Правительство Российской Федерации
Федеральное государственное автономное
образовательное учреждение высшего образования
«Национальный исследовательский университет «Высшая школа экономики»

Факультет компьютерных наук
Образовательная программа бакалавриата 01.03.02 «Прикладная математика и информатика»

ОТЧЕТ
по учебной практике

на факультете компьютерных наук НИУ ВШЭ
(название организации, предприятия)

Выполнил(а) студент(ка)
группы БПМИ191
А. М. Асланов
(инициалы, фамилия)

[ДАННЫЕ УДАЛЕНЫ]
(подпись)

Руководитель практики

Старший преподаватель Базовой кафедры Яндекс
(подразделение ФКН, должность)

Яруллин Рамиль Ильдарович
(ФИО руководителя практики)

Дата 10 сентября 2021 г. 9 (оценка) (подпись)

Москва – 2021

Оглавление

1. Постановка задачи 3
2. Описание решения 4
3. Анализ результатов 5
4. Выводы 5

1. Постановка задачи

В рамках учебной практики я прослушал мини-курс «Анализ и обработка веб-данных на Python» и выполнил по нему домашнее задание в виде программного проекта.

В курсе были рассмотрены следующие темы:

- Разведочный анализ данных, библиотеки numpy и pandas;
- Визуализация данных с помощью matplotlib и seaborn;
- Работа с API, библиотека requests;
- Парсинг html-кода с помощью BeautifulSoup и веб-скрапинг;
- Основы работы с фреймворком Flask, создание небольшого веб-приложения.

В качестве ДЗ было предложено написать какую-то версию своего программного продукта (связанного с использованием веб-данных и/или сервисов) на языке Python. Это могло быть веб-приложение на Flask, бот в ТГ или ВК, бот, ведущий публичную страницу в социальной сети, web-scraper/crawler с использованием BeautifulSoup, который собрал бы какие-то интересные/полезные данные, или что-то ещё по согласованию с руководителем.

Работа проверялась в соответствии со следующими критериями:

- Надо использовать какое-то API (желательно несколько) или/и BeautifulSoup;
- Надо как-то поработать с данными;
- Использовать хороший объектно-ориентированный стиль в коде: классы, статические и call-методы, комментарии к коду, использование typing и создание собственных типов с наследованием от NamedTuple, отлавливание ошибок с использованием кастомизированных исключений и т.д. Будет хорошо, если проверить стиль на PEP8;
- Реализация должна быть не очень простой и одноходовой, чтобы ощущался личный вклад в проект (скрипта на 50 строчек может быть недостаточно на отличную оценку);
- Проект должен быть новым для нас. Если уже есть что-то готовое, то это принимается только если как-то качественно дополнить;
- Это должно быть легально в т. ч. с точки зрения terms & conditions сервисов, которые используются.

Я решил написать web-scraper, который соберет информацию обо всех образовательных программах бакалавриата московского кампуса ВШЭ и будет парсить страницы с рейтингами студентов этих программ, представляя их в удобном для обработки и анализа виде. С помощью этого скрапера и пакета matplotlib я собрал, проанализировал и визуализировал данные о среднем показателе GPA (Grade Point Average), а также о среднем количестве отчисленных студентов по всем факультетам московской Вышки за 2018/2019, 2019/2020 и 2020/2021 учебные годы.

2. Описание решения

2.1. Данные

Все данные были собраны с сайта hse.ru. Меня интересовала страница со списком всех бакалаврских ОП московского кампуса, а также страницы самих ОП, на которых и хранится рейтинг.

2.2. Используемые методы

Бо́льшая часть ДЗ была выполнена в jupyter-тетрадке, там я в интерактивном формате описал, как была выполнена работа.

Сперва было необходимо взглянуть на данные, с которыми предстояло работать. Прежде всего, нужно было обработать страницу со списком всех образовательных программ. Обработка подразумевала получение html-кода страницы с помощью библиотеки *requests* и вычленение из него необходимой информации (название ОП, факультет реализации, ссылка на страницу ОП) сперва методом пристального взгляда, а затем программно с помощью *BeautifulSoup*. Возникали небольшие трудности с тем, что страница содержала много лишней информации, но она, тем не менее, легко фильтровалась теми же методами.

Далее я разбирался с тем, как устроено хранение рейтингов на сайтах ОП. Очень сыграло на руку то, что для (почти) всех ОП информация хранится в едином формате (начиная с 2018/2019 учебного года). Это и позволяло описать унифицированный интерфейс, способный обрабатывать бо́льшую часть образовательных программ. Исключение составили три ОП с ФКМД, их страницы устроены по-другому. Собранные рейтинги представлялись в виде *pandas*-овских датафреймов. На примере последнего (на тот момент) рейтинга ПМИ был проведен небольшой разведочный анализ данных.

Затем я спроектировал набор классов, которые осуществляли обработку и получение конкретного рейтинга по данному запросу. При тестировании обнаружились некоторые странности в полученных данных (например, у некоторых студентов средний балл был меньше минимального). Для чистоты анализа их пришлось отфильтровать. Также я написал функции, анализирующие полученные данные: вычисление среднего GPA и количества отчисленных студентов (считалось, что студент отчислен, если в некотором текущем рейтинге до пересдач запись о нем имеется, а в текущем рейтинге того же периода после пересдач — не имеется).

В последней части своей работы я написал (очень долго работавший) фрагмент кода, который вычислял и усреднял для каждого факультета описанные в предыдущем абзаце показатели по всем реализуемым на нём ОП. Ну и самое интересное — это визуализация полученных данных (с помощью *matplotlib*) и выводы (об этом в следующем разделе).

Также, как было рекомендовано руководителем, я вынес составляющую, которая отвечает за сбор данных, в отдельный скрипт. Он запрашивает параметры рейтинга, который интересует пользователя, и (если таковой нашелся) сохраняет его в формате .csv.

2.3. Код

Весь код находится в репозитории https://github.com/alexis852/hse-python-practice-2021.

3. Анализ результатов

По собранным данным и получившимся картинкам (см. jupyter-тетрадку) можно сделать следующие выводы:

- Можно сказать, что в среднем (если не считать МИЭФ) студенты учатся примерно на 7. Самые высокие оценки в среднем на ФКМД и ШИЯ. А на родном ФКНе оценки в среднем ниже, чем на большинстве факультетов;
- Что касается выбросов по МИЭФу, то при первом взгляде на один из самых свежих рейтингов, сразу складывается ощущение, что это определенно какие-то технические ошибки, с этим ничего не поделать, данные просто некорректны;
- В общем и целом (если не считать отдельные выбросы по тому же МИЭФу и 4 курс), картина получилась довольно реалистичная. Можно сказать, что в среднем меньше всего отчисляются на 3 курсе. Но примечательно, что довольно много студентов отчисляются на 4 курсе. Это может быть связано с теми же техническими ошибками. Или же моё, в общем-то, довольно логичное предположение о том, что рейтинг выпускников продолжает текущий рейтинг II семестра на 4 курсе неверное. В таком случае, у нас неполная картина и данные, стало быть, достоверными считать нельзя. Наконец, это вполне себе может быть реальностью, и тогда многие студенты, наверное, просто не выдерживают работу над дипломом ¯_(ツ)_/¯

4. Выводы

Благодаря пройденной практике, я освоил работу с основными библиотеками для анализа и визуализации данных, научился собирать и парсить данные с веб-страниц, освежил и пополнил свои знания языка Python.

В качестве продолжения моего проекта можно было бы, например, написать бота в ТГ, который бы отвечал на запросы по динамике в рейтинге конкретного студента, а именно — выдавал бы красивый график, где наглядно показана успеваемость студента за выбранный период обучения.