Министерство науки и высшего образования Российской Федерации

федеральное государственное автономное образовательное учреждение высшего образования

«Национальный исследовательский университет ИТМО»

Факультет инфокоммуникационных технологий

**ПРАКТИЧЕСКАЯ РАБОТА №3**

**Выполнил**:

студент 4 курса ИМРиП

Группы D34101

Ф.И.О. Лю Ифэн

**Руководитель**:

Чернышева Анастасия Вадимовна

Санкт-Петербург

2024

Объедините данные, полученные с веб-сайтов

手机屏幕截图

描述已自动生成

Сначала я подготовил данные, собранные на сайте.Поскольку я получил данные из нескольких городов, для более удобной их обработки я объединил их в один файл, а чтобы использовать код для их обработки, я сохранил формат файла как формат csv

Из-за политики конфиденциальности веб-сайта мне не удалось получить имя пользователя, но мой код был разработан с учетом функции получения имени пользователя, поэтому столбец имени полученных данных был пустым значением, поэтому я удалил столбец имени.

<https://github.com/l769030551/lab2>

Возврат каретки и перевод строки удалены, благодаря чему данные выглядят аккуратно.

Поскольку данные, собранные на веб-странице, представляют собой комментарии, оставленные пользователями. В комментариях, написанных пользователями, возврат каретки используется для переноса строк из-за слишком большого количества контента. Однако данные, которые мы собираем на веб-сайте, не будут автоматически удалять переводы строк и возвраты каретки, что приведет к путанице в данных. Один фрагмент данных может занимать пространство нескольких строк, что приводит к искажению данных.

Поэтому я удалил символы новой строки и возврат каретки в данных.

Далее я проверил целостность данных. Чтобы предотвратить потерю данных из-за колебаний сети во время сбора данных.

Сначала я проверил данные в каждом столбце. Если данные отсутствуют, строка с отсутствующими данными будет удалена, и мне будет сообщено об удаленных данных. Если данные полные, мне сообщат, что проверка данных прошла.

После выполнения операции проверка данных прошла

Обработанные данные полностью исключают возможные пропущенные данные, повышают целостность данных и улучшают читаемость данных.