Nhóm6_28_2_2024

February 28, 2024

1 Thực hành các công cụ thu thập dữ liệu

1.1 Cài đặt

```
[ ]: !pip install scrapy
```

```
Collecting scrapy
  Downloading Scrapy-2.11.1-py2.py3-none-any.whl (287 kB)
                           287.8/287.8

kB 9.1 MB/s eta 0:00:00
Collecting Twisted>=18.9.0 (from scrapy)
  Downloading twisted-23.10.0-py3-none-any.whl (3.2 MB)
                           3.2/3.2 MB
22.2 MB/s eta 0:00:00
Requirement already satisfied: cryptography>=36.0.0 in
/usr/local/lib/python3.10/dist-packages (from scrapy) (42.0.3)
Collecting cssselect>=0.9.1 (from scrapy)
  Downloading cssselect-1.2.0-py2.py3-none-any.whl (18 kB)
Collecting itemloaders>=1.0.1 (from scrapy)
  Downloading itemloaders-1.1.0-py3-none-any.whl (11 kB)
Collecting parsel>=1.5.0 (from scrapy)
  Downloading parsel-1.8.1-py2.py3-none-any.whl (17 kB)
Requirement already satisfied: pyOpenSSL>=21.0.0 in
/usr/local/lib/python3.10/dist-packages (from scrapy) (24.0.0)
Collecting queuelib>=1.4.2 (from scrapy)
  Downloading queuelib-1.6.2-py2.py3-none-any.whl (13 kB)
Collecting service-identity>=18.1.0 (from scrapy)
  Downloading service_identity-24.1.0-py3-none-any.whl (12 kB)
Collecting w3lib>=1.17.0 (from scrapy)
  Downloading w3lib-2.1.2-py3-none-any.whl (21 kB)
Collecting zope.interface>=5.1.0 (from scrapy)
  Downloading zope.interface-6.2-cp310-cp310-
manylinux_2_5_x86_64.manylinux1_x86_64.manylinux_2_17_x86_64.manylinux2014_x86_6
4.whl (247 kB)
                           247.3/247.3

kB 32.5 MB/s eta 0:00:00
Collecting protego>=0.1.15 (from scrapy)
```

```
  Downloading Protego-0.3.0-py2.py3-none-any.whl (8.5 kB)
Collecting itemadapter>=0.1.0 (from scrapy)
  Downloading itemadapter-0.8.0-py3-none-any.whl (11 kB)
Requirement already satisfied: setuptools in /usr/local/lib/python3.10/dist-
packages (from scrapy) (67.7.2)
Requirement already satisfied: packaging in /usr/local/lib/python3.10/dist-
packages (from scrapy) (23.2)
Collecting tldextract (from scrapy)
  Downloading tldextract-5.1.1-py3-none-any.whl (97 kB)
                           97.7/97.7 kB
15.1 MB/s eta 0:00:00
Requirement already satisfied: lxml>=4.4.1 in
/usr/local/lib/python3.10/dist-packages (from scrapy) (4.9.4)
Collecting PyDispatcher>=2.0.5 (from scrapy)
  Downloading PyDispatcher-2.0.7-py3-none-any.whl (12 kB)
Requirement already satisfied: cffi>=1.12 in /usr/local/lib/python3.10/dist-
packages (from cryptography>=36.0.0->scrapy) (1.16.0)
Collecting jmespath>=0.9.5 (from itemloaders>=1.0.1->scrapy)
  Downloading jmespath-1.0.1-py3-none-any.whl (20 kB)
Requirement already satisfied: attrs>=19.1.0 in /usr/local/lib/python3.10/dist-
packages (from service-identity>=18.1.0->scrapy) (23.2.0)
Requirement already satisfied: pyasn1 in /usr/local/lib/python3.10/dist-packages
(from service-identity>=18.1.0->scrapy) (0.5.1)
Requirement already satisfied: pyasn1-modules in /usr/local/lib/python3.10/dist-
packages (from service-identity>=18.1.0->scrapy) (0.3.0)
Collecting automat>=0.8.0 (from Twisted>=18.9.0->scrapy)
  Downloading Automat-22.10.0-py2.py3-none-any.whl (26 kB)
Collecting constantly>=15.1 (from Twisted>=18.9.0->scrapy)
  Downloading constantly-23.10.4-py3-none-any.whl (13 kB)
Collecting hyperlink>=17.1.1 (from Twisted>=18.9.0->scrapy)
  Downloading hyperlink-21.0.0-py2.py3-none-any.whl (74 kB)
                           74.6/74.6 kB
12.2 MB/s eta 0:00:00
Collecting incremental>=22.10.0 (from Twisted>=18.9.0->scrapy)
  Downloading incremental-22.10.0-py2.py3-none-any.whl (16 kB)
Requirement already satisfied: typing-extensions>=4.2.0 in
/usr/local/lib/python3.10/dist-packages (from Twisted>=18.9.0->scrapy) (4.9.0)
Requirement already satisfied: idna in /usr/local/lib/python3.10/dist-packages
(from tldextract->scrapy) (3.6)
Requirement already satisfied: requests>=2.1.0 in
/usr/local/lib/python3.10/dist-packages (from tldextract->scrapy) (2.31.0)
Collecting requests-file>=1.4 (from tldextract->scrapy)
  Downloading requests_file-2.0.0-py2.py3-none-any.whl (4.2 kB)
Requirement already satisfied: filelock>=3.0.8 in
/usr/local/lib/python3.10/dist-packages (from tldextract->scrapy) (3.13.1)
Requirement already satisfied: six in /usr/local/lib/python3.10/dist-packages
(from automat>=0.8.0->Twisted>=18.9.0->scrapy) (1.16.0)
Requirement already satisfied: pycparser in /usr/local/lib/python3.10/dist-
```

```
packages (from cffi>=1.12->cryptography>=36.0.0->scrapy) (2.21)
Requirement already satisfied: charset-normalizer<4,>=2 in
/usr/local/lib/python3.10/dist-packages (from
requests>=2.1.0->tldextract->scrapy) (3.3.2)
Requirement already satisfied: urllib3<3,>=1.21.1 in
/usr/local/lib/python3.10/dist-packages (from
requests>=2.1.0->tldextract->scrapy) (2.0.7)
Requirement already satisfied: certifi>=2017.4.17 in
/usr/local/lib/python3.10/dist-packages (from
requests>=2.1.0->tldextract->scrapy) (2024.2.2)
Installing collected packages: PyDispatcher, incremental, zope.interface, w3lib,
queuelib, protego, jmespath, itemadapter, hyperlink, cssselect, constantly,
automat, Twisted, requests-file, parsel, tldextract, service-identity,
itemloaders, scrapy
Successfully installed PyDispatcher-2.0.7 Twisted-23.10.0 automat-22.10.0
constantly-23.10.4 cssselect-1.2.0 hyperlink-21.0.0 incremental-22.10.0
itemadapter-0.8.0 itemloaders-1.1.0 jmespath-1.0.1 parsel-1.8.1 protego-0.3.0
queuelib-1.6.2 requests-file-2.0.0 scrapy-2.11.1 service-identity-24.1.0
tldextract-5.1.1 w3lib-2.1.2 zope.interface-6.2
```

1.2 Các bước sử dụng công cụ

1. Tạo một project Scrapy mới bằng lệnh: `scrapy startproject myproject`
2. Tạo một file spider mới: Di chuyển đến thư mục spiders của project (`cd myproject/myproject/spiders`) và chạy lệnh sau: `scrapy genspider spider_name website_url`
3. Chỉnh sửa Spider: Mở file Spider (spiders/spider_name.py) bằng trình soạn thảo văn bản và chỉnh sửa nội dung của phương thức parse để xử lý dữ liệu từ trang web.
4. Chạy spider để bắt đầu quá trình thu thập dữ liệu: Di chuyển đến thư mục gốc của project và chạy lệnh sau: `scrapy crawl spider_name -o output_name`

1.3 Ví dụ minh họa

[107]:

```
!scrapy startproject oral_cancer_images
```

```
New Scrapy project 'oral_cancer_images', using template directory
'/usr/local/lib/python3.10/dist-packages/scrapy/templates/project', created in:
    /content/oral_cancer_images

You can start your first spider with:
    cd oral_cancer_images
    scrapy genspider example example.com
```

[108]:

```
%cd oral_cancer_images/oral_cancer_images/spiders/
```

```
/content/oral_cancer_images/oral_cancer_images/spiders
```

[109]:

```
!scrapy genspider images_spider "https://oralcancerfoundation.org/dental/
  ↪oral-cancer-images"
```

```
Created spider 'images_spider' using template 'basic' in module:
  oral_cancer_images.spiders.images_spider
```

[112]:

```
%cd ~
%cd /content/oral_cancer_images
```

```
/root
/content/oral_cancer_images
```

[114]:

```
!scrapy crawl images_spider -o output.json
```

```
2024-02-28 08:48:08 [scrapy.utils.log] INFO: Scrapy 2.11.1 started (bot:
oral_cancer_images)
2024-02-28 08:48:08 [scrapy.utils.log] INFO: Versions: lxml 4.9.4.0, libxml2
2.10.3, cssselect 1.2.0, parsel 1.8.1, w3lib 2.1.2, Twisted 23.10.0, Python
3.10.12 (main, Nov 20 2023, 15:14:05) [GCC 11.4.0], pyOpenSSL 24.0.0 (OpenSSL
3.2.1 30 Jan 2024), cryptography 42.0.3, Platform Linux-6.1.58+-x86_64-with-
glibc2.35
2024-02-28 08:48:08 [scrapy.addons] INFO: Enabled addons:
[]
2024-02-28 08:48:08 [asyncio] DEBUG: Using selector: EpollSelector
2024-02-28 08:48:08 [scrapy.utils.log] DEBUG: Using reactor:
twisted.internet.asyncioreactor.AsyncioSelectorReactor
2024-02-28 08:48:08 [scrapy.utils.log] DEBUG: Using asyncio event loop:
asyncio.unix_events._UnixSelectorEventLoop
2024-02-28 08:48:08 [scrapy.extensions.telnet] INFO: Telnet Password:
3ce9b96c1634dad0
2024-02-28 08:48:08 [scrapy.middleware] INFO: Enabled extensions:
['scrapy.extensions.corestats.CoreStats',
 'scrapy.extensions.telnet.TelnetConsole',
 'scrapy.extensions.memusage.MemoryUsage',
 'scrapy.extensions.feedexport.FeedExporter',
 'scrapy.extensions.logstats.LogStats']
2024-02-28 08:48:08 [scrapy.crawler] INFO: Overridden settings:
{'BOT_NAME': 'oral_cancer_images',
 'FEED_EXPORT_ENCODING': 'utf-8',
 'NEWSPIDER_MODULE': 'oral_cancer_images.spiders',
 'REQUEST_FINGERPRINTER_IMPLEMENTATION': '2.7',
 'ROBOTSTXT_OBEY': True,
 'SPIDER_MODULES': ['oral_cancer_images.spiders'],
 'TWISTED_REACTOR': 'twisted.internet.asyncioreactor.AsyncioSelectorReactor'}
2024-02-28 08:48:08 [scrapy.middleware] INFO: Enabled downloader middlewares:
['scrapy.downloadermiddlewares.robotstxt.RobotsTxtMiddleware',
 'scrapy.downloadermiddlewares.httpauth.HttpAuthMiddleware',
 'scrapy.downloadermiddlewares.downloadtimeout.DownloadTimeoutMiddleware',
 'scrapy.downloadermiddlewares.defaultheaders.DefaultHeadersMiddleware',
 'scrapy.downloadermiddlewares.useragent.UserAgentMiddleware',
 'scrapy.downloadermiddlewares.retry.RetryMiddleware',
 'scrapy.downloadermiddlewares.redirect.MetaRefreshMiddleware',
 'scrapy.downloadermiddlewares.httpcompression.HttpCompressionMiddleware',
 'scrapy.downloadermiddlewares.redirect.RedirectMiddleware',
 'scrapy.downloadermiddlewares.cookies.CookiesMiddleware',
 'scrapy.downloadermiddlewares.httpproxy.HttpProxyMiddleware',
 'scrapy.downloadermiddlewares.stats.DownloaderStats']
2024-02-28 08:48:08 [scrapy.middleware] INFO: Enabled spider middlewares:
['scrapy.spidermiddlewares.httperror.HttpErrorMiddleware',
 'scrapy.spidermiddlewares.offsite.OffsiteMiddleware',
 'scrapy.spidermiddlewares.referer.RefererMiddleware',
 'scrapy.spidermiddlewares.urllength.UrlLengthMiddleware',
 'scrapy.spidermiddlewares.depth.DepthMiddleware']
2024-02-28 08:48:08 [scrapy.middleware] INFO: Enabled item pipelines:
[]
2024-02-28 08:48:08 [scrapy.core.engine] INFO: Spider opened
2024-02-28 08:48:08 [scrapy.extensions.logstats] INFO: Crawled 0 pages (at 0
pages/min), scraped 0 items (at 0 items/min)
2024-02-28 08:48:08 [scrapy.extensions.telnet] INFO: Telnet console listening on
127.0.0.1:6023
2024-02-28 08:48:08 [scrapy.core.engine] DEBUG: Crawled (200) <GET
https://oralcancerfoundation.org/robots.txt> (referer: None)
2024-02-28 08:48:10 [scrapy.downloadermiddlewares.redirect] DEBUG: Redirecting
(301) to <GET https://oralcancerfoundation.org/dental/oral-cancer-images/> from
<GET https://oralcancerfoundation.org/dental/oral-cancer-images>
2024-02-28 08:48:10 [scrapy.core.engine] DEBUG: Crawled (200) <GET
https://oralcancerfoundation.org/dental/oral-cancer-images/> (referer: None)
2024-02-28 08:48:10 [scrapy.core.scraper] DEBUG: Scraped from <200
https://oralcancerfoundation.org/dental/oral-cancer-images/>
{'image_url': 'https://oralcancerfoundation.org/wp-content/gallery/oral-cancer-
images/slide1.jpg', 'description': 'An elderly woman with advanced sun-induced
degenerative alterations of face and lips. Presented with midline ulcerated
mass.', 'diagnosis': 'Diagnosis: Squamous cell carcinoma, arising actinic
cheilitis'}
2024-02-28 08:48:10 [scrapy.core.scraper] DEBUG: Scraped from <200
https://oralcancerfoundation.org/dental/oral-cancer-images/>
{'image_url': 'https://oralcancerfoundation.org/wp-content/gallery/oral-cancer-
images/slide2.jpg', 'description': '64-year-old male with a long history of
cigarette smoking presents with an asymmetrical nodular mass in the retromolar
region.', 'diagnosis': 'Diagnosis: Squamous cell carcinoma, nodular/exothytic'}
2024-02-28 08:48:10 [scrapy.core.scraper] DEBUG: Scraped from <200
https://oralcancerfoundation.org/dental/oral-cancer-images/>
{'image_url': 'https://oralcancerfoundation.org/wp-content/gallery/oral-cancer-
images/slide3.jpg', 'description': 'A 47-year-old male presents with a tender,
well-defined ulceration of the left ventral tongue of 2 weeks duration.',
'diagnosis': 'Diagnosis: Erosive lichen planus'}
```

```
2024-02-28 08:48:10 [scrapy.core.scraper] DEBUG: Scraped from <200
https://oralcancerfoundation.org/dental/oral-cancer-images/>
{'image_url': 'https://oralcancerfoundation.org/wp-content/gallery/oral-cancer-
images/slide4.jpg', 'description': 'A 59-year-old female with a painless
papillary mass of the left posterior mandibular alveolar ridge.', 'diagnosis':
'Diagnosis: Verrucous carcinoma'}
2024-02-28 08:48:10 [scrapy.core.scraper] DEBUG: Scraped from <200
https://oralcancerfoundation.org/dental/oral-cancer-images/>
{'image_url': 'https://oralcancerfoundation.org/wp-content/gallery/oral-cancer-
images/slide5.jpg', 'description': 'A 38-year-old a male presented with a white
friable lesion of the maxillary gingiva which wiped off with a cotton swab,
leaving a raw red base.', 'diagnosis': 'Diagnosis: Aspirin Burn'}
2024-02-28 08:48:10 [scrapy.core.scraper] DEBUG: Scraped from <200
https://oralcancerfoundation.org/dental/oral-cancer-images/>
{'image_url': 'https://oralcancerfoundation.org/wp-content/gallery/oral-cancer-
images/slide6.jpg', 'description': 'A 31 year old female noticed a flat
grey/black asymptomatic alteration in the anterior floor of her mouth, of
unknown duration.', 'diagnosis': 'Diagnosis: Amalgam tattoo'}
2024-02-28 08:48:10 [scrapy.core.scraper] DEBUG: Scraped from <200
https://oralcancerfoundation.org/dental/oral-cancer-images/>
{'image_url': 'https://oralcancerfoundation.org/wp-content/gallery/oral-cancer-
images/slide7.jpg', 'description': 'The patient is a 64-year-old male with
granular nodular partially ulcerated mass of the anterior mandibular gingiva.',
'diagnosis': 'Diagnosis: Squamous cell carcinoma'}
2024-02-28 08:48:10 [scrapy.core.scraper] DEBUG: Scraped from <200
https://oralcancerfoundation.org/dental/oral-cancer-images/>
{'image_url': 'https://oralcancerfoundation.org/wp-content/gallery/oral-cancer-
images/slide8.jpg', 'description': 'A bilaterally symmetrical ulcerative process
noted by this 41-year-old female, present for several months. The patient is a
non-smoker and uses no alcohol.', 'diagnosis': 'Diagnosis: Squamous cell
carcinoma'}
2024-02-28 08:48:10 [scrapy.core.scraper] DEBUG: Scraped from <200
https://oralcancerfoundation.org/dental/oral-cancer-images/>
{'image_url': 'https://oralcancerfoundation.org/wp-content/gallery/oral-cancer-
images/slide9.jpg', 'description': 'Red and white surface alteration which is
centrally indurated/firm was noted on routine examination in this 70-year-old
male. The patient has a long time history of tobacco and alcohol abuse.',
'diagnosis': 'Diagnosis: Squamous cell carcinoma'}
2024-02-28 08:48:10 [scrapy.core.scraper] DEBUG: Scraped from <200
https://oralcancerfoundation.org/dental/oral-cancer-images/>
{'image_url': 'https://oralcancerfoundation.org/wp-content/gallery/oral-cancer-
images/slide10.jpg', 'description': 'On a routine examination of a 52-year-old
female a well defined red velvety lesion of the lower-left ventral tongue was
observed.', 'diagnosis': 'Clinical Diagnosis: Erythroplakia Microscopic |
Diagnosis: Squamous cell carcinoma'}
2024-02-28 08:48:10 [scrapy.core.scraper] DEBUG: Scraped from <200
https://oralcancerfoundation.org/dental/oral-cancer-images/>
{'image_url': 'https://oralcancerfoundation.org/wp-content/gallery/oral-cancer-
images/slide11.jpg', 'description': 'Slightly elevated crater form and firm
asymmetrical lesion were found on routine oral examination of the left lateral
soft palette.', 'diagnosis': 'Diagnosis: Squamous cell carcinoma'}
2024-02-28 08:48:10 [scrapy.core.scraper] DEBUG: Scraped from <200
https://oralcancerfoundation.org/dental/oral-cancer-images/>
{'image_url': 'https://oralcancerfoundation.org/wp-content/gallery/oral-cancer-
images/slide12.jpg', 'description': 'A 69-year female presented with a sharply
defined palatal ulceration.', 'diagnosis': 'Diagnosis: Squamous cell carcinoma'}
2024-02-28 08:48:10 [scrapy.core.scraper] DEBUG: Scraped from <200
https://oralcancerfoundation.org/dental/oral-cancer-images/>
{'image_url': 'https://oralcancerfoundation.org/wp-content/gallery/oral-cancer-
images/slide13.jpg', 'description': 'A 48-year-old female presents with a hard
gingival swelling that on X-ray consisted of dense bone which was contiguous
with the cortex.', 'diagnosis': 'Diagnosis: Osteoma'}
2024-02-28 08:48:10 [scrapy.core.scraper] DEBUG: Scraped from <200
https://oralcancerfoundation.org/dental/oral-cancer-images/>
{'image_url': 'https://oralcancerfoundation.org/wp-content/gallery/oral-cancer-
images/slide14.jpg', 'description': 'An 8-year-old child evaluated for the
superficial, ulcerated, irregular lesion.', 'diagnosis': 'Diagnosis: Tramatic
ulcer (facticious)'}
2024-02-28 08:48:10 [scrapy.core.scraper] DEBUG: Scraped from <200
https://oralcancerfoundation.org/dental/oral-cancer-images/>
{'image_url': 'https://oralcancerfoundation.org/wp-content/gallery/oral-cancer-
images/slide15.jpg', 'description': 'The patient has a white lesion with
irregular margins on the left ventral tongue. At the inferior aspect, there is a
prominent red patch of tissue.', 'diagnosis': 'Diagnosis: Carcinoma in situ'}
2024-02-28 08:48:10 [scrapy.core.scraper] DEBUG: Scraped from <200
https://oralcancerfoundation.org/dental/oral-cancer-images/>
{'image_url': 'https://oralcancerfoundation.org/wp-content/gallery/oral-cancer-
images/slide16.jpg', 'description': 'Blood-based, firm, asymptomatic nodule of
long duration was noted along the right buccal mucosa during routine
examination.', 'diagnosis': 'Diagnosis: Irritational fibroma'}
2024-02-28 08:48:10 [scrapy.core.scraper] DEBUG: Scraped from <200
https://oralcancerfoundation.org/dental/oral-cancer-images/>
{'image_url': 'https://oralcancerfoundation.org/wp-content/gallery/oral-cancer-
images/slide18.jpg', 'description': 'This 46-year-old female presents with
irregular ulceration of the ventral lateral tongue, which is surrounded by
leukoplakia.', 'diagnosis': 'Diagnosis Edge biopsy revealed squamous cell
carcinoma'}
2024-02-28 08:48:10 [scrapy.core.scraper] DEBUG: Scraped from <200
https://oralcancerfoundation.org/dental/oral-cancer-images/>
{'image_url': 'https://oralcancerfoundation.org/wp-content/gallery/oral-cancer-
images/slide19.jpg', 'description': 'A flat, painless lesion of the mid-third of
the tongue, showed sharply defined borders. This lesion was noted subsequent to
a long course of antibiotic therapy.', 'diagnosis': 'Diagnosis: Median rhomboid
glossitis'}
2024-02-28 08:48:10 [scrapy.core.scraper] DEBUG: Scraped from <200
https://oralcancerfoundation.org/dental/oral-cancer-images/>
```

```
{'image_url': 'https://oralcancerfoundation.org/wp-content/gallery/oral-cancer-
images/slide20.jpg', 'description': 'A 55-year-old male noted loosening of the
lower incisors. On exam, a granular to warty mass was noted on the gingiva
extending into the labial fold.', 'diagnosis': 'Diagnosis: Well-differentiated
squamous cell carcinoma'}
2024-02-28 08:48:10 [scrapy.core.scraper] DEBUG: Scraped from <200
https://oralcancerfoundation.org/dental/oral-cancer-images/>
{'image_url': 'https://oralcancerfoundation.org/wp-content/gallery/oral-cancer-
images/slide21.jpg', 'description': 'An elongated, heterogeneous, red, well-
defined ulceration was observed on the left lateral aspect of the tongue of a
48-year-old male.', 'diagnosis': 'Diagnosis: Biopsy at several locations
revealed dysplasia and carcinoma in situ.'}
2024-02-28 08:48:10 [scrapy.core.scraper] DEBUG: Scraped from <200
https://oralcancerfoundation.org/dental/oral-cancer-images/>
{'image_url': 'https://oralcancerfoundation.org/wp-content/gallery/oral-cancer-
images/slide22.jpg', 'description': 'In this patient, the anterior floor of the
mouth, bilateral to the midline, demonstrated a vaguely marginated lesion, which
was firm on palpation.', 'diagnosis': 'Diagnosis: Marginally invasive squamous
cell carcinoma'}
2024-02-28 08:48:10 [scrapy.core.scraper] DEBUG: Scraped from <200
https://oralcancerfoundation.org/dental/oral-cancer-images/>
{'image_url': 'https://oralcancerfoundation.org/wp-content/gallery/oral-cancer-
images/slide23.jpg', 'description': 'In this patient, the buccal mucosa
bilaterally showed red and white surface changes, with delicate keratotic striae
enclosing a thin, but intact red area.', 'diagnosis': 'Diagnosis: Lichen
Planus'}
2024-02-28 08:48:10 [scrapy.core.scraper] DEBUG: Scraped from <200
https://oralcancerfoundation.org/dental/oral-cancer-images/>
{'image_url': 'https://oralcancerfoundation.org/wp-content/gallery/oral-cancer-
images/slide24.jpg', 'description': 'This ulceration of the buccal mucosa was
noticed during a routine oral examination after a dental extraction.',
'diagnosis': 'Diagnosis: Traumatic ulcer'}
2024-02-28 08:48:10 [scrapy.core.scraper] DEBUG: Scraped from <200
https://oralcancerfoundation.org/dental/oral-cancer-images/>
{'image_url': 'https://oralcancerfoundation.org/wp-content/gallery/oral-cancer-
images/slide25.jpg', 'description': 'The right lateral tongue of this patient
demonstrated the presence of an indurated, painless ulcer of unknown duration.',
'diagnosis': 'Diagnosis: Early-stage squamous cell carcinoma'}
2024-02-28 08:48:10 [scrapy.core.scraper] DEBUG: Scraped from <200
https://oralcancerfoundation.org/dental/oral-cancer-images/>
{'image_url': 'https://oralcancerfoundation.org/wp-content/gallery/oral-cancer-
images/slide26.jpg', 'description': 'A 78-year-old female was evaluated for
multiple macular pigmented lesions of the floor of the mouth and attached
gingiva.', 'diagnosis': 'Diagnosis: Malignant melanoma'}
2024-02-28 08:48:10 [scrapy.core.scraper] DEBUG: Scraped from <200
https://oralcancerfoundation.org/dental/oral-cancer-images/>
{'image_url': 'https://oralcancerfoundation.org/wp-content/gallery/oral-cancer-
images/slide27.jpg', 'description': 'A densely papillary, white, exothitic mass
```

```
with posterior ulceration was noted on the mandibular gingiva of this 72 year
old male with a 55 pack/year history of cigarette smoking and heavy alcohol
consumption.', 'diagnosis': 'Diagnosis: Papillary squamous cell carcinoma'}
2024-02-28 08:48:10 [scrapy.core.scraper] DEBUG: Scraped from <200
https://oralcancerfoundation.org/dental/oral-cancer-images/>
{'image_url': 'https://oralcancerfoundation.org/wp-content/gallery/oral-cancer-
images/slide28.jpg', 'description': 'A 68 year old female presented with a
slightly white nodular alteration of the retromolar region.', 'diagnosis':
'Diagnosis: Squamous cell carcinoma'}
2024-02-28 08:48:10 [scrapy.core.scraper] DEBUG: Scraped from <200
https://oralcancerfoundation.org/dental/oral-cancer-images/>
{'image_url': 'https://oralcancerfoundation.org/wp-content/gallery/oral-cancer-
images/slide29.jpg', 'description': 'A bluish-red nodular mass was detected in
the maxillary right quadrant of this 54 year old male.', 'diagnosis':
'Diagnosis: Peripheral giant cell granuloma'}
2024-02-28 08:48:10 [scrapy.core.scraper] DEBUG: Scraped from <200
https://oralcancerfoundation.org/dental/oral-cancer-images/>
{'image_url': 'https://oralcancerfoundation.org/wp-content/gallery/oral-cancer-
images/slide30.jpg', 'description': 'A 78 year old woman with a long history of
tobacco and alcohol abuse, has difficulty opening her mouth. Examination
revealed a foul smelling lesion extending superiorly and to the posterior.',
'diagnosis': 'Diagnosis: Squamous cell carcinoma'}
2024-02-28 08:48:10 [scrapy.core.engine] INFO: Closing spider (finished)
2024-02-28 08:48:10 [scrapy.extensions.feedexport] INFO: Stored json feed (29
items) in: output.json
2024-02-28 08:48:10 [scrapy.statscollectors] INFO: Dumping Scrapy stats:
{'downloader/request_bytes': 736,
 'downloader/request_count': 3,
 'downloader/request_method_count/GET': 3,
 'downloader/response_bytes': 154058,
 'downloader/response_count': 3,
 'downloader/response_status_count/200': 2,
 'downloader/response_status_count/301': 1,
 'elapsed_time_seconds': 2.338591,
 'feedexport/success_count/FileFeedStorage': 1,
 'finish_reason': 'finished',
 'finish_time': datetime.datetime(2024, 2, 28, 8, 48, 10, 878409,
tzinfo=datetime.timezone.utc),
 'httpcompression/response_bytes': 1109061,
 'httpcompression/response_count': 2,
 'item_scraped_count': 29,
 'log_count/DEBUG': 35,
 'log_count/INFO': 11,
 'memusage/max': 212217856,
 'memusage/startup': 212217856,
 'response_received_count': 2,
 'robotstxt/request_count': 1,
 'robotstxt/response_count': 1,
 'robotstxt/response_status_count/200': 1,
 'scheduler/dequeued': 2,
 'scheduler/dequeued/memory': 2,
 'scheduler/enqueued': 2,
 'scheduler/enqueued/memory': 2,
 'start_time': datetime.datetime(2024, 2, 28, 8, 48, 8, 539818,
tzinfo=datetime.timezone.utc)}
2024-02-28 08:48:10 [scrapy.core.engine] INFO: Spider closed (finished)
```

2 Ứng dụng cho project của nhóm (Cancer Diagnosis)

Nhóm em sử dụng ví dụ minh họa trên để lấy ra link ảnh, mô tả thông tin bệnh nhân cũng như chuẩn đoán của bác sĩ từ bộ dữ liệu **Oral Cancer Images**.