

ДОМАШНЕЕ ЗАДАНИЕ

До 23:59 01.09.2026

1 Общие требования

Работа выполняется индивидуально. Сдайте одним архивом три файла:

1. Ноутбук `.ipynb`. Все ячейки должны быть выполнены, а код - запускаться сверху вниз без ошибок. Комментарии и ответы на вопросы оформите в markdown-ячейках рядом с соответствующим кодом.
2. Папку `cache/` с исходными HTML-страницами. Первый запуск кода может обращаться напрямую к источнику информации (сайту). Повторный запуск кода должен использовать сохраненные страницы и не отправлять запросы к сайту заново. Работа без папки `cache/` не принимается.
3. Итоговый датасет: это могут быть данные по нескольким эмитентам в один день, так и для одного эмитента за несколько дней. Идеальный вариант – попробовать собрать панельные данные. Включите колонки `ticker`, `date`, `news_tone` и число публикаций, вошедших в дневной показатель.

Собирайте данные только из разделов, в которых автоматический сбор разрешен. Не используйте прокси-ротацию, обход CAPTCHA, подмену отпечатка браузера или параллельные запросы без пауз. Если источник не предоставляет данные для автоматической загрузки, выберите другой источник.

2 Задание

1. Выберите один открытый источник: новостную ленту финансового СМИ, ленту раскрытий, форум или сообщество инвесторов либо раздел статистики Банка России. Выберите одну или несколько публичных компаний, по которым будете собирать данные.
2. Соберите публикации по выбранным компаниям. Для каждой публикации сохраните дату, заголовок и ссылку. Ориентир - не менее 60 записей. Если за выбранный период публикаций меньше, увеличьте период и укажите это в работе.

Каждую страницу скачивайте один раз и сохраняйте в кэш. По возможности код должен читать ее с диска, а не обращаться к сайту. Сохраните результат в `DataFrame` и покажите `.shape` и `.head()`.
3. Для не менее чем 20 публикаций перейдите по ссылке и извлеките текст. Учтите, что часть страниц может иметь другую разметку, например реклама или анонсы. Такие страницы не должны останавливать выполнение кода. Используйте `.get()` и проверки на `None`.
4. Подготовьте данные к анализу. Покажите результат до и после очистки.
5. Составьте семантический словарь минимум из 40 слов - по 20 слов для положительной и отрицательной тональности. Кратко объясните, почему выбранные слова подходят для вашего источника. Нормируйте показатель тональности на длину текста. Сохраните результат в `panel.csv`.

6. * На одной временной оси постройте два графика: показатель тональности и цену или доходность выбранной бумаги за тот же период. Опишите, видна ли между ними связь.

Связи может не быть. В этом случае укажите не менее трех возможных причин, например состав словаря, размер выборки, выбор окна агрегирования или запаздывание новости относительно события.

7. В 2-3 абзацах опишите проблемы, возникшие при сборе данных, и способы их решения. Например, публикации могут располагаться на нескольких страницах, даты могут быть записаны в относительном формате, а данные - загружаться только при прокрутке страницы.

Если данные видны в инспекторе, но отсутствуют в исходном коде страницы `Ctrl+U`, покажите это на примере. Решать проблему динамической загрузки в рамках задания не требуется - соответствующие инструменты будут рассмотрены на следующих занятиях.