Web Crawler

Code and Documentation Written By
Steven Fan

Overview of System

The web crawler is written in Python 3 and mainly uses the Scrapy architecture for web crawling. The Windows operating system is utilized in building and testing this program. The following documentation will document the source code itself.

Import Code Blocks

```
import sys
```
Obtain user parameters passed in through the command line
```
import os
```
Creation and manipulation of files and directories
```
import scrapy
```
Web crawling
```
from scrapy.spiders import CrawlSpider, Rule
```
Utilization of a specific spider type: *CrawlSpider* and allows specifications of rules in the behavior of the spider
```
from scrapy.linkextractors import LinkExtractor
```
Extraction of web links from HTML to traverse the web
```
from scrapy.crawler import CrawlerProcess
```
Sets up the environment and process for the spider to function in
```
from scrapy.utils.project import get_project_settings
```
Extraction of spider settings to manipulate functionality of the spider
```
from scrapy.exceptions import CloseSpider
```
Shutting down of the spider, especially due to failed file writes or page limit exceeded

crawler Class Code Blocks

```
rules = (Rule\
    (
        LinkExtractor(allow = (r'^https?:\/\/(www\.)?[^.]*\.gov[^.]*$')),
        callback = 'parse_page',
        follow = True
    ),)
```

Within the *crawler* class, the rules are initialized for the spider. The *LinkExtractor* contains a regular expression that parses only web links that are government websites. The *callback* and *follow* rules allow the spider to crawl pages from these obtained links. By default, the Scrapy spider will also not fetch duplicate web links. Note that the current iteration of the web crawler only crawls through government websites. This is denoted by the regular expression that represents the respective syntax of these government websites.

```
def __init__(self, url = None):
    super().__init__()
    self.start_urls = url
```

In the *__init__* function in the *crawler* class, the function itself is imposed with *super()* in order to prevent inheritance issues. The *__init__* function also initializes the spider with the seed URLs provided by the user.

```
def parse_page(self, response):
```

The *parse_page* function deals with handling the HTML page obtained. The function writes the HTML page to its respective data file in the data folder. The function also counts how many pages (if applicable) the spider should generate depending on the user input. If a file fails to write or the limit of pages crawled is reached, the spider raises a flag through *CloseSpider()* and exits. The function obtains user input information through global variables.

Main Code Blocks

Lines 40 through 69 of the code handles user input and feedback. The code block simply parses the user input and determines the necessary parameters: seed file name and starting URLs, the number of pages the spider should crawl, the number of levels or depth the spider should be limited to crawling, and the folder and file name where the data should be stored in. The code block consists of *try* and *except* data structures in order to effectively provide feedback to the user on incorrect or useful input.

```
try:
    os.makedirs(directory, exist_ok = True)
except Exception:
    print('[ERROR] Failed to establish output directory')
    exit()
i = 0
filename = directory + '/' + directory + str(i) + '.html'
while os.path.isfile(filename):
    i += 1
    filename = filename[0:len(directory + '/' + directory)] + str(i) + '.html'
```

This code block handles how HTML data is stored. The *try* block creates (if it does not exist) the output directory that the user specifies. The program stores data as *<template>/<template><int>*. For example, if the user specifies *data* as the output directory, then the first HTML page will be stored in *data/data0*. The integer value will increment on each HTML page to be saved. The *while* loop allows the program to append to an already existing output directory. Therefore, if the output directory contains existing data, the program will create files where it last left off. For example, if files *data/data0* to *data/data50* exist, the program starts writing files at *data/data51*.

```
settings = get_project_settings()
settings.update({'DEPTH_LIMIT' : levels})
process = CrawlerProcess(settings)
spider = crawler()
process.crawl(spider, seed)
process.start()
```

The last code block deals with initialization of the spider. First, the settings for the Scrapy spider are obtained. The settings are updated to allow restrictions in the levels or depth of fetched URLs from the seed URLs. The spider process is then generated with these settings and the appropriate spider type: *CrawlSpider* and class *crawler* are initialized. The spider process is then initialized with the respective spider along with another parameter, the seed URLs, which will be passed into the __*init*__ function in the *crawler* class. The spider process is then executed, running the spider.

Limitations

The web crawler program contains a few limitations. Scrapy is built on a single-threaded system, and thus can not handle multi-threading of a spider class. However, Scrapy does allow parallelism where the next *GET* request for a web link can be initiated before the previous *GET* request has been returned. Another limitation is that the spider may fail to obtain information from particular websites. Generally, these issues include lack of authentication or limitations of *GET* requests imposed by the hosts themselves. The other limitation is that seed files must be prepended with *http://* or *https://* in order for them to be evaluated correctly.

Instructions

The web crawler is built on Scrapy. Thus, installation of Scrapy is required. It is recommended to install Scrapy on Anaconda when installing for Windows. The program files contain the source code *crawler.py*, and the seed URLs file *seed.txt*. The *seed.txt* is an optional file and may be manipulated in name, file type, and data. The user can also create their own seed file. To execute the program, the user enters the following command:

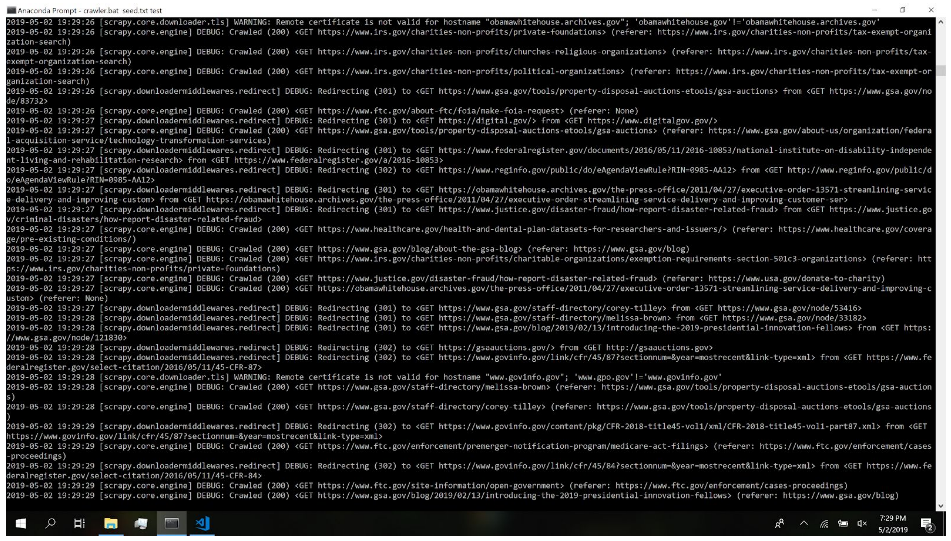
```
python crawler.py <seed file> [# pages] [# levels] <output directory>
```

where <> are required and `[]` are optional

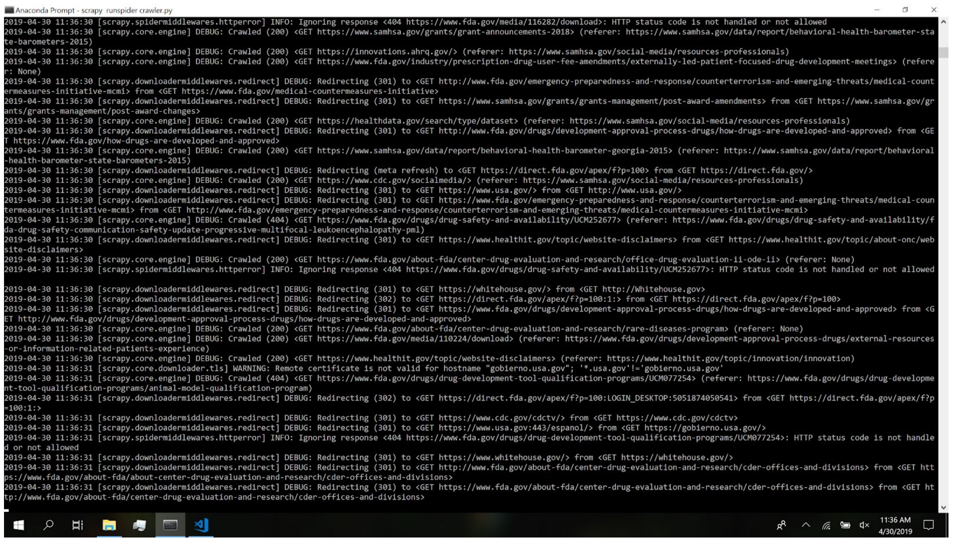
Note that inputting negative integers for pages and levels will be interpreted as infinite. The output directory will be the name of the folder and files the HTML code will be stored in.

Note: The web crawler program is compatible with the Web Document Index and Search program: https://github.com/steventfan/Web-Document-Index-and-Search

Screenshots

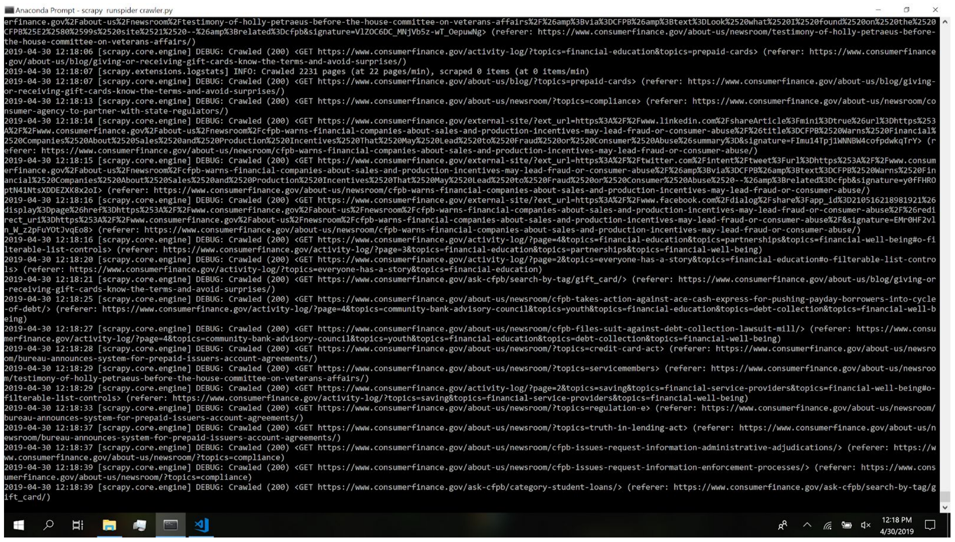


```
2019-05-02 19:29:26 [scrapy.core.downloader.tls] WARNING: Remote certificate is not valid for hostname "obamawhitehouse.archives.gov"; 'obamawhitehouse.gov'!='obamawhitehouse.archives.gov'
2019-05-02 19:29:26 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.irs.gov/charities-non-profits/private-foundations> (referer: https://www.irs.gov/charities-non-profits/tax-exempt-organi
zation-search)
2019-05-02 19:29:26 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.irs.gov/charities-non-profits/churches-religious-organizations> (referer: https://www.irs.gov/charities-non-profits/tax-
exempt-organization-search)
2019-05-02 19:29:26 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.irs.gov/charities-non-profits/political-organizations> (referer: https://www.irs.gov/charities-non-profits/tax-exempt-or
ganization-search)
2019-05-02 19:29:26 [scrapy.downloadermiddlewares.redirect] DEBUG: Redirecting (301) to <GET https://www.gsa.gov/tools/property-disposal-auctions-etools/gsa-auctions> from <GET https://www.gsa.gov/no
de/83732>
2019-05-02 19:29:26 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.ftc.gov/about-ftc/foia/make-foia-request> (referer: None)
2019-05-02 19:29:27 [scrapy.downloadermiddlewares.redirect] DEBUG: Redirecting (301) to <GET https://digital.gov/> from <GET https://www.digitalgov.gov/>
2019-05-02 19:29:27 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.gsa.gov/tools/property-disposal-auctions-etools/gsa-auctions> (referer: https://www.gsa.gov/about-us/organization/federa
l-acquisition-service/technology-transformation-services)
2019-05-02 19:29:27 [scrapy.downloadermiddlewares.redirect] DEBUG: Redirecting (301) to <GET https://www.federalregister.gov/documents/2016/05/11/2016-10853/national-institute-on-disability-independe
nt-living-and-rehabilitation-research> from <GET https://www.federalregister.gov/a/2016-10853>
2019-05-02 19:29:27 [scrapy.downloadermiddlewares.redirect] DEBUG: Redirecting (302) to <GET https://www.reginfo.gov/public/do/eAgendaViewRule?RIN=0985-AA12> from <GET http://www.reginfo.gov/public/d
o/eAgendaViewRule?RIN=0985-AA12>
2019-05-02 19:29:27 [scrapy.downloadermiddlewares.redirect] DEBUG: Redirecting (301) to <GET https://obamawhitehouse.archives.gov/the-press-office/2011/04/27/executive-order-13571-streamlining-servic
e-delivery-and-improving-custom> from <GET https://obamawhitehouse.archives.gov/the-press-office/2011/04/27/executive-order-streamlining-service-delivery-and-improving-customer-ser>
2019-05-02 19:29:27 [scrapy.downloadermiddlewares.redirect] DEBUG: Redirecting (301) to <GET https://www.justice.gov/disaster-fraud/how-report-disaster-related-fraud> from <GET https://www.justice.go
v/criminal-disasters/how-report-disaster-related-fraud>
2019-05-02 19:29:27 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.healthcare.gov/health-and-dental-plan-datasets-for-researchers-and-issuers/> (referer: https://www.healthcare.gov/covera
ge/pre-existing-conditions/)
2019-05-02 19:29:27 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.gsa.gov/blog/about-the-gsa-blog> (referer: https://www.gsa.gov/blog)
2019-05-02 19:29:27 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.irs.gov/charities-non-profits/charitable-organizations/exemption-requirements-section-501c3-organizations> (referer: htt
ps://www.irs.gov/charities-non-profits/private-foundations)
2019-05-02 19:29:27 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.justice.gov/disaster-fraud/how-report-disaster-related-fraud> (referer: https://www.usa.gov/donate-to-charity)
2019-05-02 19:29:27 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://obamawhitehouse.archives.gov/the-press-office/2011/04/27/executive-order-13571-streamlining-service-delivery-and-improving-c
ustom> (referer: None)
2019-05-02 19:29:27 [scrapy.downloadermiddlewares.redirect] DEBUG: Redirecting (301) to <GET https://www.gsa.gov/staff-directory/corey-tilley> from <GET https://www.gsa.gov/node/53416>
2019-05-02 19:29:28 [scrapy.downloadermiddlewares.redirect] DEBUG: Redirecting (301) to <GET https://www.gsa.gov/staff-directory/melissa-brown> from <GET https://www.gsa.gov/node/33182>
2019-05-02 19:29:28 [scrapy.downloadermiddlewares.redirect] DEBUG: Redirecting (301) to <GET https://www.gsa.gov/blog/2019/02/13/introducing-the-2019-presidential-innovation-fellows> from <GET https:
//www.gsa.gov/node/121830>
2019-05-02 19:29:28 [scrapy.downloadermiddlewares.redirect] DEBUG: Redirecting (302) to <GET https://gsaauctions.gov/> from <GET http://gsaauctions.gov>
2019-05-02 19:29:28 [scrapy.downloadermiddlewares.redirect] DEBUG: Redirecting (302) to <GET https://www.govinfo.gov/link/cfr/45/87?sectionnum=&year=mostrecent&link-type=xml> from <GET https://www.fe
deralregister.gov/select-citation/2016/05/11/45-CFR-87>
2019-05-02 19:29:28 [scrapy.core.downloader.tls] WARNING: Remote certificate is not valid for hostname "www.govinfo.gov"; 'www.gpo.gov'!='www.govinfo.gov'
2019-05-02 19:29:28 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.gsa.gov/staff-directory/melissa-brown> (referer: https://www.gsa.gov/tools/property-disposal-auctions-etools/gsa-auction
s)
2019-05-02 19:29:28 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.gsa.gov/staff-directory/corey-tilley> (referer: https://www.gsa.gov/tools/property-disposal-auctions-etools/gsa-auctions
)
2019-05-02 19:29:29 [scrapy.downloadermiddlewares.redirect] DEBUG: Redirecting (302) to <GET https://www.govinfo.gov/content/pkg/CFR-2018-title45-vol1/xml/CFR-2018-title45-vol1-part87.xml> from <GET
https://www.govinfo.gov/link/cfr/45/87?sectionnum=&year=mostrecent&link-type=xml>
2019-05-02 19:29:29 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.ftc.gov/enforcement/premerger-notification-program/medicare-act-filings> (referer: https://www.ftc.gov/enforcement/cases
-proceedings)
2019-05-02 19:29:29 [scrapy.downloadermiddlewares.redirect] DEBUG: Redirecting (302) to <GET https://www.govinfo.gov/link/cfr/45/84?sectionnum=&year=mostrecent&link-type=xml> from <GET https://www.fe
deralregister.gov/select-citation/2016/05/11/45-CFR-84>
2019-05-02 19:29:29 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.ftc.gov/site-information/open-government> (referer: https://www.ftc.gov/enforcement/cases-proceedings)
2019-05-02 19:29:29 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.gsa.gov/blog/2019/02/13/introducing-the-2019-presidential-innovation-fellows> (referer: https://www.gsa.gov/blog)
```

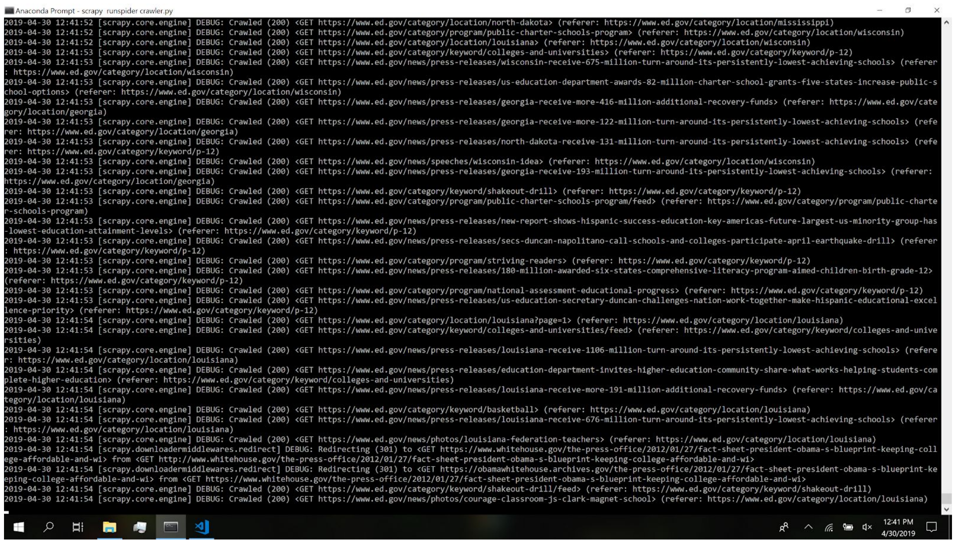


```
2019-04-30 11:36:30 [scrapy.spidermiddlewares.httperror] INFO: Ignoring response <404 https://www.fda.gov/media/116282/download>: HTTP status code is not handled or not allowed
2019-04-30 11:36:30 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.samhsa.gov/grants/grant-announcements-2018> (referer: https://www.samhsa.gov/data/report/behavioral-health-barometer-sta
te-barometers-2015)
2019-04-30 11:36:30 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://innovations.ahrq.gov/> (referer: https://www.samhsa.gov/social-media/resources-professionals)
2019-04-30 11:36:30 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.fda.gov/industry/prescription-drug-user-fee-amendments/externally-led-patient-focused-drug-development-meetings> (refere
r: None)
2019-04-30 11:36:30 [scrapy.downloadermiddlewares.redirect] DEBUG: Redirecting (301) to <GET http://www.fda.gov/emergency-preparedness-and-response/counterterrorism-and-emerging-threats/medical-count
ermeasures-initiative-mcmi> from <GET https://www.fda.gov/medical-countermeasures-initiative>
2019-04-30 11:36:30 [scrapy.downloadermiddlewares.redirect] DEBUG: Redirecting (301) to <GET https://www.samhsa.gov/grants/grants-management/post-award-amendments> from <GET https://www.samhsa.gov/gr
ants/grants-management/post-award-changes>
2019-04-30 11:36:30 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://healthdata.gov/search/type/dataset> (referer: https://www.samhsa.gov/social-media/resources-professionals)
2019-04-30 11:36:30 [scrapy.downloadermiddlewares.redirect] DEBUG: Redirecting (301) to <GET http://www.fda.gov/drugs/development-approval-process-drugs/how-drugs-are-developed-and-approved> from <GE
T https://www.fda.gov/how-drugs-are-developed-and-approved>
2019-04-30 11:36:30 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.samhsa.gov/data/report/behavioral-health-barometer-georgia-2015> (referer: https://www.samhsa.gov/data/report/behavioral
-health-barometer-state-barometers-2015)
2019-04-30 11:36:30 [scrapy.downloadermiddlewares.redirect] DEBUG: Redirecting (meta refresh) to <GET https://direct.fda.gov/apex/f?p=100> from <GET https://direct.fda.gov/>
2019-04-30 11:36:30 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.cdc.gov/socialmedia/> (referer: https://www.samhsa.gov/social-media/resources-professionals)
2019-04-30 11:36:30 [scrapy.downloadermiddlewares.redirect] DEBUG: Redirecting (301) to <GET https://www.usa.gov/> from <GET http://www.usa.gov/>
2019-04-30 11:36:30 [scrapy.downloadermiddlewares.redirect] DEBUG: Redirecting (301) to <GET https://www.fda.gov/emergency-preparedness-and-response/counterterrorism-and-emerging-threats/medical-coun
termeasures-initiative-mcmi> from <GET http://www.fda.gov/emergency-preparedness-and-response/counterterrorism-and-emerging-threats/medical-countermeasures-initiative-mcmi>
2019-04-30 11:36:30 [scrapy.core.engine] DEBUG: Crawled (404) <GET https://www.fda.gov/drugs/drug-safety-and-availability/UCM252677> (referer: https://www.fda.gov/drugs/drug-safety-and-availability/f
da-drug-safety-communication-safety-update-progressive-multifocal-leukoencephalopathy-pml)
2019-04-30 11:36:30 [scrapy.downloadermiddlewares.redirect] DEBUG: Redirecting (301) to <GET https://www.healthit.gov/topic/website-disclaimers> from <GET https://www.healthit.gov/topic/about-onc/web
site-disclaimers>
2019-04-30 11:36:30 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.fda.gov/about-fda/center-drug-evaluation-and-research/office-drug-evaluation-ii-ode-ii> (referer: None)
2019-04-30 11:36:30 [scrapy.spidermiddlewares.httperror] INFO: Ignoring response <404 https://www.fda.gov/drugs/drug-safety-and-availability/UCM252677>: HTTP status code is not handled or not allowed

2019-04-30 11:36:30 [scrapy.downloadermiddlewares.redirect] DEBUG: Redirecting (301) to <GET https://whitehouse.gov/> from <GET http://Whitehouse.gov>
2019-04-30 11:36:30 [scrapy.downloadermiddlewares.redirect] DEBUG: Redirecting (302) to <GET https://direct.fda.gov/apex/f?p=100:1:> from <GET https://direct.fda.gov/apex/f?p=100>
2019-04-30 11:36:30 [scrapy.downloadermiddlewares.redirect] DEBUG: Redirecting (301) to <GET https://www.fda.gov/drugs/development-approval-process-drugs/how-drugs-are-developed-and-approved> from <G
ET http://www.fda.gov/drugs/development-approval-process-drugs/how-drugs-are-developed-and-approved>
2019-04-30 11:36:30 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.fda.gov/about-fda/center-drug-evaluation-and-research/rare-diseases-program> (referer: None)
2019-04-30 11:36:30 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.fda.gov/media/110224/download> (referer: https://www.fda.gov/drugs/development-approval-process-drugs/external-resources
-or-information-related-patients-experience)
2019-04-30 11:36:30 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.healthit.gov/topic/website-disclaimers> (referer: https://www.healthit.gov/topic/innovation/innovation)
2019-04-30 11:36:31 [scrapy.core.downloader.tls] WARNING: Remote certificate is not valid for hostname "gobierno.usa.gov"; '*.usa.gov'!='gobierno.usa.gov'
2019-04-30 11:36:31 [scrapy.core.engine] DEBUG: Crawled (404) <GET https://www.fda.gov/drugs/drug-development-tool-qualification-programs/UCM077254> (referer: https://www.fda.gov/drugs/drug-developme
nt-tool-qualification-programs/animal-model-qualification-program)
2019-04-30 11:36:31 [scrapy.downloadermiddlewares.redirect] DEBUG: Redirecting (302) to <GET https://direct.fda.gov/apex/f?p=100:LOGIN_DESKTOP:5051874050541> from <GET https://direct.fda.gov/apex/f?p
=100:1:>
2019-04-30 11:36:31 [scrapy.downloadermiddlewares.redirect] DEBUG: Redirecting (301) to <GET https://www.cdc.gov/cdctv/> from <GET https://www.cdc.gov/cdctv>
2019-04-30 11:36:31 [scrapy.downloadermiddlewares.redirect] DEBUG: Redirecting (301) to <GET https://www.usa.gov:443/espanol/> from <GET https://gobierno.usa.gov/>
2019-04-30 11:36:31 [scrapy.spidermiddlewares.httperror] INFO: Ignoring response <404 https://www.fda.gov/drugs/drug-development-tool-qualification-programs/UCM077254>: HTTP status code is not handle
d or not allowed
2019-04-30 11:36:31 [scrapy.downloadermiddlewares.redirect] DEBUG: Redirecting (301) to <GET https://www.whitehouse.gov/> from <GET https://whitehouse.gov/>
2019-04-30 11:36:31 [scrapy.downloadermiddlewares.redirect] DEBUG: Redirecting (301) to <GET http://www.fda.gov/about-fda/center-drug-evaluation-and-research/cder-offices-and-divisions> from <GET htt
ps://www.fda.gov/about-fda/about-center-drug-evaluation-and-research/cder-offices-and-divisions>
2019-04-30 11:36:31 [scrapy.downloadermiddlewares.redirect] DEBUG: Redirecting (301) to <GET https://www.fda.gov/about-fda/center-drug-evaluation-and-research/cder-offices-and-divisions> from <GET ht
tp://www.fda.gov/about-fda/center-drug-evaluation-and-research/cder-offices-and-divisions>
```



```
erfinance.gov%2Fabout-us%2Fnewsroom%2Ftestimony-of-holly-petraeus-before-the-house-committee-on-veterans-affairs%2F%26amp%3Bvia%3DCFPB%26amp%3Btext%3DLook%2520what%2520I%2520found%2520on%2520the%2520
CFPB%25E2%2580%2599s%2520site%2521%2520--%26amp%3Brelated%3Dcfpb&signature=VlZOC6DC_MNjVb5z-wT_OepuwNg> (referer: https://www.consumerfinance.gov/about-us/newsroom/testimony-of-holly-petraeus-before-
the-house-committee-on-veterans-affairs/)
2019-04-30 12:18:06 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.consumerfinance.gov/activity-log/?topics=financial-education&topics=prepaid-cards> (referer: https://www.consumerfinance
.gov/about-us/blog/giving-or-receiving-gift-cards-know-the-terms-and-avoid-surprises/)
2019-04-30 12:18:07 [scrapy.extensions.logstats] INFO: Crawled 2231 pages (at 22 pages/min), scraped 0 items (at 0 items/min)
2019-04-30 12:18:07 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.consumerfinance.gov/about-us/blog/?topics=prepaid-cards> (referer: https://www.consumerfinance.gov/about-us/blog/giving-
or-receiving-gift-cards-know-the-terms-and-avoid-surprises/)
2019-04-30 12:18:13 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.consumerfinance.gov/about-us/newsroom/?topics=compliance> (referer: https://www.consumerfinance.gov/about-us/newsroom/co
nsumer-agency-to-partner-with-state-regulators/)
2019-04-30 12:18:14 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.consumerfinance.gov/external-site/?ext_url=https%3A%2F%2Fwww.linkedin.com%2FshareArticle%3Fmini%3Dtrue%26url%3Dhttps%253
A%2F%2Fwww.consumerfinance.gov%2Fabout-us%2Fnewsroom%2Fcfpb-warns-financial-companies-about-sales-and-production-incentives-may-lead-fraud-or-consumer-abuse%2F%26title%3DCFPB%2520Warns%2520Financial%
2520Companies%2520About%2520Sales%2520and%2520Production%2520Incentives%2520That%2520May%2520Lead%2520to%2520Fraud%2520or%2520Consumer%2520Abuse%26summary%3D&signature=FImu14Tpj1WNNBW4cofpdwkqTrY> (r
eferer: https://www.consumerfinance.gov/about-us/newsroom/cfpb-warns-financial-companies-about-sales-and-production-incentives-may-lead-fraud-or-consumer-abuse/)
2019-04-30 12:18:15 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.consumerfinance.gov/external-site/?ext_url=https%3A%2F%2Ftwitter.com%2Fintent%2Ftweet%3Furl%3Dhttps%253A%2F%2Fwww.consum
erfinance.gov%2Fabout-us%2Fnewsroom%2Fcfpb-warns-financial-companies-about-sales-and-production-incentives-may-lead-fraud-or-consumer-abuse%2F%26amp%3Bvia%3DCFPB%26amp%3Btext%3DCFPB%2520Warns%2520Fin
ancial%2520Companies%2520About%2520Sales%2520and%2520Production%2520Incentives%2520That%2520May%2520Lead%2520to%2520Fraud%2520or%2520Consumer%2520Abuse%2520--%26amp%3Brelated%3Dcfpb&signature=y0fFHRO
ptN41NtsXDDEZXK8x2oI> (referer: https://www.consumerfinance.gov/about-us/newsroom/cfpb-warns-financial-companies-about-sales-and-production-incentives-may-lead-fraud-or-consumer-abuse/)
2019-04-30 12:18:16 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.consumerfinance.gov/external-site/?ext_url=https%3A%2F%2Fwww.facebook.com%2Fdialog%2Fshare%3Fapp_id%3D210516218981921%26
display%3Dpage%26href%3Dhttps%253A%2F%2Fwww.consumerfinance.gov%2Fabout-us%2Fnewsroom%2Fcfpb-warns-financial-companies-about-sales-and-production-incentives-may-lead-fraud-or-consumer-abuse%2F%26redi
rect_uri%3Dhttps%253A%2F%2Fwww.consumerfinance.gov%2Fabout-us%2Fnewsroom%2Fcfpb-warns-financial-companies-about-sales-and-production-incentives-may-lead-fraud-or-consumer-abuse%2F&signature=EMr0HF2vl
n_W_z2pFuYOtJvqEo8> (referer: https://www.consumerfinance.gov/about-us/newsroom/cfpb-warns-financial-companies-about-sales-and-production-incentives-may-lead-fraud-or-consumer-abuse/)
2019-04-30 12:18:16 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.consumerfinance.gov/activity-log/?page=4&topics=financial-education&topics=partnerships&topics=financial-well-being#o-fi
lterable-list-controls> (referer: https://www.consumerfinance.gov/activity-log/?page=3&topics=financial-education&topics=partnerships&topics=financial-well-being)
2019-04-30 12:18:20 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.consumerfinance.gov/activity-log/?page=2&topics=everyone-has-a-story&topics=financial-education#o-filterable-list-contro
ls> (referer: https://www.consumerfinance.gov/activity-log/?topics=everyone-has-a-story&topics=financial-education)
2019-04-30 12:18:21 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.consumerfinance.gov/ask-cfpb/search-by-tag/gift_card/> (referer: https://www.consumerfinance.gov/about-us/blog/giving-or
-receiving-gift-cards-know-the-terms-and-avoid-surprises/)
2019-04-30 12:18:25 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.consumerfinance.gov/about-us/newsroom/cfpb-takes-action-against-ace-cash-express-for-pushing-payday-borrowers-into-cycle
-of-debt/> (referer: https://www.consumerfinance.gov/activity-log/?page=4&topics=community-bank-advisory-council&topics=youth&topics=financial-education&topics=debt-collection&topics=financial-well-b
eing)
2019-04-30 12:18:27 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.consumerfinance.gov/about-us/newsroom/cfpb-files-suit-against-debt-collection-lawsuit-mill/> (referer: https://www.consu
merfinance.gov/activity-log/?page=4&topics=community-bank-advisory-council&topics=youth&topics=financial-education&topics=debt-collection&topics=financial-well-being)
2019-04-30 12:18:28 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.consumerfinance.gov/about-us/newsroom/?topics=credit-card-act> (referer: https://www.consumerfinance.gov/about-us/newsro
om/bureau-announces-system-for-prepaid-issuers-account-agreements/)
2019-04-30 12:18:29 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.consumerfinance.gov/about-us/newsroom/?topics=servicemembers> (referer: https://www.consumerfinance.gov/about-us/newsroo
m/testimony-of-holly-petraeus-before-the-house-committee-on-veterans-affairs/)
2019-04-30 12:18:29 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.consumerfinance.gov/activity-log/?page=2&topics=saving&topics=financial-service-providers&topics=financial-well-being#o-
filterable-list-controls> (referer: https://www.consumerfinance.gov/activity-log/?topics=saving&topics=financial-service-providers&topics=financial-well-being)
2019-04-30 12:18:33 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.consumerfinance.gov/about-us/newsroom/?topics=regulation-e> (referer: https://www.consumerfinance.gov/about-us/newsroom/
bureau-announces-system-for-prepaid-issuers-account-agreements/)
2019-04-30 12:18:37 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.consumerfinance.gov/about-us/newsroom/?topics=truth-in-lending-act> (referer: https://www.consumerfinance.gov/about-us/n
ewsroom/bureau-announces-system-for-prepaid-issuers-account-agreements/)
2019-04-30 12:18:37 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.consumerfinance.gov/about-us/newsroom/cfpb-issues-request-information-administrative-adjudications/> (referer: https://w
ww.consumerfinance.gov/about-us/newsroom/?topics=compliance)
2019-04-30 12:18:39 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.consumerfinance.gov/about-us/newsroom/cfpb-issues-request-information-enforcement-processes/> (referer: https://www.cons
umerfinance.gov/about-us/newsroom/?topics=compliance)
2019-04-30 12:18:39 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.consumerfinance.gov/ask-cfpb/category-student-loans/> (referer: https://www.consumerfinance.gov/ask-cfpb/search-by-tag/g
ift_card/)
```



```
2019-04-30 12:41:52 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.ed.gov/category/location/north-dakota> (referer: https://www.ed.gov/category/location/mississippi)
2019-04-30 12:41:52 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.ed.gov/category/program/public-charter-schools-program> (referer: https://www.ed.gov/category/location/wisconsin)
2019-04-30 12:41:52 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.ed.gov/category/location/louisiana> (referer: https://www.ed.gov/category/location/wisconsin)
2019-04-30 12:41:53 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.ed.gov/category/keyword/colleges-and-universities> (referer: https://www.ed.gov/category/keyword/p-12)
2019-04-30 12:41:53 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.ed.gov/news/press-releases/wisconsin-receive-675-million-turn-around-its-persistently-lowest-achieving-schools> (referer
: https://www.ed.gov/category/location/wisconsin)
2019-04-30 12:41:53 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.ed.gov/news/press-releases/us-education-department-awards-82-million-charter-school-grants-five-states-increase-public-s
chool-options> (referer: https://www.ed.gov/category/location/wisconsin)
2019-04-30 12:41:53 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.ed.gov/news/press-releases/georgia-receive-more-416-million-additional-recovery-funds> (referer: https://www.ed.gov/cate
gory/location/georgia)
2019-04-30 12:41:53 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.ed.gov/news/press-releases/georgia-receive-more-122-million-turn-around-its-persistently-lowest-achieving-schools> (refe
rer: https://www.ed.gov/category/location/georgia)
2019-04-30 12:41:53 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.ed.gov/news/press-releases/north-dakota-receive-131-million-turn-around-its-persistently-lowest-achieving-schools> (refe
rer: https://www.ed.gov/category/keyword/p-12)
2019-04-30 12:41:53 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.ed.gov/news/speeches/wisconsin-idea> (referer: https://www.ed.gov/category/location/wisconsin)
2019-04-30 12:41:53 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.ed.gov/news/press-releases/georgia-receive-193-million-turn-around-its-persistently-lowest-achieving-schools> (referer:
https://www.ed.gov/category/location/georgia)
2019-04-30 12:41:53 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.ed.gov/category/keyword/shakeout-drill> (referer: https://www.ed.gov/category/keyword/p-12)
2019-04-30 12:41:53 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.ed.gov/category/program/public-charter-schools-program/feed> (referer: https://www.ed.gov/category/program/public-charte
r-schools-program)
2019-04-30 12:41:53 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.ed.gov/news/press-releases/new-report-shows-hispanic-success-education-key-americas-future-largest-us-minority-group-has
-lowest-education-attainment-levels> (referer: https://www.ed.gov/category/keyword/p-12)
2019-04-30 12:41:53 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.ed.gov/news/press-releases/secs-duncan-napolitano-call-schools-and-colleges-participate-april-earthquake-drill> (referer
: https://www.ed.gov/category/keyword/p-12)
2019-04-30 12:41:53 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.ed.gov/category/program/striving-readers> (referer: https://www.ed.gov/category/keyword/p-12)
2019-04-30 12:41:53 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.ed.gov/news/press-releases/180-million-awarded-six-states-comprehensive-literacy-program-aimed-children-birth-grade-12>
(referer: https://www.ed.gov/category/keyword/p-12)
2019-04-30 12:41:53 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.ed.gov/category/program/national-assessment-educational-progress> (referer: https://www.ed.gov/category/keyword/p-12)
2019-04-30 12:41:53 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.ed.gov/news/press-releases/us-education-secretary-duncan-challenges-nation-work-together-make-hispanic-educational-excel
lence-priority> (referer: https://www.ed.gov/category/keyword/p-12)
2019-04-30 12:41:54 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.ed.gov/category/location/louisiana?page=1> (referer: https://www.ed.gov/category/location/louisiana)
2019-04-30 12:41:54 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.ed.gov/category/keyword/colleges-and-universities/feed> (referer: https://www.ed.gov/category/keyword/colleges-and-unive
rsities)
2019-04-30 12:41:54 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.ed.gov/news/press-releases/louisiana-receive-1106-million-turn-around-its-persistently-lowest-achieving-schools> (refere
r: https://www.ed.gov/category/location/louisiana)
2019-04-30 12:41:54 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.ed.gov/news/press-releases/education-department-invites-higher-education-community-share-what-works-helping-students-com
plete-higher-education> (referer: https://www.ed.gov/category/keyword/colleges-and-universities)
2019-04-30 12:41:54 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.ed.gov/news/press-releases/louisiana-receive-more-191-million-additional-recovery-funds> (referer: https://www.ed.gov/ca
tegory/location/louisiana)
2019-04-30 12:41:54 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.ed.gov/category/keyword/basketball> (referer: https://www.ed.gov/category/location/louisiana)
2019-04-30 12:41:54 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.ed.gov/news/press-releases/louisiana-receive-676-million-turn-around-its-persistently-lowest-achieving-schools> (referer
: https://www.ed.gov/category/location/louisiana)
2019-04-30 12:41:54 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.ed.gov/news/photos/louisiana-federation-teachers> (referer: https://www.ed.gov/category/location/louisiana)
2019-04-30 12:41:54 [scrapy.downloadermiddlewares.redirect] DEBUG: Redirecting (301) to <GET https://www.whitehouse.gov/the-press-office/2012/01/27/fact-sheet-president-obama-s-blueprint-keeping-coll
ege-affordable-and-wi> from <GET http://www.whitehouse.gov/the-press-office/2012/01/27/fact-sheet-president-obama-s-blueprint-keeping-college-affordable-and-wi>
2019-04-30 12:41:54 [scrapy.downloadermiddlewares.redirect] DEBUG: Redirecting (301) to <GET https://obamawhitehouse.archives.gov/the-press-office/2012/01/27/fact-sheet-president-obama-s-blueprint-ke
eping-college-affordable-and-wi> from <GET https://www.whitehouse.gov/the-press-office/2012/01/27/fact-sheet-president-obama-s-blueprint-keeping-college-affordable-and-wi>
2019-04-30 12:41:54 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.ed.gov/category/keyword/shakeout-drill/feed> (referer: https://www.ed.gov/category/keyword/shakeout-drill)
2019-04-30 12:41:54 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.ed.gov/news/photos/courage-classroom-js-clark-magnet-school> (referer: https://www.ed.gov/category/location/louisiana)
```