

Programación para *Data Science*

Unidad 5: Adquisición de datos en Python - Ejercicios y preguntas

Ejercicio 1

Seleccionad y descargad cualquier fichero de datos de alguno de los portales comentados en el Notebook de esta unidad.

Cargad los datos en una variable Python (podéis usar [pandas \(http://pandas.pydata.org/\)](http://pandas.pydata.org/) si queréis). Mostrad el número de muestras que contiene el conjunto de datos y los atributos disponibles para cada muestra.

```
In [1]: # Respuesta
```

Pregunta 1

Queremos saber cuántas personas hay en el espacio en un momento dado. Identificad qué método de la [siguiente API \(http://open-notify.org/Open-Notify-API\)](http://open-notify.org/Open-Notify-API) podemos utilizar para obtener este dato y contestad a las siguientes preguntas:

- 1. ¿A qué URL realizaremos la petición?
- 2. ¿Qué tipo de petición HTTP (qué acción) tendremos que realizar a la API para obtener los datos deseados?
- 3. ¿En qué formato obtendremos la respuesta de la API seleccionada?
- 4. ¿En qué campo de la respuesta encontraremos la información que buscamos?

Respuesta

Ejercicio 2

Implementad una función que devuelva el número de personas que hay actualmente en el espacio.

```
In [2]: # Respuesta
```

Ejercicio 3

Programad una función que muestre la fecha y hora de los diez próximos pases de la estación espacial internacional (ISS) por encima de una localización concreta (especificada por su dirección postal).

Pista: ¡pensad que podéis combinar resultados de varias API!

```
In [4]: # Respuesta
```

Pregunta 2

¿Qué es la *scrapy shell*? Describid para qué sirve y poned algún ejemplo de su uso.

Respuesta

Ejercicio 4

[KDNuggets \(http://www.kdnuggets.com/\)](http://www.kdnuggets.com/) es una web con contenido sobre *Data Science*. Entre otros, la web mantiene un listado de ofertas de trabajo relacionadas con el ámbito.

Programad un *crawler* que extraiga los títulos de todas las ofertas de trabajo de KDNuggets.

Para hacerlo, utilizad la estructura de *crawler* que hemos visto en el Notebook de esta unidad, **modificando únicamente dos líneas de código**:

- La URL de inicio.
- La expresión XPath que selecciona el contenido que hay que capturar.

Nota: si la ejecución del *crawler* os devuelve un error `ReactorNotRestartable`, reiniciad el *kernel* del Notebook (en el menú, Kernel - Restart).

```

In [1]: # Importamos scrapy.
import scrapy
from scrapy.crawler import CrawlerProcess

# Creamos la araña.
class uoc_spider(scrapy.Spider):

    # Asignamos un nombre a la araña.
    name = "uoc_spider"

    # Indicamos la URL que queremos analizar en primer lugar
    # Includ aquí la URL de inicio:
    #####
    start_urls = [
        ""
    ]
    #####

    # Definimos el analizador.
    def parse(self, response):
        # Extraemos el título del grado.
        # Includ aquí la expresión 'xpath' que nos devuelve los títulos de las ofertas de trabajo.
        #####
        for t in response.xpath(''):
            #####
            yield {
                'title': t.extract()
            }

if __name__ == "__main__":

    # Creamos un crawler.
    process = CrawlerProcess({
        'USER_AGENT': 'Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.1)',
        'DOWNLOAD_HANDLERS': {'s3': None},
        'LOG_ENABLED': False
    })

    # Inicializamos el crawler con nuestra araña.
    process.crawl(uoc_spider)

    # Lanzamos la araña.
    process.start()

```

```
INFO:scrapy.utils.log:Scrapy 1.0.3 started (bot: scrapybot)
INFO:scrapy.utils.log:Optional features available: ssl, http11, boto
INFO:scrapy.utils.log:Overridden settings: {'USER_AGENT': 'Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.1)', 'LOG_ENABLED': False}
INFO:scrapy.middleware:Enabled extensions: CloseSpider, TelnetConsole, LogStats, CoreStats, SpiderState
INFO:scrapy.middleware:Enabled downloader middlewares: HttpAuthMiddleware, DownloadTimeoutMiddleware, UserAgentMiddleware, RetryMiddleware, DefaultHeadersMiddleware, MetaRefreshMiddleware, HttpCompressionMiddleware, RedirectMiddleware, CookiesMiddleware, ChunkedTransferMiddleware, DownloaderStats
INFO:scrapy.middleware:Enabled spider middlewares: HttpErrorMiddleware, OffsiteMiddleware, RefererMiddleware, UrlLengthMiddleware, DepthMiddleware
INFO:scrapy.middleware:Enabled item pipelines:
INFO:scrapy.core.engine:Spider opened
INFO:scrapy.extensions.logstats:Crawled 0 pages (at 0 pages/min), scraped 0 items (at 0 items/min)
DEBUG:scrapy.telnet:Telnet console listening on 127.0.0.1:6024
DEBUG:scrapy.core.engine:Crawled (200) <GET http://www.kdnuggets.com/jobs/index.html> (referer: None)
DEBUG:scrapy.core.scrapers:Scraped from <200 http://www.kdnuggets.com/jobs/index.html>
{'title': u'Chubb: Customer Experience & Marketing Analytics Manager'}
DEBUG:scrapy.core.scrapers:Scraped from <200 http://www.kdnuggets.com/jobs/index.html>
{'title': u'Live.xyz: Senior Data Scientist'}
DEBUG:scrapy.core.scrapers:Scraped from <200 http://www.kdnuggets.com/jobs/index.html>
{'title': u'Turner: Sr. Operations Research Analyst / Data Scientist'}
DEBUG:scrapy.core.scrapers:Scraped from <200 http://www.kdnuggets.com/jobs/index.html>
{'title': u'Turner: Advisor Analytics Architect'}
DEBUG:scrapy.core.scrapers:Scraped from <200 http://www.kdnuggets.com/jobs/index.html>
{'title': u'Apple: Software Engineer \u2013 Local Search'}
DEBUG:scrapy.core.scrapers:Scraped from <200 http://www.kdnuggets.com/jobs/index.html>
{'title': u'PeachIH: Data Scientist, Machine Learning Engineer'}
DEBUG:scrapy.core.scrapers:Scraped from <200 http://www.kdnuggets.com/jobs/index.html>
{'title': u'Eindhoven University of Technology: Full Professor Database Technology'}
DEBUG:scrapy.core.scrapers:Scraped from <200 http://www.kdnuggets.com/jobs/index.html>
{'title': u'Corios: IT Systems Administrator'}
DEBUG:scrapy.core.scrapers:Scraped from <200 http://www.kdnuggets.com/jobs/index.html>
{'title': u'Aetna: Lead Data Scientist'}
DEBUG:scrapy.core.scrapers:Scraped from <200 http://www.kdnuggets.com/jobs/index.html>
{'title': u'Accenture: Artificial Intelligence Experienced Researcher'}
DEBUG:scrapy.core.scrapers:Scraped from <200 http://www.kdnuggets.com/jobs/index.html>
{'title': u'LeapYear: Lead Data Scientist'}
DEBUG:scrapy.core.scrapers:Scraped from <200 http://www.kdnuggets.com/jobs/index.html>
{'title': u'IRIS Advanced Engineering: TecnioSpring Data Mining Fellowship'}
DEBUG:scrapy.core.scrapers:Scraped from <200 http://www.kdnuggets.com/jobs/index.html>
{'title': u'Schwab: Senior Data Scientist'}
DEBUG:scrapy.core.scrapers:Scraped from <200 http://www.kdnuggets.com/jobs/index.html>
{'title': u'Stylight: Senior Data Scientist'}
DEBUG:scrapy.core.scrapers:Scraped from <200 http://www.kdnuggets.com/jobs/index.html>
{'title': u'Freeletics: Data Scientist'}
DEBUG:scrapy.core.scrapers:Scraped from <200 http://www.kdnuggets.com/jobs/index.html>
{'title': u'Apple: Data Mining Scientist'}
DEBUG:scrapy.core.scrapers:Scraped from <200 http://www.kdnuggets.com/jobs/index.html>
{'title': u'Metis: Evangelist'}
DEBUG:scrapy.core.scrapers:Scraped from <200 http://www.kdnuggets.com/jobs/index.html>
{'title': u'Altria: Applied Statistician, Regulatory Affairs'}
DEBUG:scrapy.core.scrapers:Scraped from <200 http://www.kdnuggets.com/jobs/index.html>
{'title': u'Netflix: Manager, Content Programming Science & Algorithms'}
DEBUG:scrapy.core.scrapers:Scraped from <200 http://www.kdnuggets.com/jobs/index.html>
{'title': u'Aetna: Principal Data Scientist'}
DEBUG:scrapy.core.scrapers:Scraped from <200 http://www.kdnuggets.com/jobs/index.html>
{'title': u'ACI Worldwide: Sr. Data Scientist'}
DEBUG:scrapy.core.scrapers:Scraped from <200 http://www.kdnuggets.com/jobs/index.html>
{'title': u'R2: DevOps for Data Science'}
DEBUG:scrapy.core.scrapers:Scraped from <200 http://www.kdnuggets.com/jobs/index.html>
{'title': u'NEC Labs: Researcher'}
DEBUG:scrapy.core.scrapers:Scraped from <200 http://www.kdnuggets.com/jobs/index.html>
{'title': u'89degrees: SAS Campaign Developer'}
DEBUG:scrapy.core.scrapers:Scraped from <200 http://www.kdnuggets.com/jobs/index.html>
{'title': u'89degrees: SAS Administrator'}
DEBUG:scrapy.core.scrapers:Scraped from <200 http://www.kdnuggets.com/jobs/index.html>
{'title': u'89degrees: Enterprise Data Architect'}
DEBUG:scrapy.core.scrapers:Scraped from <200 http://www.kdnuggets.com/jobs/index.html>
{'title': u'Scripps: Director, Data Science'}
DEBUG:scrapy.core.scrapers:Scraped from <200 http://www.kdnuggets.com/jobs/index.html>
{'title': u'Scripps: Data Scientist'}
DEBUG:scrapy.core.scrapers:Scraped from <200 http://www.kdnuggets.com/jobs/index.html>
{'title': u'Dstillery: Senior Data Scientist'}
DEBUG:scrapy.core.scrapers:Scraped from <200 http://www.kdnuggets.com/jobs/index.html>
{'title': u'Senior Data Scientist (Inv / Risk / Marketing) NY & BOS | Recent PhD'}
DEBUG:scrapy.core.scrapers:Scraped from <200 http://www.kdnuggets.com/jobs/index.html>
{'title': u'FeatureX: Software Engineer'}
DEBUG:scrapy.core.scrapers:Scraped from <200 http://www.kdnuggets.com/jobs/index.html>
{'title': u'FeatureX: Machine Learning Research Scientist'}
DEBUG:scrapy.core.scrapers:Scraped from <200 http://www.kdnuggets.com/jobs/index.html>
{'title': u'FeatureX: Computer Vision Research Scientist'}
DEBUG:scrapy.core.scrapers:Scraped from <200 http://www.kdnuggets.com/jobs/index.html>
{'title': u'Aetna: Lead Data Scientist'}
DEBUG:scrapy.core.scrapers:Scraped from <200 http://www.kdnuggets.com/jobs/index.html>
{'title': u'U. of Calgary: Professor, Tier 1 Canada Research Chair, Biomedical Engineering'}
DEBUG:scrapy.core.scrapers:Scraped from <200 http://www.kdnuggets.com/jobs/index.html>
{'title': u'Aetna: Senior Data Scientist'}
DEBUG:scrapy.core.scrapers:Scraped from <200 http://www.kdnuggets.com/jobs/index.html>
{'title': u'Aetna: Principal Data Scientist'}
DEBUG:scrapy.core.scrapers:Scraped from <200 http://www.kdnuggets.com/jobs/index.html>
{'title': u'Stylight: Senior Data Scientist'}
DEBUG:scrapy.core.scrapers:Scraped from <200 http://www.kdnuggets.com/jobs/index.html>
{'title': u'Ferris State University: Assistant Professor Business Data Analytics'}
DEBUG:scrapy.core.scrapers:Scraped from <200 http://www.kdnuggets.com/jobs/index.html>
{'title': u'Accenture: Data Scientist'}
```

```
INFO:scrapy.core.engine:Closing spider (finished)
INFO:scrapy.statscollectors:Dumping Scrapy stats:
{'downloader/request_bytes': 247,
 'downloader/request_count': 1,
 'downloader/request_method_count/GET': 1,
 'downloader/response_bytes': 10204,
 'downloader/response_count': 1,
 'downloader/response_status_count/200': 1,
 'finish_reason': 'finished',
 'finish_time': datetime.datetime(2017, 3, 29, 9, 24, 51, 977869),
 'item_scraped_count': 40,
 'log_count/DEBUG': 42,
 'log_count/INFO': 7,
 'response_received_count': 1,
 'scheduler/dequeued': 1,
 'scheduler/dequeued/memory': 1,
 'scheduler/enqueued': 1,
 'scheduler/enqueued/memory': 1,
 'start_time': datetime.datetime(2017, 3, 29, 9, 24, 51, 448811)}
INFO:scrapy.core.engine:Spider closed (finished)
```

Pregunta 3

¿Qué es OAuth? Explicad una situación donde sería útil usar OAuth en vez de un protocolo de autenticación tradicional basado en usuario y contraseña.

Respuesta