Домашнее задание 1. Сбор текстовых данных.

Дедлайн: 04 ноября 23.59.

Формат сдачи: материалы (jupyter-ноутбук с готовым парсером и полученный датасет в формате .xlsx) выгрузите на любое облачное хранилище с предоставлением доступа по ссылке. Jupyter-ноутбук и данные необходимо назвать в формате ***Фамилия_Имя.ipynb*** и ***Фамилия_Имя.xlsx*** соответственно. Ссылку на облачное хранилище отправить до дедлайна на почту vbaklanova@hse.ru с темой письма ***Д31_Фамилия_Имя*.**

Инструкция:

1. Выберите 1 криптовалюту для анализа.
2. Выберите 1 источник текстовых данных: сайт/телеграм-канал, где обсуждение ведется только по Вашей криптовалюте. Внимание! Если Вы выбираете парсинг телеграм-канала, Ваша максимальная оценка – 6 баллов.
3. Впишите их напротив ФИО в ведомости на Github. Внимание! Связки "криптовалюта-сайт" не должны повторяться между студентами.
4. Выберите гранулярность данных:
   - если моделирование планируется на дневной основе, период парсинга должен составлять 1 год.
   - если моделирование планируется на часовой основе, период парсинга должен составлять 1 месяц. Внимание! Этот вариант может быть выбран только в том случае, если Вы анализируете популярную криптовалюту. Количественно это означает, что каждый час должно быть в среднем минимум 100 комментариев.
5. Скачайте котировки выбранной в п.1 криптовалюты с выбранными в п.3 гранулярностью и периодом. Эти данные Вам понадобятся в конце курса.
6. Приступайте к парсингу комментариев из выбранного в п. 2 источника данных за выбранный в п.3 период. Внимание! Пользоваться можно только изученными в ходе курса библиотеками. Важно! В начале jupyter-ноутбука необходимо прописать Вашу логику парсинга (то есть что и по какому принципу вы парсите, как ставите временную метку к комментарию и т.д.). Вы вольны выбрать любую логику парсинга. Это означает, что Вы можете качать как все комментарии подряд, так и задействуя определенный принцип (например, парсить комментарии только в самых популярных обсуждениях). Особенно такой подход может быть применен, если Вы выбрали широко известную криптовалюту, которая имеет массу обсуждений и нужно отфильтровать самые ценные из них, для того а) чтобы получить текстовые данные, которые с бОльшей вероятностью будут коррелировать с рынком; б) чтобы Ваш компьютер не умер во время парсинга.
7. В результате у Вас должен получиться датафрейм со следующим минимальным набором столбцов: [datetime, comment]. Внимание! Иногда, когда сайт имеет структуру «тема обсуждения-комментарии к ней», дата и время комментариев не отображаются. В таком случае, позволительно записывать дату и время создания темы обсуждения в качестве даты и времени для всех комментариев к этой теме. Также приветствуется сбор косвенных признаков. Например, полезно собирать количество лайков к комментарию и т.п. Подобные доп. атрибуты позволят Вам в дальнейшем составить более «чувствительный» к динамике рынка индекс сентимента.
8. Проверьте Ваш jupyter-ноутбук на чистоту и аккуратность, отсутствие лишнего кода. Добавьте пояснения к каждой смысловой части кода.

Будут оцениваться воспроизводимость, корректность и универсальность написанного парсера, а также качество пояснений к коду и логике парсинга. УДАЧИ!

N.B. Может понадобиться обойти капчу, для этого обратитесь к https://rucaptcha.com/lang/python