ДЗ #2.

Цель

Разработать алгоритм, имитирующий приоритезацию спайдера на основе алгоритма «Секитей» с качеством, превышающим «Жадный» алгоритм.

Описание.

Используя алгоритм «Секитей» извлечения признаков из урлов и любой из алгоритмов кластеризации определить нужно ли качать входящий урл или нет. Датасет разделен на два множества тренировочное, три сайта, и валидационное, два сайта. Для каждого сайта нужно будет максимально эффективно выбрать доступную квоту. Квота – максимальное количество урлов, которое может быть взято с данного сайта. Это число передается в качестве параметра на вход алгоритму для принятия решения о важности урла.

Реализация.

Студенту нужно реализовать две функции

1. ***define_segments***, выделяющая сегменты из сайта и квоты для них. На вход получает 500 урлов с кулинками, и 500 урлов без кулинок, а также значение квоты для всего сайта в целом.
2. ***fetch_url*** , определяющая нужность урла. На вход принимает урл как параметр на выход должна вернуть истину если урл нужно положить в индекс иначе – ложь.

Как все работает

Тест взывает метод определения сегментов для сайта (define_segments) с некоторой рандомной выборкой урлов для инициализации (по 500 урлов для каждого класса). Для оставшихся урлов вызывается метод, определяющий ценность урла, если урл нужно положить в индекс функция должна вернуть True, в противном случае False. Если функция fetch_url возвращает True, то тест уменьшает значение квоты. Обработка заканчивается, когда достигнута граница квоты или больше нет урлов для выборки.

Метрики

В качестве метрики используется F1 мера

$$\mathrm{F1} = 2 * \frac{precition * recall}{precition + recall}$$

Где:

precition – полнота выборки

reacall - точность выборки

$$\mathrm{precition} = \frac{1}{T}\sum_{i=0}^{T}\frac{N_{fetched}}{N_{quota}}$$

Nfetched – количество выбранных документов

Nquota - количество документов разрешенных квотой

T – количество тестов (сайтов)

$$\text{recall} = \frac{1}{\text{T}}\sum_{i}^{T}\frac{N_{qfetched}}{N_{qtotal}}$$

Nqfetched – количество отобранных документов с кулинками

Nqtotal - количество документов с кулинками всего

Значение F меры	Балы
➢ **0.7**	2
➢ **0.8**	5
➢ **0.9**	10

На каждый сайт должно уходить не более 15 секунд.

Прототипы функций

define_segments(QLINK_URLS, UNKNOWN_URLS, QUOTA)

QLINK_URLS – массив урлов с кулинками

UNKNOWN _URLS – массив урлов без кулинок

QUOTA – размер квоты для сайтов

fetch_url(url)

url – урл для оценки

возврат:

True - урл нужно положить в индекс

False – урл не нужен