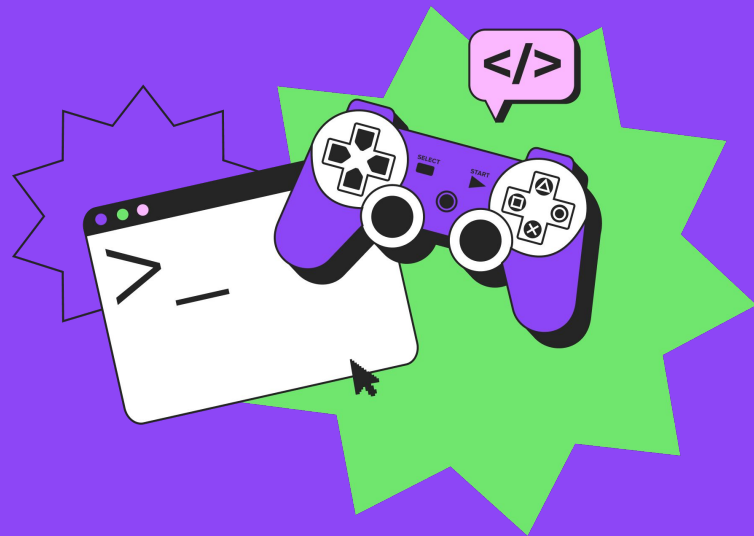


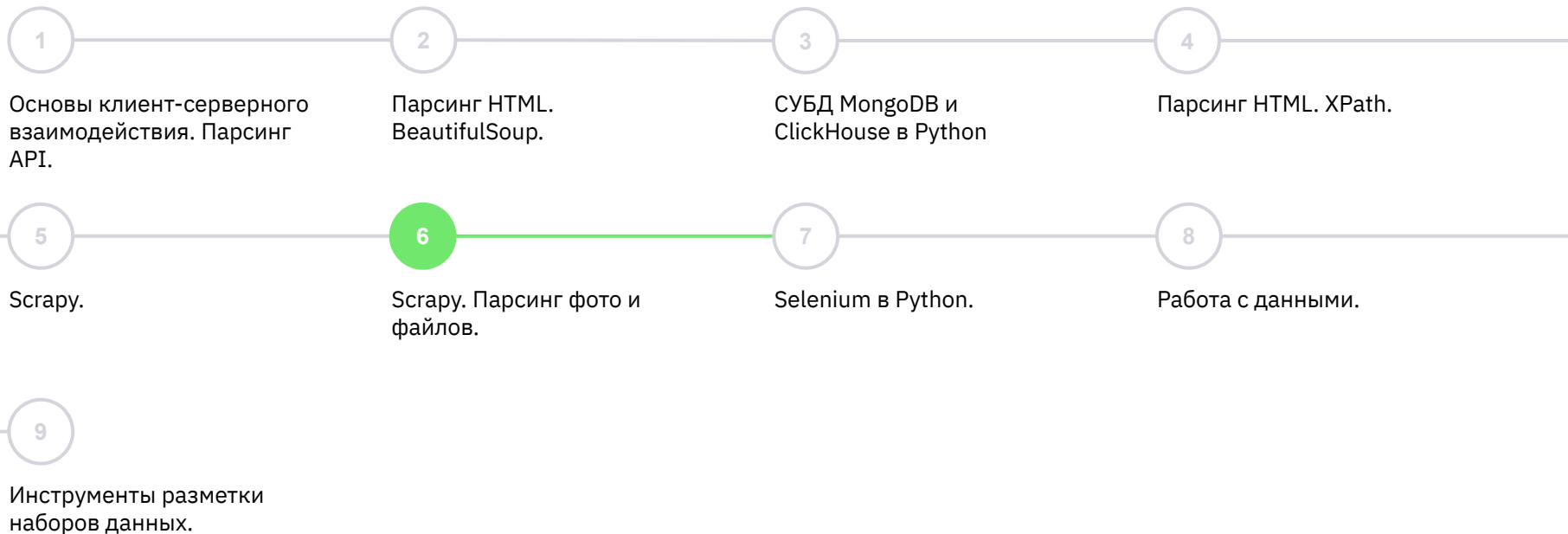
Scrapy. Парсинг фото и файлов.

Семинар 6
Сбор и разметка данных








Сбор и разметка данных



Что будет на уроке сегодня

-  Создание скриптов Python для извлечения и загрузки изображений с помощью пауков Scrapy.
-  Использование Scrapy Items и ItemLoaders для управления сбором и обработкой данных.
-  Настройка и реализация пайплайнов в Scrapy для различных задач обработки данных.





Викторина



Что из перечисленного ниже не является компонентом Scrapy?

1. Spiders
2. Items
3. Pipelines
4. Controllers



Что из перечисленного ниже не является компонентом Scrapy?

1. Spiders
2. Items
3. Pipelines
4. Controllers



Каково назначение ключевого слова `yield` в пауке Scrapy?

1. Остановить выполнение паука.
2. Вернуть результат и приостановить выполнение паука до его повторного вызова.
3. Вывести результат на консоль.
4. Вернуть результат и завершить выполнение паука.



Каково назначение ключевого слова `yield` в пауке Scrapy?

1. Остановить выполнение паука.
2. Вернуть результат и приостановить выполнение паука до его повторного вызова.
3. Вывести результат на консоль.
4. Вернуть результат и завершить выполнение паука.



Каково назначение параметра `ITEM_PIPELINES` в Scrapy?

1. Чтобы указать порядок применения пайплайнов.
2. Чтобы определить начальный URL-адрес паука.
3. Чтобы задать пользовательский агент паука.
4. Чтобы указать место загрузки файлов.



Каково назначение параметра `ITEM_PIPELINES` в Scrapy?

1. Чтобы указать порядок применения пайплайнов.
2. Чтобы определить начальный URL-адрес паука.
3. Чтобы задать пользовательский агент паука.
4. Чтобы указать место загрузки файлов.



Каково назначение параметра `ITEM_PIPELINES` в Scrapy?

1. Чтобы указать порядок применения пайплайнов.
2. Чтобы определить начальный URL-адрес паука.
3. Чтобы задать пользовательский агент паука.
4. Чтобы указать место загрузки файлов.



Каково назначение Item в Scrapy?

1. Для определения структуры данных, которые будет извлекать паук.
2. Для навигации по HTML-структуре веб-страницы.
3. Чтобы указать место загрузки файлов.
4. Для определения начального URL-адреса паука.



Каково назначение Item в Scrapy?

1. Для определения структуры данных, которые будет извлекать паук.
2. Для навигации по HTML-структуре веб-страницы.
3. Чтобы указать место загрузки файлов.
4. Для определения начального URL-адреса паука.



Что делает метод `response.follow()` в Scrapy?

1. Посылает новый HTTP-запрос и вызывает метод `callback` с ответом.
2. Останавливает выполнение паука.
3. Возвращает ответ на предыдущий HTTP-запрос.
4. Печатает ответ в консоль.



Что делает метод `response.follow()` в Scrapy?

1. **Посылает новый HTTP-запрос и вызывает метод callback с ответом.**
2. Останавливает выполнение паука.
3. Возвращает ответ на предыдущий HTTP-запрос.
4. Печатает ответ в консоль.



Как Scrapy может обрабатывать загрузку файлов?

1. С помощью пользовательского пайплайна, который использует ImagesPipeline или FilesPipeline.
2. С помощью встроенного метода download() в пауке.
3. С помощью ключевого слова yield в пауке.
4. Scrapy не может обрабатывать загрузку файлов.



Как Scrapy может обрабатывать загрузку файлов?

1. С помощью пользовательского пайплайна, который использует ImagesPipeline или FilesPipeline.
2. С помощью встроенного метода download() в пауке.
3. С помощью ключевого слова yield в пауке.
4. Scrapy не может обрабатывать загрузку файлов.



Вопросы?

Вопросы?



Вопросы?





Практика



Знакомство с целевым веб-сайтом

https://commons.wikimedia.org/wiki/Category:Featured_pictures_on_Wikimedia_Commons

Паук должен делать следующее:

- Начать с URL-адреса категории "Featured pictures on Wikimedia Commons".
- Для каждого изображения на странице перейдите по ссылке на страницу изображения.
- На странице каждого изображения извлеките следующие данные: Description in English, Date, Source, Author и URL-адрес самого изображения.
- Сохраните извлеченные данные в файле JSON, а также загрузите изображение.



Задание 1

- Создайте новый проект Scrapy с именем "wikimedia_scraper".
- В каталоге spiders создайте нового паука с именем "wikimedia".
- Определите start_urls равным https://commons.wikimedia.org/wiki/Category:Featured_pictures_on_Wikimedia_Commons
- Создайте метод parse в пауке, который будет извлекать URL отдельных страниц изображений со страницы категории.
- Создайте метод parse_image, который будет использоваться для парсинга отдельных страниц изображений.

Подсказки:

- Используйте выражение xpath в методе парсинга для извлечения URL-адресов. Например:
`response.xpath("//li[@class='gallerybox']/div/div/div/a/@href").extract()`
- Используйте метод `response.follow()` для перехода по каждому извлеченному URL к соответствующей странице изображения и передайте его в метод `parse_image`.



20 минут



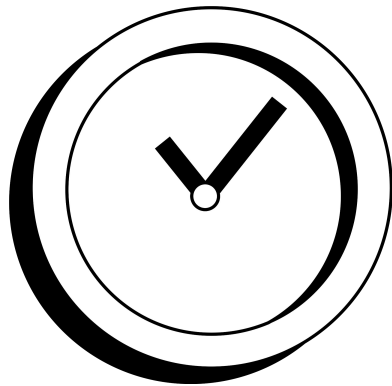
Задание 1

- Создайте новый проект Scrapy с именем "wikimedia_scraper".
- В каталоге spiders создайте нового паука с именем "wikimedia".
- Определите start_urls равным https://commons.wikimedia.org/wiki/Category:Featured_pictures_on_Wikimedia_Commons
- Создайте метод parse в пауке, который будет извлекать URL отдельных страниц изображений со страницы категории.
- Создайте метод parse_image, который будет использоваться для парсинга отдельных страниц изображений.

Подсказки:

- Используйте выражение xpath в методе парсинга для извлечения URL-адресов. Например:
`response.xpath("//li[@class='gallerybox']/div/div/div/a/@href").extract()`
- Используйте метод `response.follow()` для перехода по каждому извлеченному URL к соответствующей странице изображения и передайте его в метод `parse_image`.

<<20:00-





Задание 2

Откройте файл `settings.py` в корневом каталоге вашего проекта Scrapy.

Добавьте или измените необходимые настройки для вашего проекта. К ним относятся `BOT_NAME`, `USER_AGENT`, `ROBOTSTXT_OBEY`, `ITEM_PIPELINES` и `IMAGES_STORE`, `DOWNLOAD_DELAY`, `DOWNLOAD_DELAY_FACTOR`, `LOG_LEVEL`.



20 минут



Задание 2 - Hints

Подсказки:

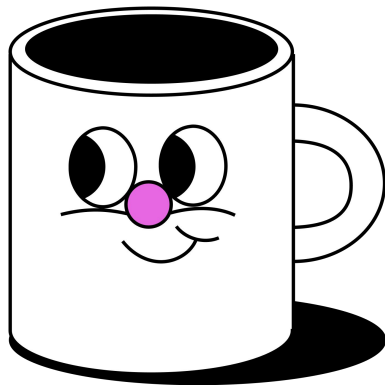
- BOT_NAME обычно является именем вашего проекта Scrapy.
- Установите ROBOTSTXT_OBEY в True, если вы хотите, чтобы ваш паук соблюдал политику robots.txt сайтов, которые он сканирует.
- ITEM_PIPELINES - это словарь, в котором ключами являются строки, представляющие пути к классам пайплайнов, а значениями - целые числа, представляющие порядок применения пайплайнов. Например:

```
ITEM_PIPELINES = {  
    'myproject.pipelines.MyImagesPipeline': 1,  
    'myproject.pipelines.JsonWriterPipeline': 2,  
}
```
- IMAGES_STORE - это строка, представляющая путь к директории, в которую вы хотите загрузить изображения.
Например: IMAGES_STORE = 'downloaded_images'.



20 минут

Перерыв



<<5:00->>



Задание 3

- В методе `parse_image` используйте XPath для извлечения необходимой информации: Description, Date, Source, Author, and Link.
- Также извлеките URL-адрес фактического изображения на странице.

Подсказки:

- Используйте `ItemLoader`, чтобы управлять извлечением данных. Вы можете создать его экземпляр в методе `parse_image` следующим образом: `l = ItemLoader(item=WikimediaItem(), response=response)`.
- Используйте методы `add_xpath()` и `add_value()` `ItemLoader` для добавления значений в поля `WikimediaItem`.
- Используйте `MapCompose(lambda i: urljoin(response.url, i))`, чтобы соединить базовый URL с относительным URL изображения.



40 минут



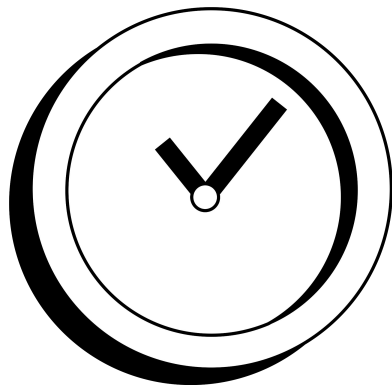
Задание 3

- В методе `parse_image` используйте XPath для извлечения необходимой информации: Description, Date, Source, Author, and Link.
- Также извлеките URL-адрес фактического изображения на странице.

Подсказки:

- Используйте `ItemLoader`, чтобы управлять извлечением данных. Вы можете создать его экземпляр в методе `parse_image` следующим образом: `l = ItemLoader(item=WikimediaItem(), response=response)`.
- Используйте методы `add_xpath()` и `add_value()` `ItemLoader` для добавления значений в поля `WikimediaItem`.
- Используйте `MapCompose(lambda i: urljoin(response.url, i))`, чтобы соединить базовый URL с относительным URL изображения.

<<40:00-





Задание 4

- Определите два класса пайплайнов в файле pipelines.py: один для загрузки изображений, а другой для записи данных в JSON-файл.

- В пайплайнах реализуйте методы, требуемые Scrapy: `get_media_requests()` и `item_completed()` для пайплайна загрузки изображений и `open_spider()`, `close_spider()` и `process_item()` для пайплайна записи JSON.



30 минут



Задание 4 - Hints

- Для пайплайна загрузки изображений используйте ImagesPipeline, предоставляемый Scrapy.
- В методе `get_media_requests()` создайте запрос для каждого URL изображения в `item['image_urls']`.
- В методе `item_completed()` обновите `item['image_urls']`, чтобы он содержал пути загруженных изображений.
- Для пайплайна записи JSON откройте файл в `open_spider()` и запишите в него каждый элемент в `process_item()`. Закройте файл - `close_spider()`.



30 минут



Домашнее задание

1. Создайте новый проект Scrapy. Дайте ему подходящее имя и убедитесь, что ваше окружение правильно настроено для работы с проектом.
2. Создайте нового паука, способного перемещаться по сайту **www.unsplash.com**. Ваш паук должен уметь перемещаться по категориям фотографий и получать доступ к страницам отдельных фотографий.
3. Определите элемент (Item) в Scrapy, который будет представлять изображение. Ваш элемент должен включать такие детали, как URL изображения, название изображения и категорию, к которой оно принадлежит.
4. Используйте Scrapy ImagesPipeline для загрузки изображений. Обязательно установите параметр IMAGES_STORE в файле settings.py. Убедитесь, что ваш паук правильно выдает элементы изображений, которые может обработать ImagesPipeline.
5. Сохраните дополнительные сведения об изображениях (название, категория) в CSV-файле. Каждая строка должна соответствовать одному изображению и содержать URL изображения, локальный путь к файлу (после загрузки), название и категорию.



Вопросы?

Вопросы?



Вопросы?





Спасибо за внимание!