Аннотация общеуниверситетского факультатива «Программирование на языке Python для сбора и анализа данных»

Автор курса: доц. каф. высшей математики Щуров И.В.

Описание курса

Эффективное использование современных компьютерных систем немыслимо без получения базовых навыков программирования и опыта их приложения к реальным задачам. Умея писать несложные программы-скрипты, исследователь может упростить или автоматизировать решение множества задач, связанных со сбором и обработкой данных, взаимодействием с различными базами данных и сервисами, поиском информации, преобразованием файлов различных форматов и т.д. Перед современным учёным, работающим в любой области знаний (да и любым современным человеком) очень часто встают задачи, которые вообще не поддаются решению без навыков программирования. Так, например, всё больше сложноструктурированных данных, интересных для исследований, становятся доступны через программные интерфейсы (API) — их использование открывает чрезвычайно широкие возможности. Получение соответствующих навыков значительно повысит конкурентоспособность выпускников ВШЭ на рынке труда. Среда Jupyter (IPython Notebook), вместе с математическими библиотеками Python (numpy, scipy, SymPy) может использоваться как свободная и эффективная альтернатива таким математическим пакетам, как MatLAB и Wolfram Mathematica.

В ходе нашего курса мы будем использовать популярный в академической среде (и не только) язык Python и интерактивную среду IPython Notebook, позволяющую мгновенно получить результат выполнения фрагмента кода без отдельного этапа компиляции программы. Использование этих инструментов позволяет снять психологический барьер перед необходимостью «написать программу» — работа в такой среде скорее выглядит как «разговор с компьютером» на языке программирования и позволяет быстро продвинуться в изучении этого языка.

Слушатели курса имеют возможность заниматься как очно, так и дистанционно, а также совмещать разные формы обучения. В рамках курса доступны следующие ресурсы:

-Еженедельные очные лекции.

-Еженедельные практические занятия.

-Онлайн-трансляция лекций.

-Видеозапись лекций и текстовые конспекты.

-Автоматически проверяемые домашние задания.

-Форум для онлайновых консультаций с участием преподавателей и между самими студентами.

В качестве промежуточного контроля выступают домашние задания. В качестве итогового контроля — подготовка итогового проекта в интересующей слушателя области.

Темы курса

1.Базовые понятия программирования. Переменные, операторы, управляющие конструкции, циклы.

2.Типы данных в Python: строки, списки, кортежи, словари, множества.

3.Функции, их определение, особенности вызова функций.

4.Простейшие алгоритмы.

5.Элементы функционального программирования: list comprehensions, лямбда-функции, рекурсия.

6.Элементы объектно-ориентированного программирования. Класс, объект, свойство, метод. Примеры классов.

7.Ввод-вывод. Работа с файлами.

8.Работа с файлами различных форматов: CSV, XML, JSON и др.

9.Регулярные выражения, использование регулярных выражений для обработки текстов и извлечения данных.

10.Чтение и обработка (парсинг) веб-страниц (файлов HTML) с помощью BeautifulSoup, извлечение данных из веб-страниц.

11.Работа с внешними API различных сервисов.

12.Математические и статистические инструменты Python.

13.Построение графиков и диаграмм с помощью Matplotlib + Pyplot, а также других графических библиотек: seaborn, plot.ly и др.

14.Обработка данных в pandas.

15.Основы машинного обучения: примеры использования библиотеки scikit-learn.

Примеры практических задач, решаемых в рамках курса

1.Преобразовать 1000 файлов из формата Excel в формат CSV за несколько секунд.

2.Забрать с сайта ЦИК данные о результатах произвольных выборов по всем участковым избирательным комиссиям, найти участки, на которых кандидат-победитель набрал более 80% голосов, построить по ним диаграмму рассеяния «явка-результат», найти коэффициенты регрессии.

3.Скачать все статьи, опубликованные в некотором СМИ за последние три месяца; построить график зависимости частоты использования ряда слов-маркеров в комментариях к статьям в зависимости от времени публикации комментария.

4.С помощью API сайта Государственной Думы, найти всех действующих и бывших депутатов, имеющих учёную степень доктора наук.

5.Найти и перечислить все статьи о людях в Википедии.

6.По данным Всемирного банка, проверить гипотезу о связи между ВВП на душу населения и качеством образования.