecting PyDispatcher>=2.0.5 (from scrapy)
  Downloading PyDispatcher-2.0.5.tar.gz
Collecting zope.interface>=4.0.2 (from Twisted>=10.0.0->scrapy)
  Downloading zope.interface-4.3.2-cp35-cp35m-macosx_10_9_x86_64.whl (133kB)
    100% |████████████████████████████████| 143kB 1.5MB/s
Collecting constantly>=15.1 (from Twisted>=10.0.0->scrapy)
  Downloading constantly-15.1.0-py2.py3-none-any.whl
Collecting incremental>=16.10.1 (from Twisted>=10.0.0->scrapy)
  Downloading incremental-16.10.1-py2.py3-none-any.whl
Collecting cryptography>=1.3.4 (from pyOpenSSL->scrapy)
  Downloading cryptography-1.6-cp35-cp35m-macosx_10_10_x86_64.whl (1.4MB)
    100% |████████████████████████████████| 1.4MB 616kB/s
Collecting attrs (from service-identity->scrapy)
  Downloading attrs-16.3.0-py2.py3-none-any.whl
Collecting pyasn1 (from service-identity->scrapy)
  Downloading pyasn1-0.1.9-py2.py3-none-any.whl
Collecting pyasn1-modules (from service-identity->scrapy)
  Downloading pyasn1_modules-0.0.8-py2.py3-none-any.whl
Requirement already satisfied: setuptools in /Library/Frameworks/Python.framework/Versions/3.5/
lib/python3.5/site-packages (from zope.interface>=4.0.2->Twisted>=10.0.0->scrapy)
Collecting idna>=2.0 (from cryptography>=1.3.4->pyOpenSSL->scrapy)
  Downloading idna-2.1-py2.py3-none-any.whl (54kB)
    100% |████████████████████████████████| 61kB 3.4MB/s
Collecting cffi>=1.4.1 (from cryptography>=1.3.4->pyOpenSSL->scrapy)
  Downloading cffi-1.9.1-cp35-cp35m-macosx_10_10_x86_64.whl (156kB)
    100% |████████████████████████████████| 163kB 1.6MB/s
Collecting pycparser (from cffi>=1.4.1->cryptography>=1.3.4->pyOpenSSL->scrapy)
  Downloading pycparser-2.17.tar.gz (231kB)
    100% |████████████████████████████████| 235kB 2.0MB/s
```

//scrapy.org/

INSTALLATION

SCRAPY

INSTALLATION



```
Installing collected packages: zope.interface, constantly, incremental, Twisted, lxml, pyasn1,
idna, pycparser, cffi, cryptography, pyOpenSSL, attrs, pyasn1-modules, service-identity, csssel
ect, w3lib, parsel, queuelib, PyDispatcher, scrapy
  Running setup.py install for Twisted ... done
  Running setup.py install for lxml ... done
  Running setup.py install for pycparser ... done
  Running setup.py install for PyDispatcher ... done
Successfully installed PyDispatcher-2.0.5 Twisted-16.6.0 attrs-16.3.0 cffi-1.9.1 constantly-15.
1.0 cryptography-1.6 cssselect-1.0.0 idna-2.1 incremental-16.10.1 lxml-3.6.4 parsel-1.1.0 pyOpe
nSSL-16.2.0 pyasn1-0.1.9 pyasn1-modules-0.0.8 pycparser-2.17 queuelib-1.4.2 scrapy-1.2.1 servic
e-identity-16.0.0 w3lib-1.16.0 zope.interface-4.3.2
```

https://scrapy.org/

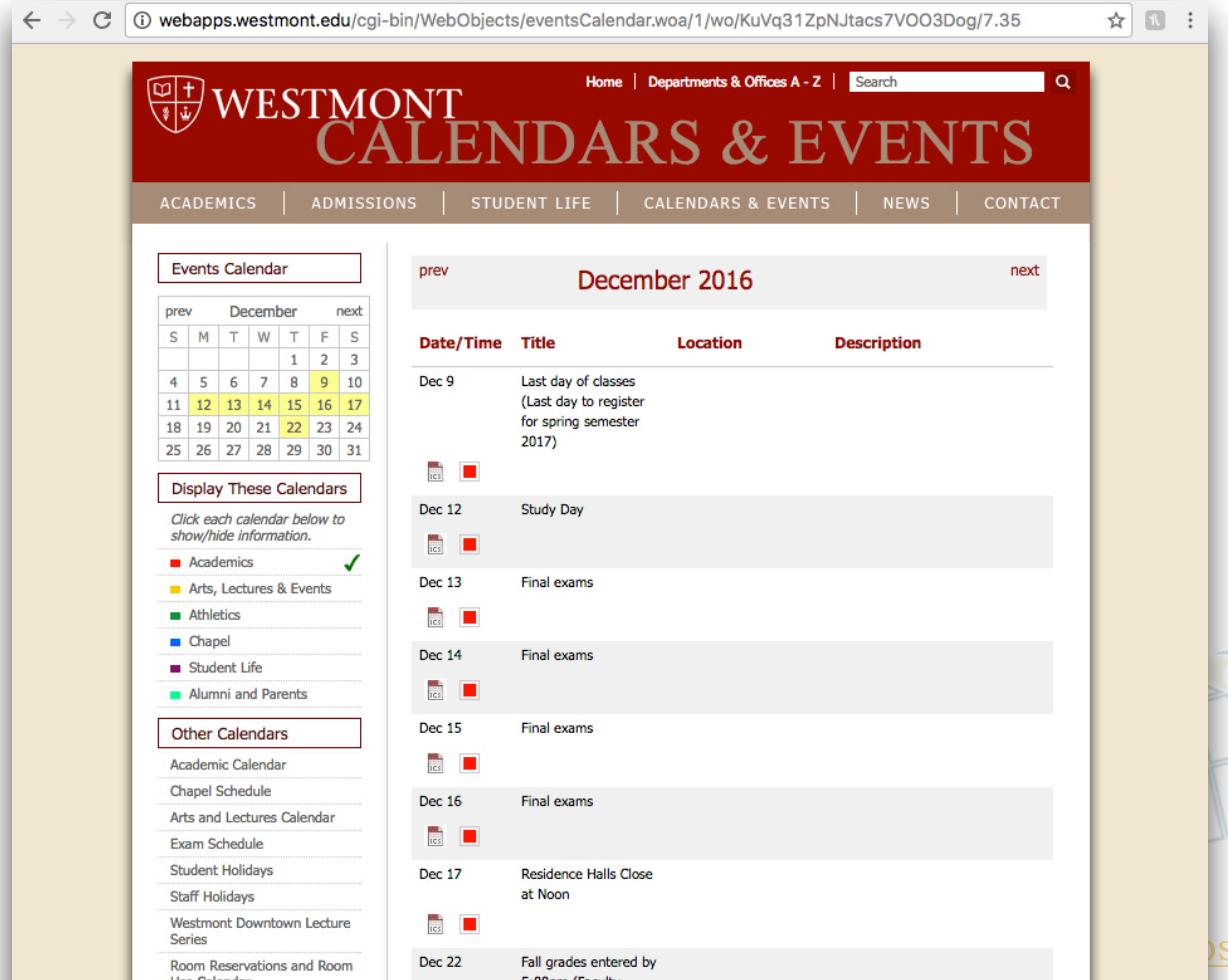
SCRAPY

INSTALLATION

https://scrapy.org/

SCRAPY - CASE STUDY 1

EXTRACT WESTMONT CALENDAR INFO



webapps.westmont.edu/cgi-bin/WebObjects/eventsCalendar.woa/1/wo/KuVq31ZpNJtacs7VOO3Dog/7.35

Home | Departments & Offices A - Z | Search
WESTMONT
CALENDARS & EVENTS
ACADEMICS | ADMISSIONS | STUDENT LIFE | CALENDARS & EVENTS | NEWS | CONTACT

Events Calendar

prev			December			next
S	M	T	W	T	F	S
				1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	31

Display These Calendars

Click each calendar below to show/hide information.

- 🟥 Academics ✓
- 🟨 Arts, Lectures & Events
- 🟩 Athletics
- 🟦 Chapel
- 🟪 Student Life
- 🟩 Alumni and Parents

Other Calendars

- Academic Calendar
- Chapel Schedule
- Arts and Lectures Calendar
- Exam Schedule
- Student Holidays
- Staff Holidays
- Westmont Downtown Lecture Series
- Room Reservations and Room

prev **December 2016** next

Date/Time	Title	Location	Description
Dec 9	Last day of classes (Last day to register for spring semester 2017)		
Dec 12	Study Day		
Dec 13	Final exams		
Dec 14	Final exams		
Dec 15	Final exams		
Dec 16	Final exams		
Dec 17	Residence Halls Close at Noon		
Dec 22	Fall grades entered by		

s://scrapy.org/

SCRAPY - CASE STUDY 1

QUICK INTRODUCTION TO XML

- XML is a text format
    - HTML, a subset, used to represent web pages
- Forms a tree structure

https://scrapy.org/

SCRAPY - CASE STUDY 1

QUICK INTRODUCTION TO XPATH

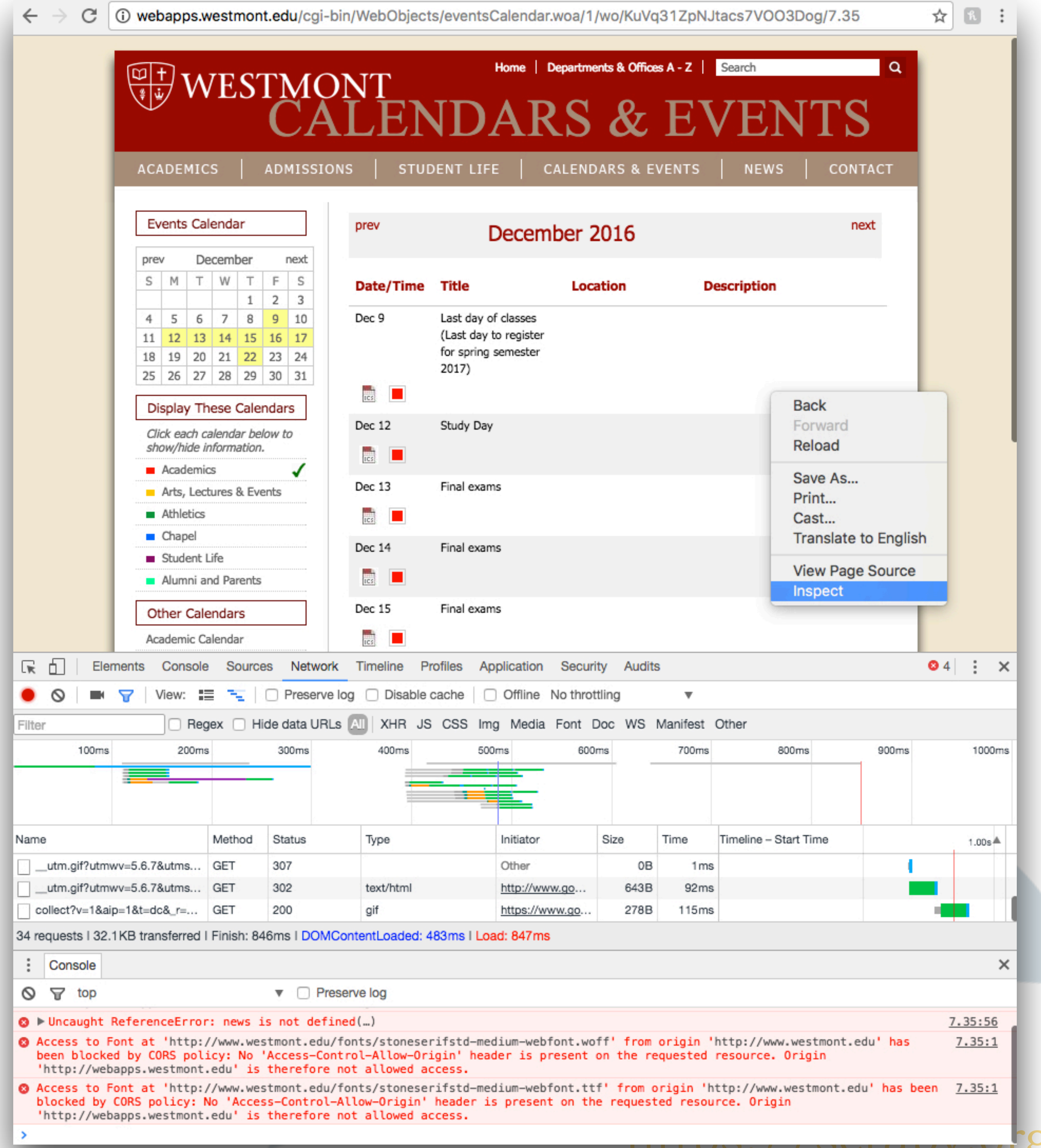
- XPATH is a query language
  - Used to select subtrees of a tree
- It looks a little like a file path

https://scrapy.org/

SCRAPY - CASE STUDY 1

QUICK INTRODUCTION TO BROWSER TOOLS

- Chrome Developer Tools
- View-Source


webapps.westmont.edu/cgi-bin/WebObjects/eventsCalendar.woa/1/wo/KuVq31ZpNJtacs7VOO3Dog/7.35
Home | Departments & Offices A - Z | Search
WESTMONT
CALENDARS & EVENTS
ACADEMICS | ADMISSIONS | STUDENT LIFE | CALENDARS & EVENTS | NEWS | CONTACT
Events Calendar
prev December next
Display These Calendars
Click each calendar below to show/hide information.
Academics ✓
Arts, Lectures & Events
Athletics
Chapel
Student Life
Alumni and Parents
Other Calendars
Academic Calendar
Back
Forward
Reload
Save As...
Print...
Cast...
Translate to English
View Page Source
Inspect
Elements Console Sources Network Timeline Profiles Application Security Audits
View: □ Preserve log □ Disable cache □ Offline No throttling
Filter □ Regex □ Hide data URLs All XHR JS CSS Img Media Font Doc WS Manifest Other
100ms 200ms 300ms 400ms 500ms 600ms 700ms 800ms 900ms 1000ms
Console
top □ Preserve log

S	M	T	W	T	F	S
				1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	31

prev December 2016 next

Date/Time	Title	Location	Description
Dec 9	Last day of classes (Last day to register for spring semester 2017)		
Dec 12	Study Day		
Dec 13	Final exams		
Dec 14	Final exams		
Dec 15	Final exams		

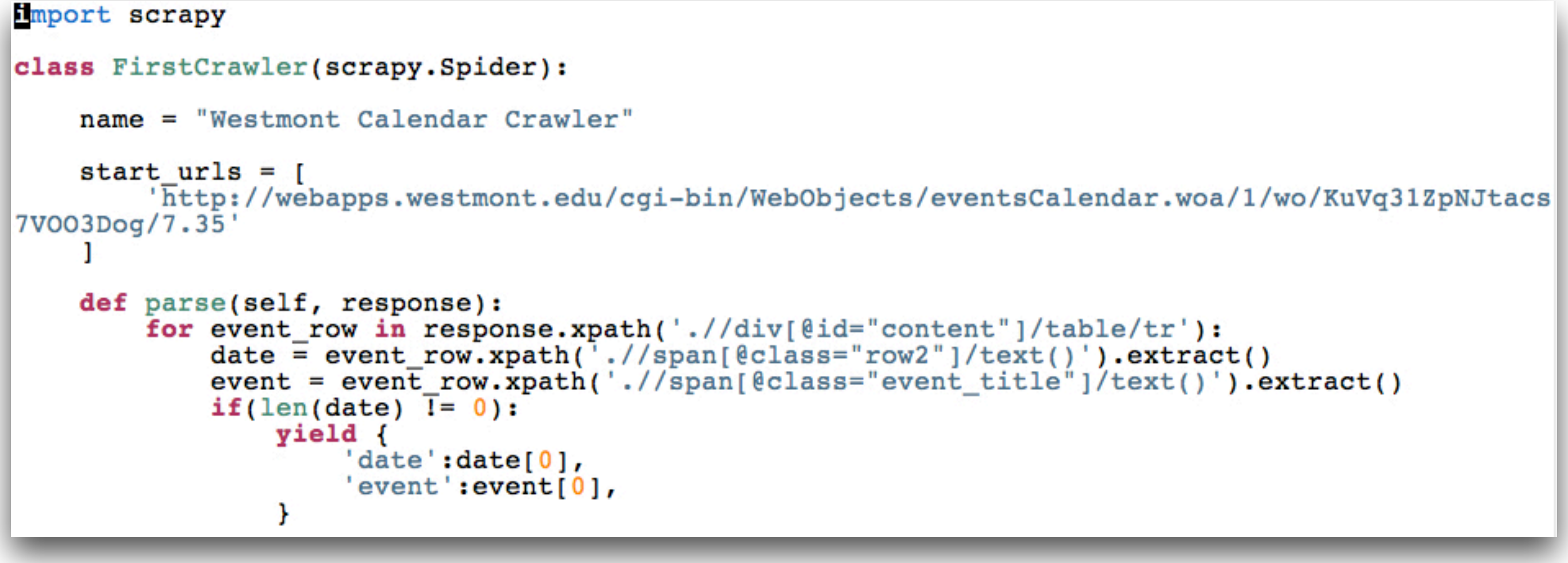
Name	Method	Status	Type	Initiator	Size	Time	Timeline – Start Time 1.00s
__utm.gif?utmwv=5.6.7&utms...	GET	307		Other	0B	1ms	
__utm.gif?utmwv=5.6.7&utms...	GET	302	text/html	http://www.go...	643B	92ms	
collect?v=1&aip=1&t=dc&_r=...	GET	200	gif	https://www.go...	278B	115ms	

34 requests | 32.1KB transferred | Finish: 846ms | DOMContentLoaded: 483ms | Load: 847ms

```
▶ Uncaught ReferenceError: news is not defined(…)                                                      7.35:56
Access to Font at 'http://www.westmont.edu/fonts/stoneserifstd-medium-webfont.woff' from origin 'http://www.westmont.edu' has   7.35:1
been blocked by CORS policy: No 'Access-Control-Allow-Origin' header is present on the requested resource. Origin
'http://webapps.westmont.edu' is therefore not allowed access.
Access to Font at 'http://www.westmont.edu/fonts/stoneserifstd-medium-webfont.ttf' from origin 'http://www.westmont.edu' has been   7.35:1
blocked by CORS policy: No 'Access-Control-Allow-Origin' header is present on the requested resource. Origin
'http://webapps.westmont.edu' is therefore not allowed access.
```

SCRAPY - CASE STUDY 1

EXTRACT WESTMONT CALENDAR INFO



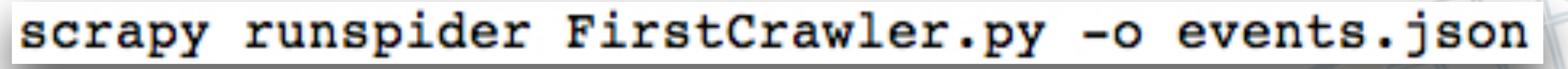
```python
import scrapy

class FirstCrawler(scrapy.Spider):

    name = "Westmont Calendar Crawler"

    start_urls = [
        'http://webapps.westmont.edu/cgi-bin/WebObjects/eventsCalendar.woa/1/wo/KuVq31ZpNJtacs
7VOO3Dog/7.35'
    ]

    def parse(self, response):
        for event_row in response.xpath('.//div[@id="content"]/table/tr'):
            date = event_row.xpath('.//span[@class="row2"]/text()').extract()
            event = event_row.xpath('.//span[@class="event_title"]/text()').extract()
            if(len(date) != 0):
                yield {
                    'date':date[0],
                    'event':event[0],
                }
```

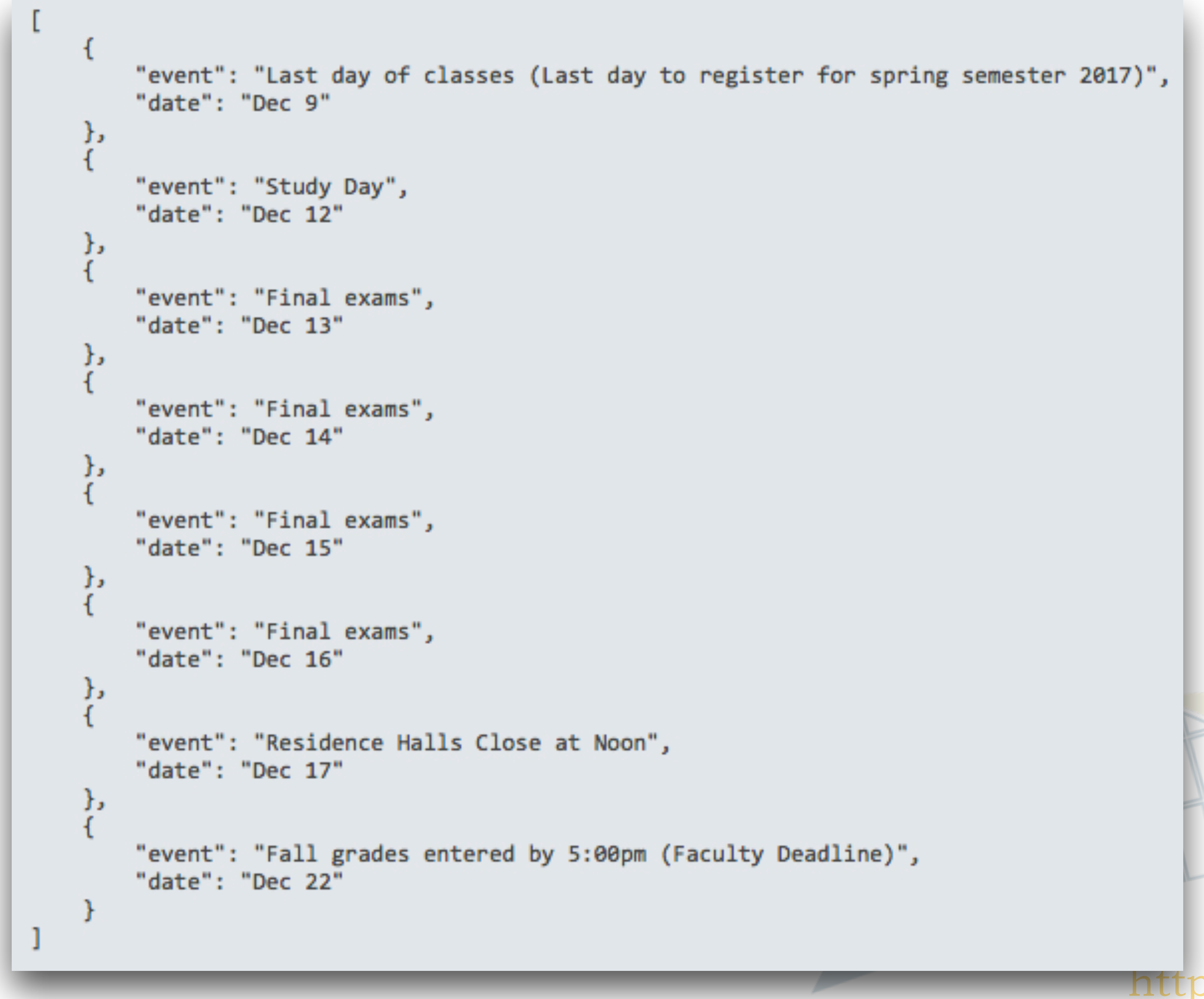


```
scrapy runspider FirstCrawler.py -o events.json
```

https://scrapy.org/

SCRAPY - CASE STUDY 1

EXTRACT WESTMONT CALENDAR INFO

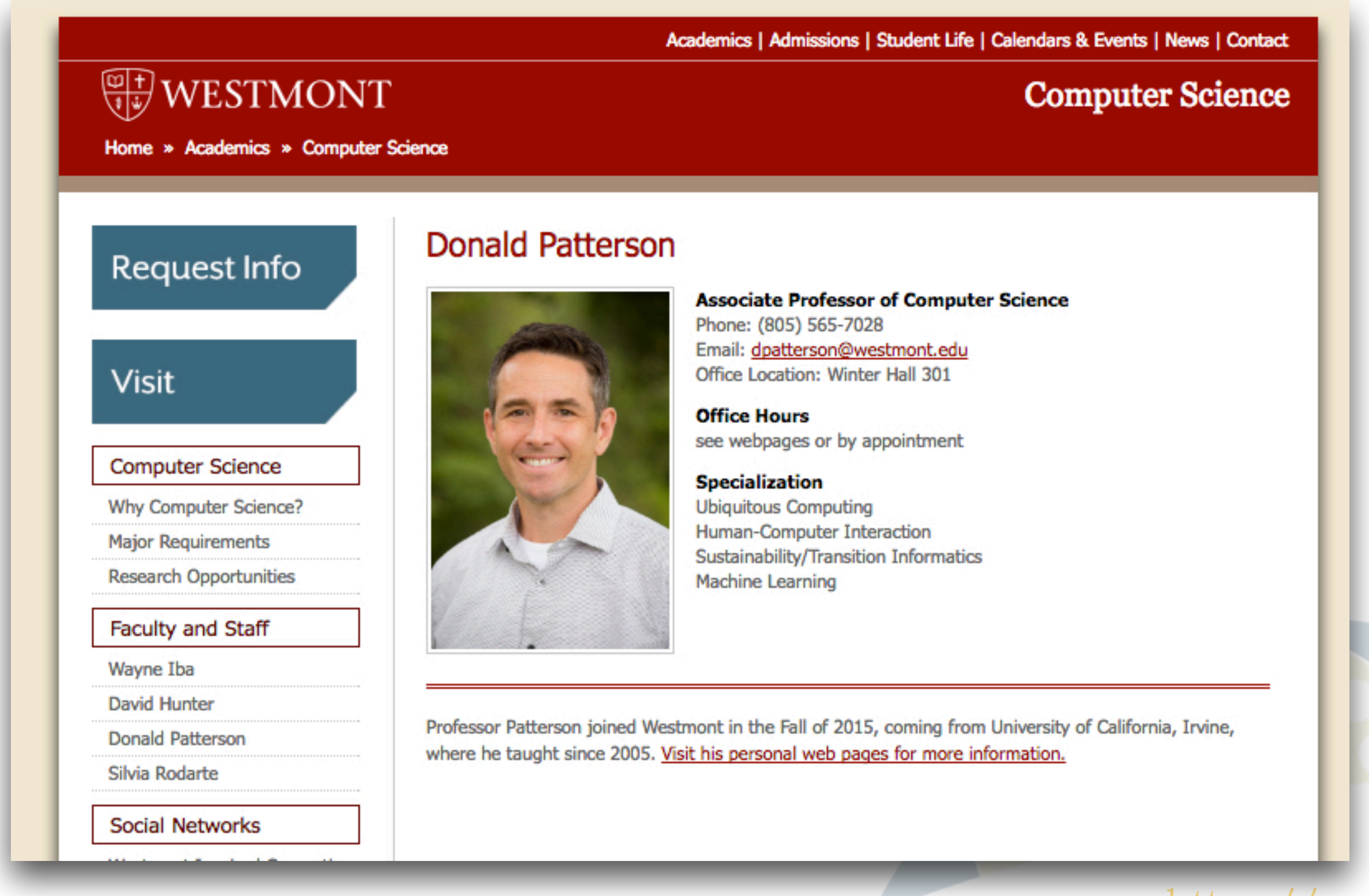


```
[
    {
        "event": "Last day of classes (Last day to register for spring semester 2017)",
        "date": "Dec 9"
    },
    {
        "event": "Study Day",
        "date": "Dec 12"
    },
    {
        "event": "Final exams",
        "date": "Dec 13"
    },
    {
        "event": "Final exams",
        "date": "Dec 14"
    },
    {
        "event": "Final exams",
        "date": "Dec 15"
    },
    {
        "event": "Final exams",
        "date": "Dec 16"
    },
    {
        "event": "Residence Halls Close at Noon",
        "date": "Dec 17"
    },
    {
        "event": "Fall grades entered by 5:00pm (Faculty Deadline)",
        "date": "Dec 22"
    }
]
```

https://scrapy.org/

SCRAPY - CASE STUDY 2

FIND ALL OCCURRENCES OF “PATTERSON”



Donald Patterson

Associate Professor of Computer Science
Phone: (805) 565-7028
Email: dpatterson@westmont.edu
Office Location: Winter Hall 301

Office Hours
see webpages or by appointment

Specialization
Ubiquitous Computing
Human-Computer Interaction
Sustainability/Transition Informatics
Machine Learning

Professor Patterson joined Westmont in the Fall of 2015, coming from University of California, Irvine, where he taught since 2005. Visit his personal web pages for more information.

https://scrapy.org/

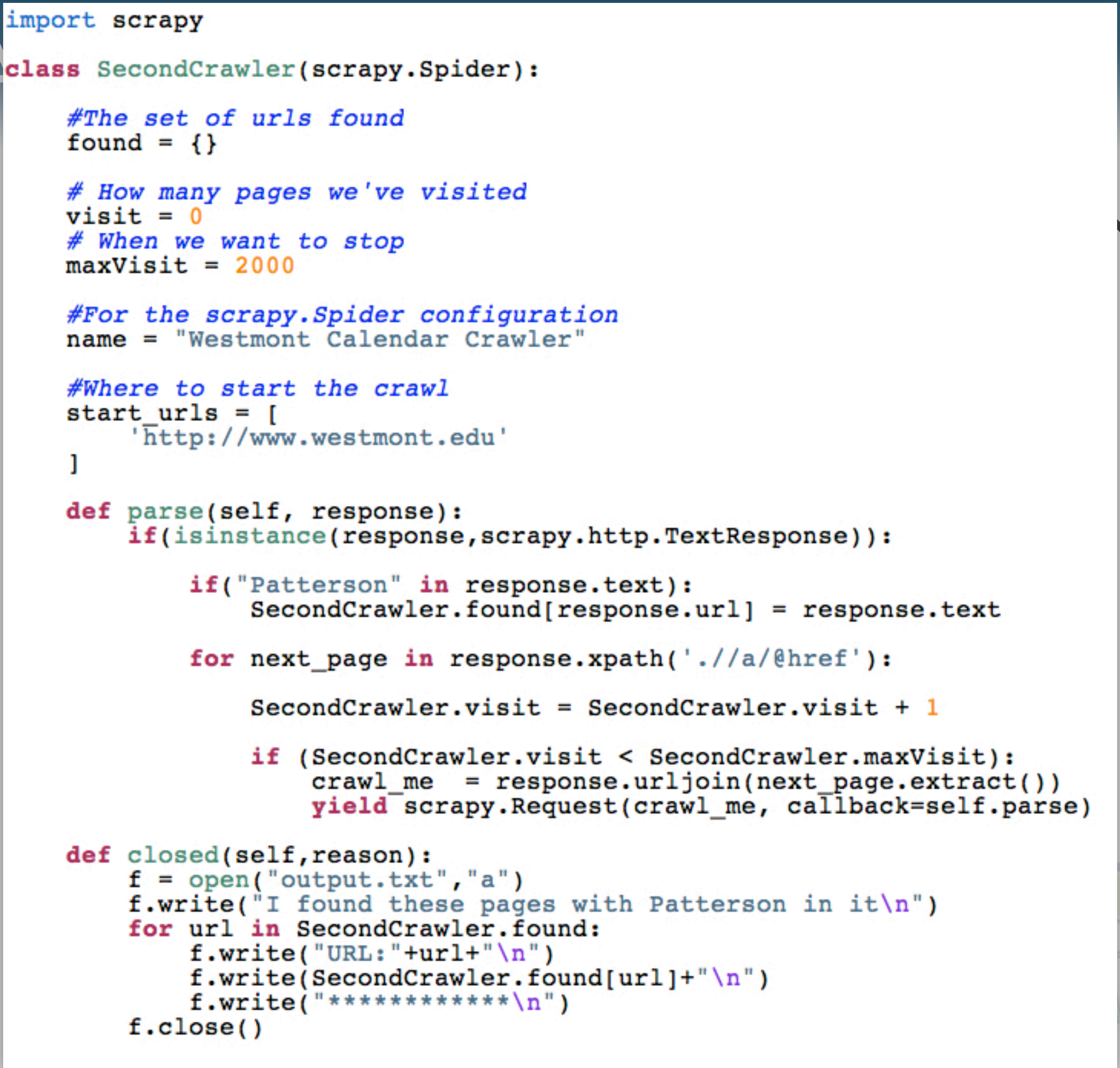
SCRAPY - CASE STUDY 2

FIND ALL OCCURRENCES OF “PATTERSON”

SCRA

FIND

N”



```
import scrapy

class SecondCrawler(scrapy.Spider):

    #The set of urls found
    found = {}

    # How many pages we've visited
    visit = 0
    # When we want to stop
    maxVisit = 2000

    #For the scrapy.Spider configuration
    name = "Westmont Calendar Crawler"

    #Where to start the crawl
    start_urls = [
        'http://www.westmont.edu'
    ]

    def parse(self, response):
        if(isinstance(response,scrapy.http.TextResponse)):

            if("Patterson" in response.text):
                SecondCrawler.found[response.url] = response.text

            for next_page in response.xpath('.//a/@href'):

                SecondCrawler.visit = SecondCrawler.visit + 1

                if (SecondCrawler.visit < SecondCrawler.maxVisit):
                    crawl_me  = response.urljoin(next_page.extract())
                    yield scrapy.Request(crawl_me, callback=self.parse)

    def closed(self,reason):
        f = open("output.txt","a")
        f.write("I found these pages with Patterson in it\n")
        for url in SecondCrawler.found:
            f.write("URL:"+url+"\n")
            f.write(SecondCrawler.found[url]+"\n")
            f.write("************\n")
        f.close()
```

https://scrapy.org/

SCRAPY - CASE STUDY 2

FIND ALL OCCURRENCES OF “PATTERSON”

https://scrapy.org/

WESTMONT INSPIRED

COMPUTING LAB