Описание

Цель

Выделить признаки из урлов через алгоритм «Секитей»

Состав

Тест содержит пять сайтов, т.е. пять наборов файлов с урлами. Файлы лежат в папке ./data/. Каждый сайт представлен парой файлов, которые содержат в себе хорошие и неизвестные ссылки. Хорошие ссылки находятся в файлах, содержащих в имени “examined”. Неизвестные урлы – в файлах, содержащих в имени «general». В папку ./check будет помещаться вывод программы extract_features, так же там находятся файлы, с которыми будут сверятся результаты теста.

Название	Описание
./data/	папка с данными
./check/	папка с результатами
./check-features.py	скрипт теста
./extract_features.py	заготовка для реализации задания

Описание

Тест содержит три открытых сайта, т.е. данные, на которых студент может производить тест. И два закрытых сайта, т.е. результаты, которые не будут доступны, и по ним будет проверятся качество работы. Результат работы теста может быть **PASSED**, что означает тест пройден, или **NOT PASSED**, что означает тест не пройден и тогда будет показана причина провала теста и имя теста.

Студенту нужно реализовать метод **extract_features** в модуле **extract_features.py**. Заменить файл в папке теста на свой.

Запуск

Распаковать архив в отдельную папку, заменить файл **extract_features.py** на свою реализацию и запустить скрипт проверки

python ./check-features.py

Результаты

Отлаженный скрипт прислать в качестве выполненного ДЗ.

Описание параметров

1. Файл *examined*
2. Файл *general*
3. Файл, в который нужно записать результаты теста.

Формат файла результатов:

<Признак>\t<Количество>\n
<Признак>\t<Количество>\n
….

Файл результатов должен быть отсортирован по количеству признаков.

Привила именования фичей:

Для сегментов:

segment_<name>_<index>:<val>

где:

name – название фичи для сегмента
index – индекс сегмента
val – значение фичи

Для параметров

param_name:<название параметра>
param:<ключ=значение>

Для описанных правил имеем получаем следующие имена фичей

1. Количество сегментов в пути - **segments:<len>**
2. Список имен параметров запросной части (может быть пустым) **param_name:<имя>**
3. Присутствие в запросной части пары <parameters=value> - **param:<parameters=value >**
4. Сегмент пути на позиции :
    a) Совпадает со значением <строка> - **segment_name_<index>:<string>**
    b) Состоит из цифр - **segment_[0-9]_<index>:1**
    c) Совпадает со значением <строка с точностью до комбинации цифр>: <строка><цифры><строка> - **segment_substr[0-9]_<index>:1**
    d) Имеет заданное расширение - **segment_ext_<index>:<extension value>**
    e) Комбинация из двух последних вариантов - **segment_ ext_substr[0-9]_<index>:<extension value>**
    f) Состоит из данного количества символов: **segment_len_<index>:<segment length>**

Знак <> означает подстановку значеня, например, <index> -означает, что нужно использовать индекс сегмента: **segment_substr[0-9]_1:1.** Первый сегмент имеет фичу **substr[0-9]**