Техническое задание на разработку инструмента

«Краулер отзывов из картографических

источников»: Google Maps, TripAdvisor, 2GIS, Yandex.maps

1. Требования к функциональности

Инструмент решает задачу сбора и парсинга данных отзывов из

источников:

• Google Maps

• 2GIS

• TripAdvisor

• Yandex.maps

1.1. Входные и выходные данные

1.1.1 Входные: Строка длины, не превышающей n в кодировке utf-8, где n=100 – длина запроса, символы. Пример: «Кофейни Москвы», «Бургерные Санкт-Петербург»

1.1.2 Выходные: модуль записи выходных данных должен быть абстрагирован и поддаваться легкой модификации для записи в различные базы данных. В текущей реализации инструмент записывает данные в .csv файл с оговоренными полями.;
А также текст отзыва должен быть минимально предобработан (очищен от html-сущностей, фрагментов кода и т.п.). Текст должен быть приведен полностью, либо, если текст превышает указанную максимальную длину текста *L*, приводиться до последнего предшествующего *L* пробела (*L* = 8000 – максимальная длина текста отзыва)

1.2. Информационные поля

Информация из источников собирается единообразно

Информация об организации в определенной точке

1. Местоположение (координаты) организации записывается в БД в поле

orgLoc в едином формате.

2. Текстовый адрес организации записывается в поле **orgAddress**, в

кириллице.

3. Название организации записывается в поле **orgName**, в кириллице.

4. Рейтинг организации в данной точке записывается в поле **orgStar**s в виде

числового значения с одним знаком после запятой.

5. Информация окатегории организации(ее сфере деятельности) в

таксономии источника записывается в поле **orgType**. В случае, если

источник присваивает организации более одного типа, они записываются в поле через разделитель.

6. Информация о сайте организации, при ее наличии, записывается в поле **orgSite**.

7. Вся информация о дополнительных признаках организации (оплата, предложения, особенности заведения и т.п.) записывается в поле **orgFeatures** в формате feature1: value1, value2; feature2: value3 и т.д.

8. Источник записывается в поле **source** в виде ссылки на организацию.

Информация об отзыве

1. Каждому отзыву присваивается уникальный идентификатор и записывается в поле **revid**.

2. Текст отзыва предобрабатывается и записывается в поле настраиваемой длины n **revText**

3. Дата и время сбора отзыва соответствует моменту отправки батча отзывов на запись в файл/БД и сохраняется в поле revDTTM

4. Дата публикации отзыва записывается в поле **revPubDate** в едином формате DD.MM.YYYY.

5. Рейтинг отзыва соответствует количеству звезд и записывается в поле **revStars**.

2. Требования к реализации

1. Модуль записи собранных данных абстрагирован от логики сбора и предобработки и модифицируется под запись в различные БД;

2. Краулер поддерживает многопоточный сбор и запись;

3. Можно запустить несколько копий краулера одновременно;

4. Краулер сдается заказчику в виде исходного кода на Python3;

5. Сбор 1000 отзывов осуществляется не более, чем за 60 сек.

6. Унификация координат – отзывы организаций из всех источников если и пишутся с координатами, то координаты берутся из Google Maps. То есть в разных источниках координаты одного и того же ресторана (объекта) могут отличаться - нужна унификация в случае присутствия данного объекта в Google Maps - то есть, если этот объект присутствует помимо, к примеру, 2gis-а, присутствует также и в Google Maps - то указывать координаты Google Maps.