Internet Information collection and data processing methods

Lesson 1

Урок 1. Введение в методы сбора и обработки данных

Задание 1. Так как мы решили делать общий проект по тематике ICO  
Ваше домашнее задание немного расширяется.   
Вам необходимо найти не менее 2х источников информации о существующих/существовавших ICO проектов. ( в материалы к уроку добавил несколько ссылок с которых можно начать, для понимания что такое ico)

клоним себе проект <https://github.com/GefMar/GB_data_mining>  
Каждый ученик создает себе свою ветку в которой он будет работать, для каждого урока новая ветка которую мы после будем сливать. (для удобства именуйте ветку по шаблону номерурока\_фамилия

Задание 2. Здесь у вас простое задание - дать нам обратную связь, чтобы мы становились лучше для вас.  
Пройдите опрос по ссылке: <https://forms.gle/XpHNLmKhznpSoBMDA>

Мы все читаем и стараемся учитывать пожелания в следующих уроках!

Lesson 2

### Урок 2. HTTP/1.1, Proxy и VPN

задание 1: Ресурс <https://icobench.com/icos>  
Необходимо выгрузить в отдельные файлы html код каждой страницы каталога ICO проектов. Шаблон имени файла icobench\_ico\_page\_{номер страницы}.html  
Задание 2: Ресурс <https://5ka.ru/special_offers/>   
Необхоидмо выгрузить все товары по акции - каждый товар в отдельный файл (csv или json), имя файла id товара, выгрузить необходимо абсолютно все товары по акции которые есть в источнике.

Задание 3 - для необязательно: с помощью поисковика <https://www.shodan.io/> необходимо найти elastic search или mongoDB или JsonApi с информацией о ICO проектах.

все домашние работы желательно сдавать через GITHUB, но можно архивом с кодом.

-. Здесь у вас простое задание - дать нам обратную связь, чтобы мы становились лучше для вас.  
Пройдите опрос по ссылке: <https://forms.gle/FMFrqg27vAFtkJjAA>

Мы все читаем и стараемся учитывать пожелания в следующих уроках!

Lesson 3

### Урок 3. СУБД MongoDB

Задача:   
Спарсить АПИ <https://icorating.com/ico/all/load/?page=1> которое есть тут.   
Сохранить данные в MongoDB и в SQlite базу (с использованием sqlalchemy)  
КОД урока уже на Гитхаб (ветка teacher\_02)

задание более высокой сложности.   
С сайта пятерочки спарсить все товары как и ранее, но сохранить все в MongoDB и SQLite (с использованием sqlalchemy)  
Усложняется задача тем что: все вложенные обьекты должны быть отдельными объектами в БД и должны сохраниться связи.

1. Здесь у вас простое задание - дать нам обратную связь, чтобы мы становились лучше для вас. Пройдите опрос по ссылке: <https://forms.gle/FMFrqg27vAFtkJjAA>

Мы все читаем и стараемся учитывать пожелания в следующих уроках!

Lesson 4

### Урок 4. REST

Итак, домашнее задание:   
С помощью бибилотеки scrapy обойти все ICO проекты на рессурсе <https://icobench.com/icos>?  
Нам необходимо собрать следующие данные:   
Название,  
Слоган  
Описание  
Теги(категории)  
ссылки на официальные источники (Оф сайт проекта, соц сети проекта)  
Рейтинги.  
содержимое секции About  
Команду и Адвизоров проекта

Необходимо собирать персон таким образом что-бы Человек в базе был отдельной записью, а проект в соответсвющих полях только ссылался на этого человека (наладить связь), учитывайте что должность человека на разных проектах может быть различной.   
У человека необходимо знать Имя, и ссылки на соц сети если таковые есть.

Необходимо что-бы вы использовали items и pipelines по назначению.

Дополнительное задание (необзязательно)  
У проектов в секции Whitepaper присутсвует файл, при помощи pipeline произвести скачивание файла и добавить путь к файлу в БД.

Lesson 5

### Урок 5. SOAP

Мы продолжаем собирать публичные данные о ICO проектах.

## **1**

На данном этапе нам необходимо расширить существующую Базу данных собираемую из источника <https://icobench.com/icos>?  
Рейтинг - необходимо поле в котором будет словарь с каждым из рейтингов (profile, team, vision, product)  
Поле с датами проведения preICO (PREICO TIME)  
Поле с датами проведения ICO(ICO TIME)  
Все поля из данной секции (<https://www.dropbox.com/s/u102astb33xbs5s/%D0%A1%D0%BA%D1%80%D0%B8%D0%BD%D1%88%D0%BE%D1%82%202019-07-26%2000.25.44.png?dl=0>)

## **2**

Пишем паука для <https://icorating.com/ico/all/>

Нам необходимо собрать следующие поля -   
Название проекта  
Описание проекта  
Сслыки на соц сети и прочие рессурсы проекта  
поля из секции (<https://www.dropbox.com/s/e9wkdb5ooup5uyu/%D0%A1%D0%BA%D1%80%D0%B8%D0%BD%D1%88%D0%BE%D1%82%202019-07-26%2000.31.41.png?dl=0>)  
Команду проекта (имена ссылки на сети и позицию)  
Даты проведения Ico

## **3**

Дополнительно рессурс [https://deadcoins.com](https://deadcoins.com/)  
Собираем с него всю информацию о проекте которая там есть (всего 4-5 полей), но данный сайт с точки зрения парсинга иногда делает сюрпризы. Это небольшой тест на внимательность. Распарсить его не сложно совсем.

# **Все задания выполнять максимально используя XPATH селекторы**

Lesson 6