Домашнее задание
по теме «Работа с Web-scraping»

Формулировка задания

Сформировать запросы для поиска изображений и выполнить из загрузку с веб-ресурса. Для выполнения работы использовать библиотеки requests, bs4, selenium и другие. Допустимо реализовать асинхронную загрузку.

Результирующий код должен быть читаемым, с единой системой отступов и адекватными названиями переменных.

Описание плана работы

Сформировать запросы для поиска изображений и выполнить из загрузку с веб-ресурса. Для выполнения работы использовать библиотеки requests, bs4, selenium и другие. Допустимо реализовать асинхронную загрузку.
Обратить внимание на работу с captcha сайта Yandex.

Желательна реализация в файлах py. Сохранить задачи (сделать коммиты для каждой) в локальном git и опубликовать в удаленном репозитории.

Для отчета по работе выполнить задание в файле.py или .ipynb. Сделать снимки экрана корректного выполнения программы в IDE.

Задача Тестирование методов класса

С использованием страницы https://yandex.ru/images/ сформировать запросы для поиска изображений, контент на которых соответствует классам **polar bear** и **brown bear**. Для каждого класса должно быть загружено не менее 1000 изображений. Изображения для каждого класса должны находиться в подпапке папки dataset с соответствующим названием.

Не допускается:

- Создание папок вручную. В коде должен быть отражен процесс создания папок и перемещения/загрузки в них файлов.
- Дублирование изображений для класса.

Примечания

- Каждое изображение должно иметь расширение .jpg
- Именовать файлы необходимо порядковым номером (от 0 до 999).

- Для дальнейшего удобства необходимо дополнять имя файла ведущими нулями (например, 0000, 0001, ..., 0999). Для этого необходимо использовать один из методов класса str.
- После загрузки всех изображений, необходимо их просмотреть на соответствие классу. В случае замеченных несоответствий необходимо будет дополнить набор данных до минимального размера. Для избежания подобных ситуаций рекомендуется загружать изображения с запасом.

Выполнение задания подразумевает два уровня сложности:

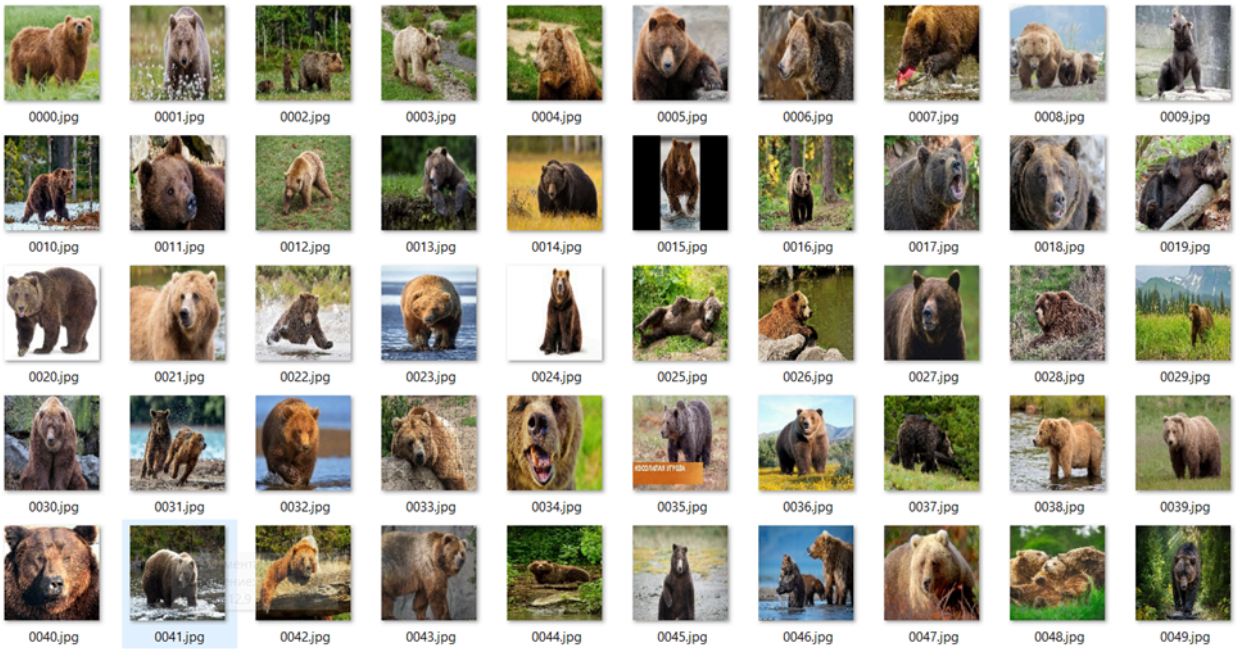
- Для первого уровня сложности достаточно загрузить лишь миниатюры изображений.
- Для второго уровня сложности необходимо загрузить полноразмерные изображения.

Для того, чтобы задание было принято, достаточно выполнить первый уровень сложности.

Обратить внимание на работу с captcha сайта Yandex. Рекомендуется использовать для обхода selenium.webdriver и аналоги. Использование прокси позволяет эффективно работать с captcha.

Выходные данные для задачи

Для класса **brown bear** первые 50 изображений. Загрузили изображения с Яндекс Картинок по запросу **brown bear** и удалось загрузить 1012 изображений.



Для класса **polar bear** первые 50 изображений. Загрузили изображения с Яндекс Картинок по запросу **polar bear** и удалось загрузить 1020 изображений.

Перечень необходимых инструментов

- Python
- requests
- bs4 (BeautifulSoup 4)
- selenium
- venv
- Jupiter Notebook
- IDE VS Code
- GigaIDE

Форма предоставления результата

1. В поле ссылки загрузить ссылку на удаленный репозиторий с доступом для наставника.
2. В поле файла загрузить архив с папкой, в которой разместить отчет со скриншотами по заданию и решение задачи. Решение должно быть представлено в формате .ipynb или .py.

Шкала оценивания

- 1.0 – отлично
- 0.7–0.9 – хорошо
- 0.5–0.6 – удовлетворительно
- Менее 0.5 – задание не выполнено