«Проведение исследований в сети интернет с использованием регулярных выражений»

1. Содержание дисциплины

1.1 Лекционная часть

|  |  |  |  |
| --- | --- | --- | --- |
| № п/п | Наименование раздела | Объем часов | |
| Лекции | Самостоятельная работа |
| 1. | Введение в предметную область анализа данных. Виды данных, структуры данных, системы обработки данных. | 1 | 2 |
| 2. | Данные в сети интернет. Понятие формата данных. Форматы xml и csv. Язык гипертекста html. | 2 | 2 |
| 3. | Исследование в интернет. Задача исследования, формат результатов исследования. Методы проведения исследований на существующих данных в сети. Скрейпинг. | 2 | 2 |
| 4. | Язык программирования Python. Основные сведения: синтаксис, структуры данных, особенности. | 2 | 2 |
| 5. | Методы и средства скрейпинга на языке Python. Библиотека BeautifulSoup. Библиотека http-запросов requests. | 2 | 2 |
| 6. | Основы работы с регулярными выражениями и производство http-запросов. Библиотека регулярных выражений re. | 2 | 2 |
| 7. | Средство представления и обработки данных pandas. Конвертация DataFrame в формат .xlsx. | 2 | 2 |
| 8. | Основные возможности pyplot из библиотеки matplotlib для визуализации результатов исследования | 1 | 2 |
| 9. | Организация проведения исследования в интернет и представление его результатов | 2 | 4 |
|  | Итого часов: | 16 | 20 |

1.2. Самостоятельная работа

|  |  |  |
| --- | --- | --- |
| № п/п | Виды самостоятельной работы | Объем часов |
| 1. | Формулировка задания на исследование в интернет. Создание формализованного технического задания на исследование и четкое формулирование источников выборки, методик обработки и формы представления данных результата исследования | 2 |
| 2. | Создание простейших скриптов на Python: преобразование форматов данных, обработка списков, использование генераторов и функций. | 2 |
| 3. | Парсинг страницы с использованием библиотеки BeautifulSoup. Получение и верификация данных. Поиск и устранение багов в сборе данных. | 2 |
| 4. | Формирование http-запросов с помощью requests. Сбор данных с разных страниц. Фильтрация данных. | 2 |
| 5. | Поиск в больших данных с использованием средств библиотеки регулярных выражений re. | 2 |
| 6. | Создание DataFrame и обработка данных в них | 2 |
| 7. | Визуализация данных средствами pyplot | 4 |
| 8 | Создание и проведение исследования полного цикла по сформулированному заданию. Презентация проекта исследования и его результатов | 2 |