**УДК 004.056**

**Парсинг данных**

***Парфенов А.К., Маврин В.В.***

ФГБОУ ВО «Воронежский государственный университет инженерных технологий» г. Воронеж, Россия

В современном мире большая часть информации хранится в интернете. Чтобы получить ещё больше пользы из информации, её можно структурировать. В этом помогает парсинг - процесс сбора и систематизирования информации, размещенной на определенных сайтах, с помощью специальных программ, автоматизирующих процесс, такие программы называются парсерами. Парсер – это специальная программа, нацеленная на сбор информации. Эта программа позволяет взять файл в одном формате и преобразовать его данные в более удобоваримую форму, которую можно использовать в своих целях. С каждым годом в мире появляется всё больше и больше новой информации. По данным аналитической компании IDC, объем данных, содержавшихся на серверах интернета в 2018 году, составлял 33 зеттабайта, рост данных в интернете IDC оценивает в 30% в год, с прогнозом 175 зеттабайта к 2025 году, парсинг позволяет работать со всей этой информацией, что делает его крайне актуальным.

Чтобы написание парсера было эффективным следует знать основные его методы: поэтапный парсинг и метод паука.

В первом методе парсинг происходит так: сначала парсятся все внешние элементы, далее парсятся элементы, содержащиеся во внешних элементах, затем это действие повторяется с полученными элементами, в итоге этот цикл повторяется до тех пор, пока мы не получим всю необходимую информацию. Разберём данный метод на примере парсинга книг интернет-библиотеки. Пусть в данной библиотеке имеются категории, жанры, поджанры и книги. Каждая книга принадлежит определенному поджанру, которая в свою очередь принадлежит жанру, который принадлежит категории. Сначала парсятся все категории, затем программа просматривает все жанры и парсит все поджанры, далее парсер заходит в каждый поджанр и парсит там все книги. Это и есть поэтапный парсинг. Из примера можно понять, что данный метод хорош в тех случаях, когда нужно собрать всю информацию, а не отдельные части. Главная проблема этого метода – это его сложность при возрастании количества уровней.

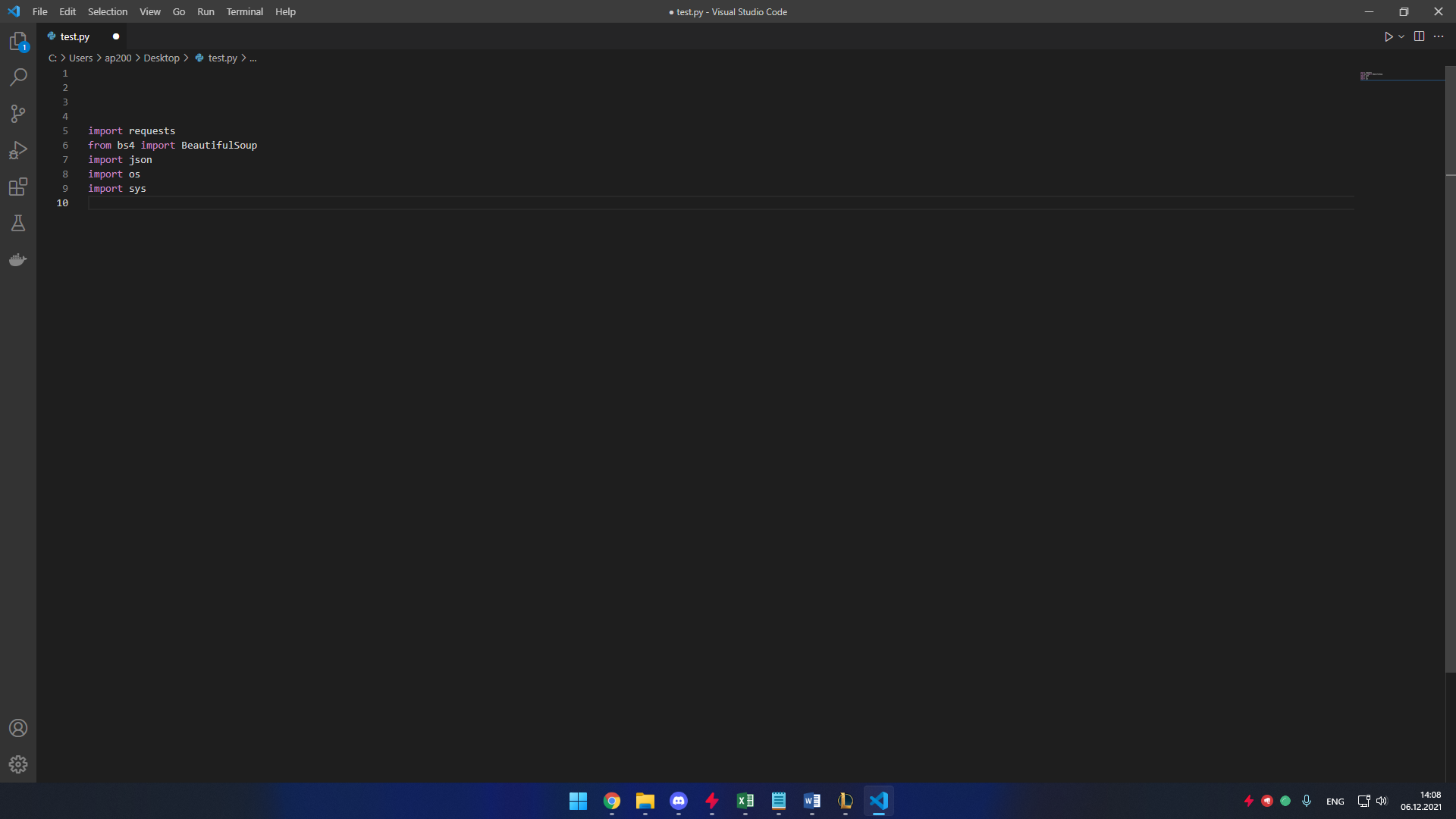
При Методе паука парсер открывает веб-страницу, собирает все ссылки с сайта, добавляет их в общую базу данных. Затем открывает первую ссылку, ищет все ссылки и добавляет в базу данных только те, которые отсутствуют. Также при открытии страницы программа разбирает контент страницы. Если этот контент нужен нам, он сохраняет его в базу данных. Как понятно из описания данный метод будет быстрее работать с большими объёмами данных по сравнению с поэтапным парсингом.

Теперь разберём пример реализованного парсера сайта ВГУИТ, который собирает информацию о факультетах данного учебного заведения.

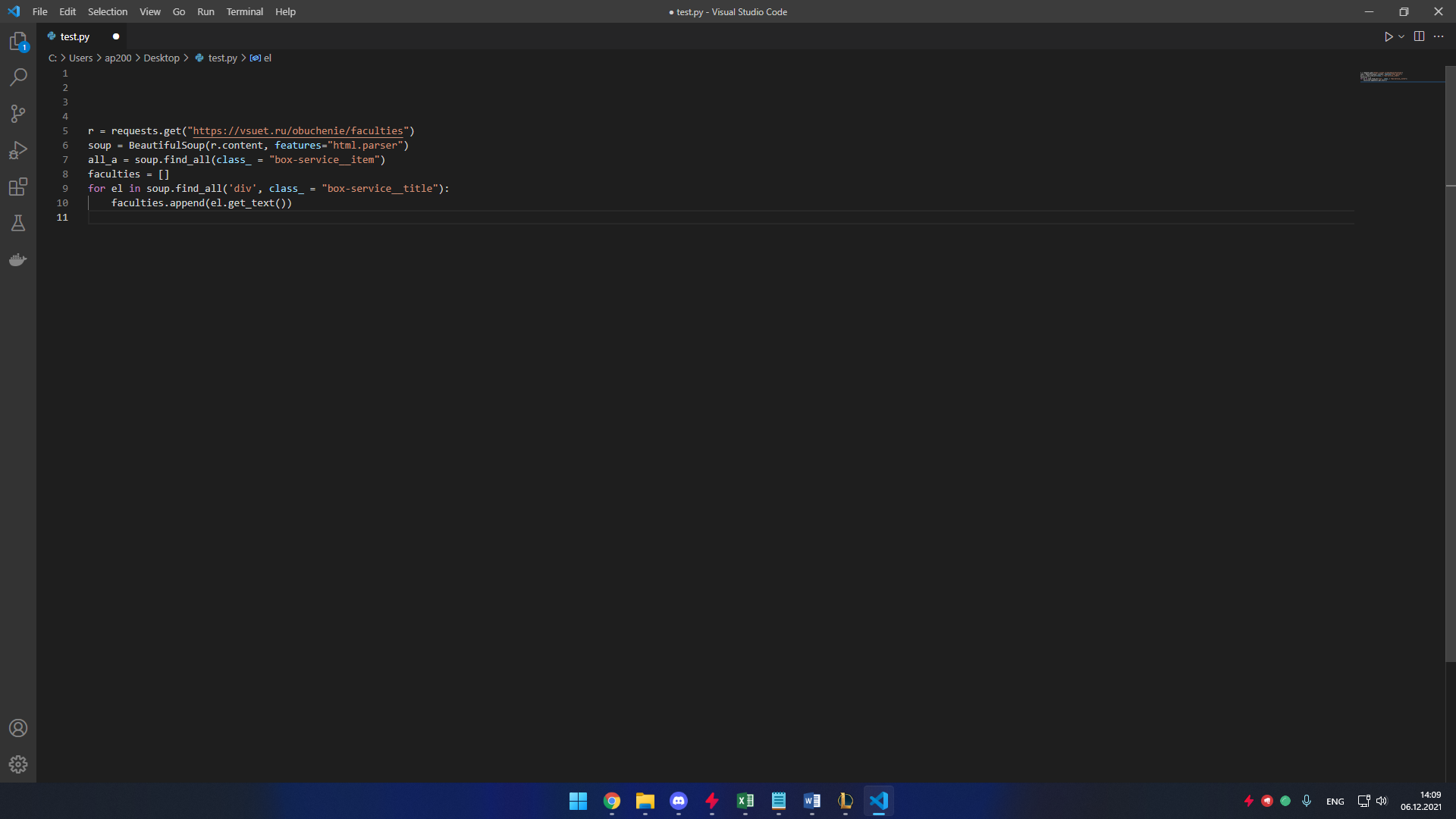
Цель данной программы – это сбор информации о факультетах. Парсер будет использоваться на сервере, а при запросе со стороны приложения полученная информация будет отправляться в виде JSON-массива для дальновидного преобразования в объекты и использования. В итоге приложение будет предоставлять информацию в структурированном и упрощённом виде.

Разберём наиболее важные части кода программы по парсингу информации факультетов ВГУИТ.

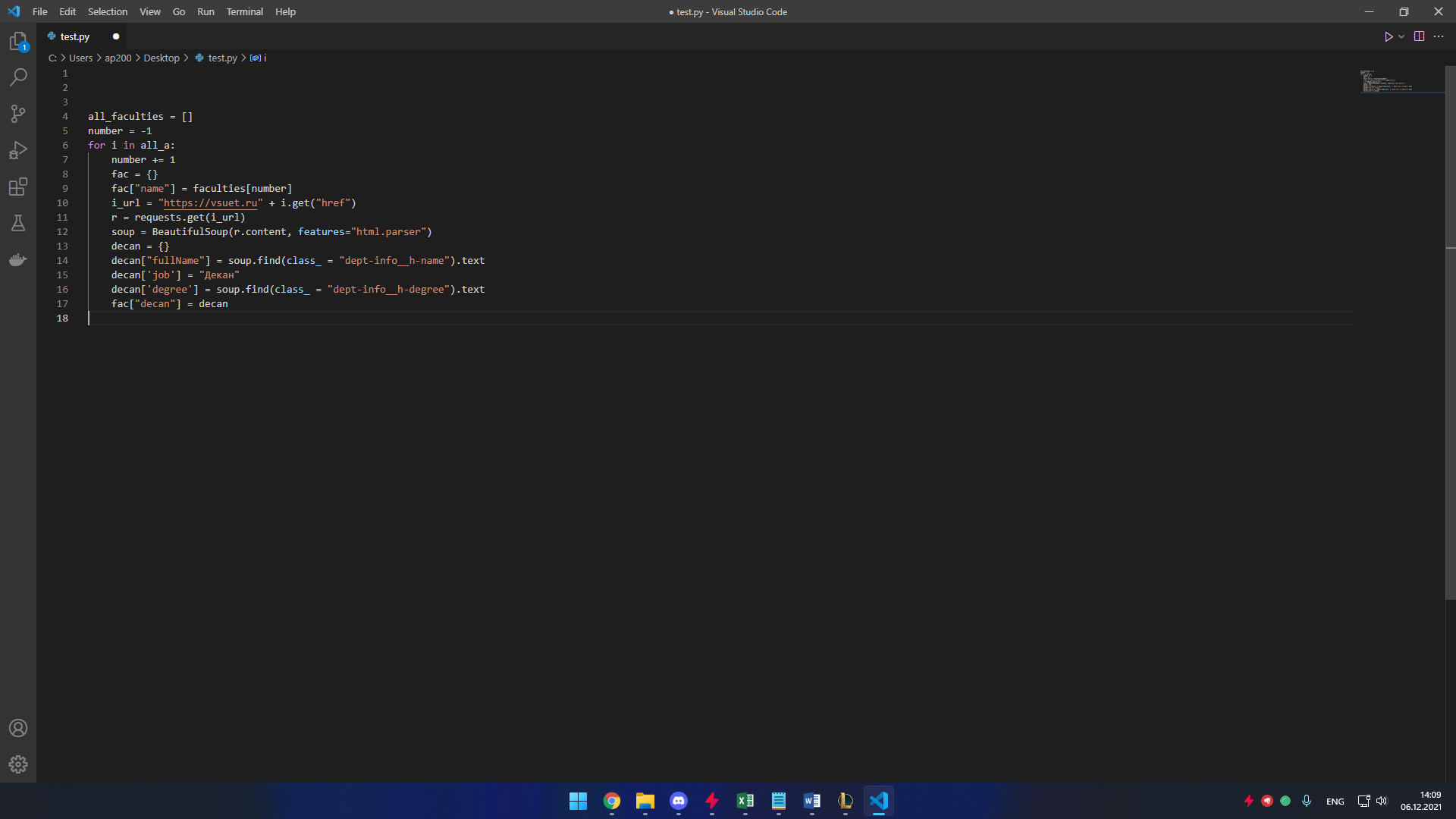
1.Импортируем нужные библиотеки для более удобной работы.



2.Подключаемся к веб-странице со всеми факультетами и собираем ссылки на них.

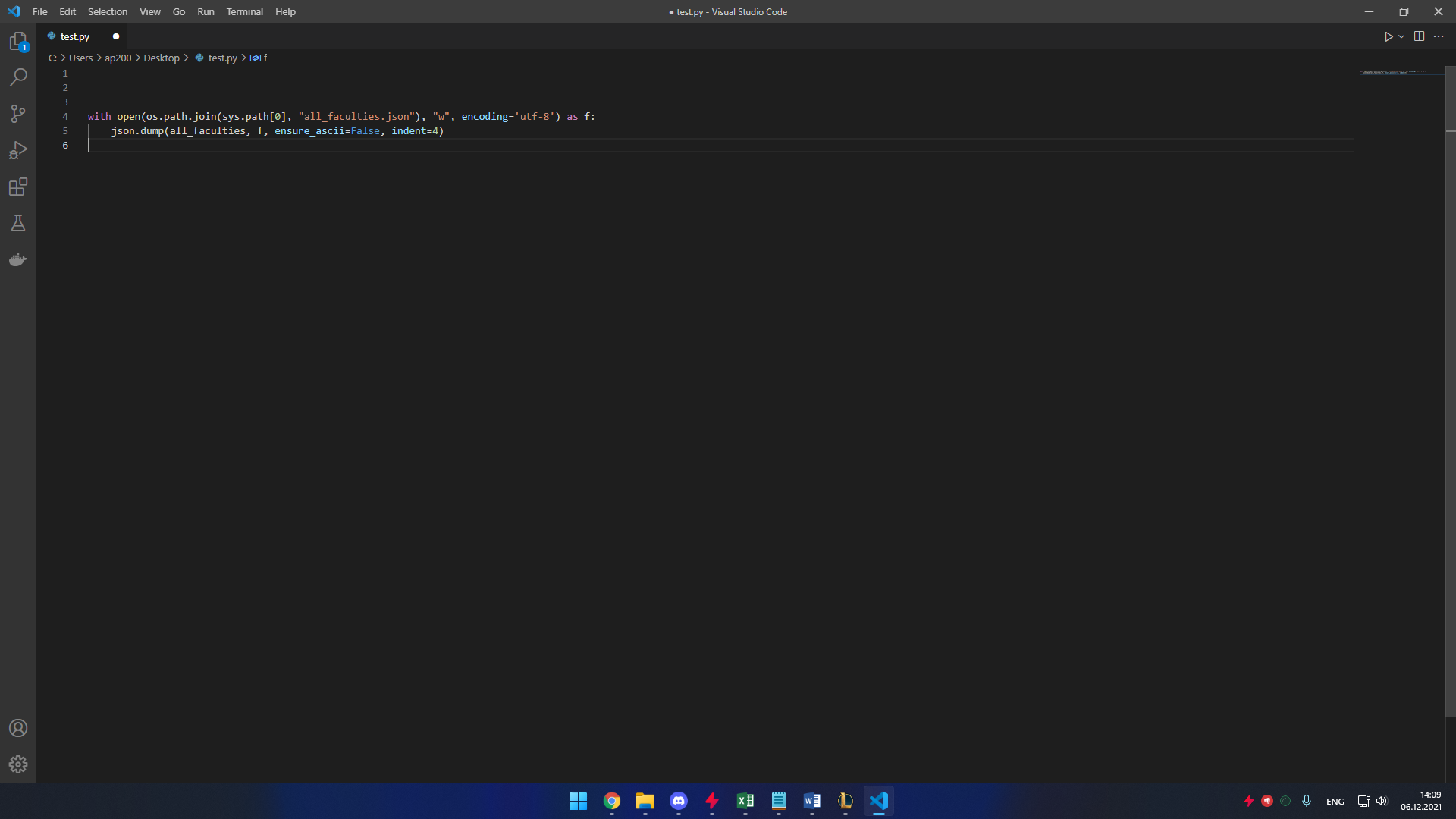
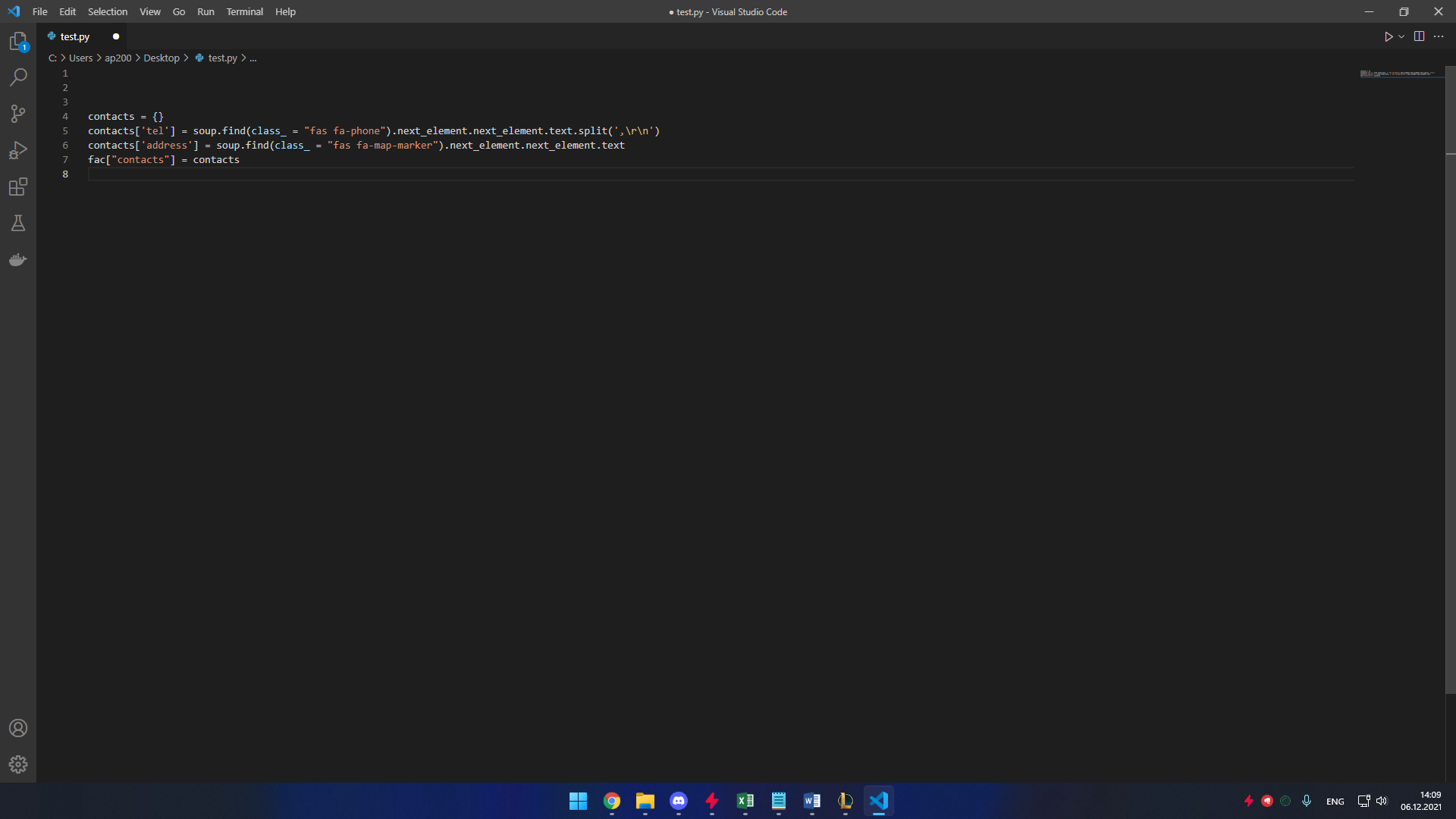


3.Создаём главный массив и проходим цикл, с помощью которого мы собираем информацию о декане каждого факультета.

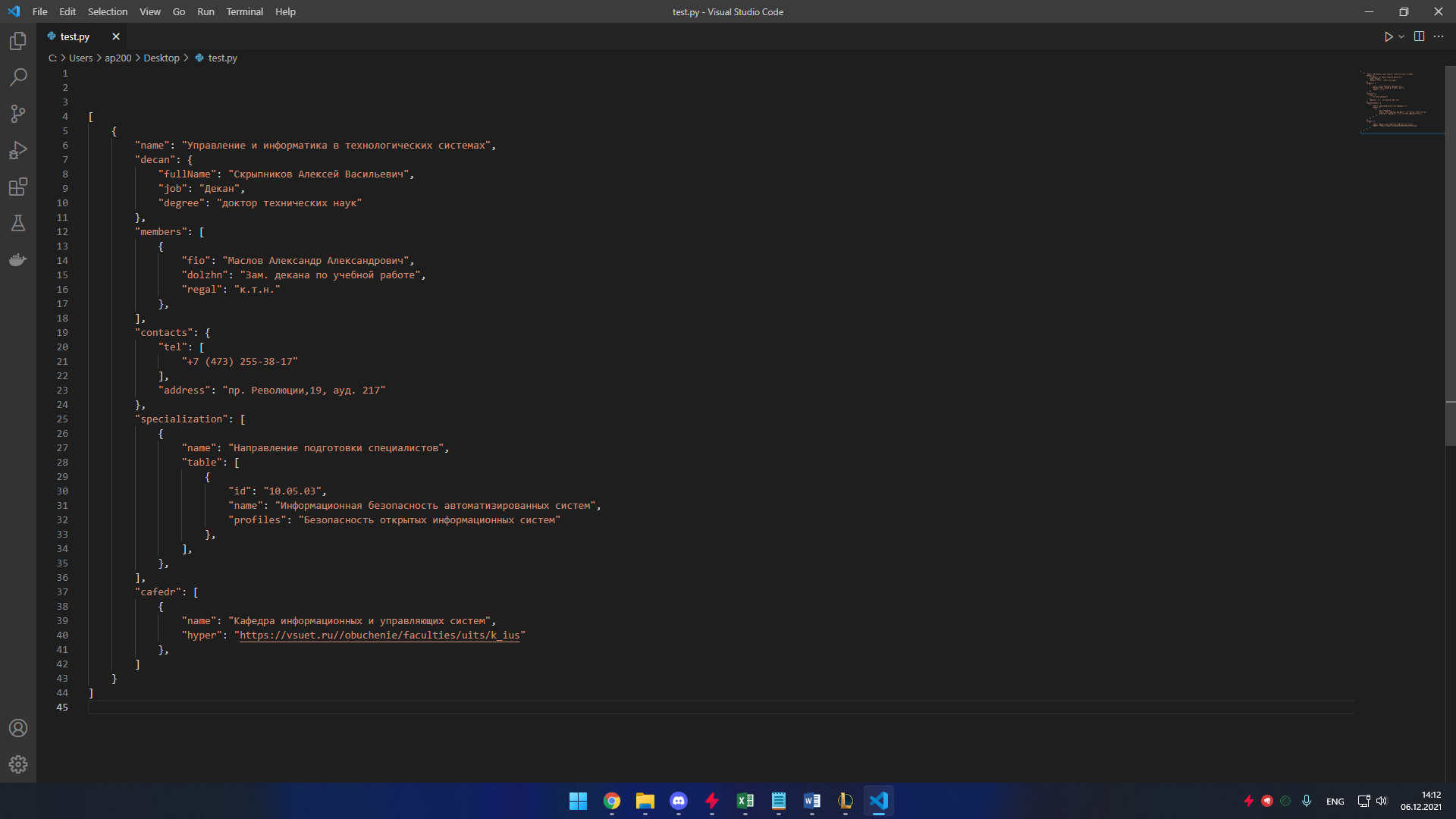


4.Собираем контактные данные факультетов

5.Преобразуем полученную информацию в формат JSON.



В итоге мы получаем JSON файл следующего вида:



Таким образом мы собрали нужные нам данные, а также преобразовали их в нужный нам формат. Всё вышесказанное позволяет сделать вывод - парсинг страниц - это ведущий способ сбора информации с сайтов. Этот способ используется почти во всех сферах бизнеса: маркетинг, сбор информации по продажам конкурентов и так далее. По исследованию Hubspot основные цели применения: контент 38%, аналитика 26%, контакты 19%, сравнение цен 16%. Одна из основных проблем информации с которой встречаются разработчики – это отсутствие структурированности информации. Справится с этой ошибкой помогает парсинг, именно из-за этого он в ближайшем будущем будет становится только популярнее.

Список литературы

1 Что такое парсинг и как правильно парсить [Электронный ресурс]. - Режим доступа: https://blog.calltouch.ru/chto-takoe-parsing/ (дата обращения: 04.12.2021).

2 Сколько «весит» вся информация в интернете? | Публикации | Вокруг Света [Электронный ресурс]. - Режим доступа: <https://www.vokrugsveta.ru/quiz/329408/> (дата обращения: 04.12.2021).

3 Поэтапный парсинг и метод паука | Трепачёв Дмитрий [Электронный ресурс]. – Режим доступа: <http://old.code.mu/books/advanced/php/parsing/poetapnyj-parsing-i-metod-pauka.html> (дата обращения: 04.12.2021).

4 Парсинг и парсер: что это такое, для чего нужно и как этим пользоваться [Электронный ресурс]. - Режим доступа: <https://romi.center/ru/learning/what-is-data-parsing> (дата обращения: 04.12.2021).