



Дисциплина «Искусственный интеллект и анализ данных»

Кураторы: [АИС Город](#) и [УлГТУ](#)

Описание задания

Работа сотрудников коммунальной сферы не терпит долгих ожиданий, особенно если дело относится к работе с отключениями воды. Ведь чем быстрее население будет оповещено об отключениях (плановых или аварийных), тем меньше будет недовольных граждан.

Закон обязывает регистрировать все отключения коммунальных ресурсов в специальной федеральной информационной системе. Эта система требует точного указания перечня домов, которые были затронуты отключением.

Но операторы диспетчерской службы не могут тратить много времени на выбор всех адресов, особенно если авария на коммунальных сетях затрагивает сразу половину района города.

Предлагается разработать такой алгоритм, который по заданным доступным адресам домов и текстовому описанию отключения максимально точно определял бы перечень домов, которые были упомянуты.



Например, диспетчер в описании аварии печатает «из п/з Д=100, без ХВС К. Либкнехта 21, 23, 23а 25», а система должна распознать следующие адреса:

- Ульяновская обл, Ульяновск г, Карла Либкнехта ул, 21;
- Ульяновская обл, Ульяновск г, Карла Либкнехта ул, 23;
- Ульяновская обл, Ульяновск г, Карла Либкнехта ул, 23А;
- Ульяновская обл, Ульяновск г, Карла Либкнехта ул, 25.

Оператору остаётся лишь подтвердить правильность распознавания или немного отредактировать список домов.

Исходные данные

1. Файл «volgait2024-semifinal-addresses.csv», в котором содержится список доступных адресов для поиска.

Столбцы:

- house_uuid - УИД (токен) дома
- house_full_address - Текстовое описание адреса

2. Файл «volgait2024-semifinal-task.csv», в котором содержатся комментарии диспетчеров, оставленные при регистрации отключения.

Столбцы:

- shutdown_id - ИД (токен) отключения;
- comment - Собственно комментарий, содержащий адреса, которые необходимо распознать из файла 1.



Формат всех файлов - CSV, разделители - «;», кодировка - UTF-8, первая строка - заголовок.

Требования к решению

1. Программа должна принимать на вход файл с заданием в таком же формате, что и «volgait2024-semifinal-task.csv». Путь до файла должен задаваться каким-либо образом в коде решения.

2. Для каждой строки (отключения) из комментария (поле «comment») должны распознаваться адреса из файла «volgait2024-semifinal-addresses.csv» и сохраняться их УИДы (поле «house_uuid»).

3. Для определения домов должны использоваться элементы ИИ.

4. По результатам работы программы рядом с файлом из п.1 должен создаваться файл «volgait2024-semifinal-result.csv» со следующими столбцами:

— shutdown_id - ИД (токен) отключения - поле из файла «volgait2024-semifinal-task.csv»;

— house_uuids - Строка, в которой будут указываться УИДы домов (поле из файла «volgait2024-semifinal-addresses.csv») через запятую.



Пример файла «volgait2024-semifinal-result.csv»:

```
"shutdown_id";"house_uuids"  
1;412823ab-c3ea-4988-a93d-9822212cddfc,6d4303b9-6182-  
48f0-af46-870d8c8f22c3  
2;ec1709ec-7951-4fcb-af6d-5136ded6918c  
3;  
4;4641ff60-b5db-4b84-b55f-8f2dc6fd68f6
```

Таким образом, в файле выше описан следующий результат:

— для отключения с ИД = 1 система распознала адреса с идентификаторами «412823ab-c3ea-4988-a93d-9822212cddfc» и «6d4303b9-6182-48f0-af46-870d8c8f22c3»;

— для отключения с ИД = 2 система распознала адрес с УИД «ec1709ec-7951-4fcb-af6d-5136ded6918c»;

— для отключения с ИД = 3 система не смогла распознать адреса домов;

— для отключения с ИД = 4 система распознала адрес с УИД «4641ff60-b5db-4b84-b55f-8f2dc6fd68f6».

Результат выполнения задания

1. Исходные тексты решения для их проверки и запуска членами жюри.
2. Если в решении используется обучение каких-либо моделей, обученные модели также должны прикладываться к решению.
3. Если запуск решения не тривиален, должна быть приложена краткая инструкция по запуску решения.



Примечания

1. Изучите более детально те комментарии об отключениях, что приложены к заданию. Можно выделить несколько паттернов, как задаются адреса. Чем их больше Ваше решение сможет обработать - тем лучше.
2. Не для всех комментариев удастся определить список домов. Это нормально.
3. Порядок домов в поле «house_uuids» не важен.