|  |  |  |
| --- | --- | --- |
|  |  |  |
| **Проектная работа**  **ДПО**  **«Компьютерная лингвистика»**  **Pump&Dump** | | |
|  |  |  |
|  |  |  |
| Высшая школа экономики | 2024 г. | Алексей Шкунов |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | |  | | |  |  |
| **Pump&Dump**, или **памп и дамп** (с [англ.](https://ru.wikipedia.org/wiki/%D0%90%D0%BD%D0%B3%D0%BB%D0%B8%D0%B9%D1%81%D0%BA%D0%B8%D0%B9_%D1%8F%D0%B7%D1%8B%D0%BA) — «накачать и сбросить») — схема манипулятивного повышения курса на рынках ценных бумаг, криптовалют или иных подобных активов с последующим их сильным обвалом. В основу заложена попытка увеличить стоимость при помощи ложных, ничем не обоснованных рекомендаций. | | **Невидимая рука рынка: как анонимы в Telegram научились двигать котировки и почему это опасно**  Вечером в пятницу 25 октября 2019 года телеграм-канал MarketTwits, на который подписано 44 000 человек, опубликовал «молнию» с новостью о том, что нефтяная компания «Сургутнефтегаз» хочет купить пакет акций «Лукойла». Пост появился в 18:53 по Москве, в нем стояла ссылка на источники информационного агентства «Интерфакс».  Вечером того же дня опровергать слухи о покупке был вынужден лично глава «Лукойла» Вагит Алекперов.  В течение трех дней перед этим акции «Сургутнефтегаза» без видимых причин выросли на 26%.[[1]](#footnote-1) | | | **«Ревущий котенок» и поход инвесторов любителей на GameStop**  Блогер Кит Билл, известный под ником «Ревущий котенок», инвестировавший в GameStop $56 тыс. и имевший в соцсетях аудиторию 50 тыс. человек, смог объединить инвесторов-любителей из соцсетей и убедил их вложиться в акции GameStop, чтобы не дать бумагам упасть и заработать на росте самим. В итоге 21 января 2021-го акции GameStop за день взлетели на 360% и продолжали расти до 1000%. Крупные хедж-фонды, которые ставили на их падение, потеряли, по меньшей мере, $5 млрд, а «Ревущий котенок» заработал 28 500% - $16 млн.[[2]](#footnote-2) | **Число физлиц с брокерскими счетами на "Мосбирже" в 2023 году выросло на 6,7 млн.**  Количество физлиц, имеющих брокерские счета на "Московской бирже", по итогам 2023 года увеличилось на 6,7 млн и достигло 29,7 млн человек, сообщается в пресс-релизе биржи.  Сделки на фондовом рынке "Мосбиржи" в 2023 году в среднем ежемесячно заключали 3 млн человек (2,2 млн в 2022 году).  Частные инвесторы в 2023 году вложили в ценные бумаги на "Мосбирже" более 1,1 трлн рублей, в том числе 178,1 млрд рублей в акции, 215,4 млрд рублей в биржевые фонды, 715 млрд рублей в облигации. |
|  | | | | | | |
| **Банк России предупреждает о рисках Pump & Dump[[3]](#footnote-3)** | | | | | **По исследованию Хазиева Г.А. (Академия управления МВД РФ) в 2020-м году было выявлено 73 кейса.[[4]](#footnote-4)** | |
| Банк России в ходе мониторинга организованных торгов выявляет публикации в отношении эмитентов или их ценных бумаг в социальных сетях и мессенджерах. Регулятор считает такую практику недобросовестной. | | | | | С активным развитием российского фондового рынка появляются новые виды недобросовестных практик и форм противоправного поведения его участников. | |
| **Сервисы мониторинга ТГ-каналов** | | | | | | |
| https://tgstat.ru/ | Каталог Telegram-каналов и чатов | | 2 190 руб. в месяц |  | | |
| https://parsum.ru/ | Cервис сарафанно-репутационного взаимодействия | | 42₽ за чат в неделю |  | | |
| https://telemetr.me/ | Сервис аналитики Телеграм-каналов | | 1 950 ₽ рублей в месяц |  | | |

**Назначение**

Автоматизированная обработка на основе правил новостей Telegram и формирование отчетности для экспертной оценки при выявлении схем Pump&Dump.

**Ключевые термины**

1. Ти́кер (ticker symbol) — краткое название в [биржевой](https://ru.wikipedia.org/wiki/%D0%91%D0%B8%D1%80%D0%B6%D0%B0) информации [котируемых](https://ru.wikipedia.org/wiki/%D0%9A%D0%BE%D1%82%D0%B8%D1%80%D0%BE%D0%B2%D0%BA%D0%B0) инструментов ([акций](https://ru.wikipedia.org/wiki/%D0%90%D0%BA%D1%86%D0%B8%D1%8F_(%D1%84%D0%B8%D0%BD%D0%B0%D0%BD%D1%81%D1%8B)), [облигаций](https://ru.wikipedia.org/wiki/%D0%9E%D0%B1%D0%BB%D0%B8%D0%B3%D0%B0%D1%86%D0%B8%D1%8F), [индексов](https://ru.wikipedia.org/wiki/%D0%A4%D0%BE%D0%BD%D0%B4%D0%BE%D0%B2%D1%8B%D0%B9_%D0%B8%D0%BD%D0%B4%D0%B5%D0%BA%D1%81))
2. Листинг — предоставление доступа к покупке/продаже ценных бумаг компании на бирже.
3. Эшелон или уровень листинга - группа акций, объединённых общими свойствами. Это условный показатель, где крупные компании попадают в верхние, относительно небольшие — в нижние.

**Характеристики**

На момент формирования Отчета по проекту корпус новостей содержит данные по 30 публичным каналам объемом 47 000 постов, 3 150 835 токена, (105 Мб).

**Структура**

1. Данные Московской биржи
2. Новости Telegram
3. Лингвистический компонент
4. Статистическая обработка
5. Отчеты

**Направления развития**

1. Мониторинг по набору тикеров
2. Поиск по новостям
3. Дискурс-анализ
4. Анализ финансовой информации
5. Визуальное представление данных
6. Машинное обучение
7. Бот «Карманный справочник»

|  |  |  |
| --- | --- | --- |
| **MOEX\_load** |  |  |
|  |  |  |
| **Библиотеки** | **Источники данных** | **Функции** |
| requests  pandas  BeautifulSoup  re  pymorphy3 | 1. Московская биржа. Листинг ценных бумаг (csv) 2. Московская биржа. Ежедневные данные по тикерам 3. Собственный корпус акций российских компаний 4. Собственный список стоп-слов 1 | 1. Загрузка актуального списка российских акций с нормализацией текста 2. Сверка с собственным корпусом для поиска новых эмитентов по ИНН 3. Загрузка дополнительной информации по эмитенту (полное и сокращенное наименование) 4. Нормализация наименования (наличие гласных, вхождение краткого в полное, стоп-слова) 5. Склонение наименования 6. Обновление корпуса акций |
|  |  |  |
| 02.02.2024 00:00:00,1187,Первый уровень,529,Акции,Акция привилегированная,акции привилегированные,SBERP,RU0009029557,20301481B,11.07.2007 00:00:00,"Публичное акционерное общество ""Сбербанк России""",7707083893,3,Рубль,1000000000,16.07.2007 00:00:00,,,,,+,,,,,,,,,"09.06.2014 Включение в Первый уровень 16.07.2007 Включение в Котировальный список ""А"" первого уровня",,,,,РПС: Акции - адрес. РЕПО в ин. валюте: USD РЕПО c облигациями Т+ Неполные лоты (акции) РЕПО с ЦК 1 день РЕПО с ЦК Адресное Т+ Акции и ДР РПС с ЦК: Акции и ДР - адрес. РЕПО в ин. валюте: EUR Поставка по СК (акции) РЕПО с ЦК 1 день (расч. в USD) РЕПО с ЦК 1 день (расч. в EUR) РЕПО с ЦК 7 дн. РЕПО с ЦК адр. (расч. в USD) РЕПО с ЦК адр.(расч. в EUR) РЕПО с ЦК 7 дней c расчетами в долларах США РЕПО с ЦК 7 дней c расчетами в евро Продажа обеспечения бирж.рынок - безадрес. РЕПО с ЦК адр. (расч. в CNY) РЕПО с ЦК 1 день (расч. в CNY) РЕПО в ин. валюте: CNY Урегулирование РЕПО с ЦК Нерезиденты - безадрес. РПС с ЦК Нерезиденты: Акции и ДР - адрес. Урегулирование с ЦК Нерезиденты - безадрес. РЕПО с ЦК 7 дн. (расч. в CNY) ОТС: Акции РПС с ЦК - адресные,"Акция обыкновенная | | |
| 7707083893|сбербанк россии|  sber,sberp|  сбербанк|  сбербанк|  сбербанке,сбербанках,сбербанками,сбербанком,сбербанков,сбербанка,сбербанк,сбербанкам,сбербанки,сбербанку|  сбер,сбера,сберу,сбером,сбере | | |
|  |  |  |

|  |  |  |
| --- | --- | --- |
| **TG\_load** |  |  |
|  |  |  |
| **Библиотеки** | **Источники данных** | **Функции** |
| pandas  requests  re  BeautifulSoup  fake\_useragent  pytz  datetime | 1. Главная страница канала в TG   https://t.me/s/" + channel   1. Страница новости   "https://t.me/" + channel + "/" + str(post) + "?embed=1"   1. Настраиваемый список обрабатываемых каналов | 1. Загрузка списка каналов 2. Парсинг главной страницы канала (крайняя новость, количество подписчиков) 3. Создание списка новостей для загрузки 4. Парсинг страницы с новостью 5. Нормализация даты-времени (+часовой пояс) 6. Загрузка текста новости с предобработкой (регулярные выражения) 7. Обновление корпусов новостей |
|  |  |  |
|  |  | tinkoff\_invest\_official|5817|02-02-2024|10:48|​​ Самолет: ужесточения условий по ипотеке — не помеха Лидер среди российских девелоперов по площади строящихся объектов — компания Самолет — поделилась операционными результатами и ключевыми финансовыми показателями за четвертый квартал и двенадцать месяцев 2023 года. |
|  |  |  |
|  | | |

|  |  |  |  |
| --- | --- | --- | --- |
| **NEWS\_post** |  | |  |
|  |  | |  |
| **Библиотеки** | **Источники данных** | | **Функции** |
| pandas  nltk  textblob | 1. Собственный список стоп-слов 2 2. Собственный корпус акций российских компаний | | 1. Предобработка текста (стоп-слова NLTK + свой список, пунктуация) 2. Токенизация 3. Дедупликация 4. Удаление рекламных постов 5. Создание дополнительных служебных полей 6. Тикеризация 7. Сентимент-анализ |
|  |  | |  |
| tinkoff\_invest\_official|5817|  02-02-2024|10:48|​​  Самолет: ужесточения условий по ипотеке — не помеха Лидер среди российских девелоперов по площади строящихся объектов — компания Самолет — поделилась операционными результатами и ключевыми финансовыми показателями за четвертый квартал и двенадцать месяцев 2023 года. | tinkoff\_invest\_official|5817|  2024-02-02|10:48|​​  Самолет: ужесточения условий по ипотеке — не помеха Лидер среди российских девелоперов по площади строящихся объектов — компания Самолет — поделилась операционными результатами и ключевыми финансовыми показателями за четвертый квартал и двенадцать месяцев 2023 года.  |2024|5|  самолет,ужесточения,условий,ипотеке,помеха,  лидер,среди,российских,девелоперов,  площади,строящихся,объектов,  компания,самолет,поделилась,операционными,результатами,  ключевыми,финансовыми,показателями,  четвертый,квартал,двенадцать,месяцев|  **SMLT|**  neutral | | 9731004688|группа компаний самолет|smlt|самолет|  самолет|самолёта,самолёте,самолёту,самолётом,самолётами,  самолёт,самолёты,самолётов,самолётах,самолётам| |
|  |  | |  |
| **Не всегда все гладко** - 9703103060|вуш холдинг|wush|вушхолднг|||вуш,вуша,вуше | | | |
| import spacy  nlp = spacy.load("ru\_core\_news\_lg")  tokens = nlp(txt)  if token.pos\_ == "PROPN":  news\_names.add(token.text.lower())  news\_names.add(token.lemma\_.lower()) | | tinkoff\_invest\_official|4944|2023-10-18|10:22|Whoosh поделится дивидендами Так ждали и вот они приехали Совет директоров Whoosh рекомендовал выплатить дивиденды размером 10,25 рублей на акцию. Дивидендная доходность — около 4,6%. Компания, специализирующаяся на шеринге самокатов, поделится с акционерами прибылью впервые с выхода на биржу. |wush | |

|  |  |  |
| --- | --- | --- |
| **NEWS\_analyze** |  |  |
|  |  |  |
| **Библиотеки** | **Источники данных** | **Функции** |
| pandas  collections  datetime  nltk | 1. Корпус новостей 2. Список каналов | 1. Количественная оценка (общее количество, доступный период, средняя длина поста, среднее количество постов в день, количество постов в периоде, количество подписчиков) 2. Синтаксический состав (частые слова, биграммы, униграммы) 3. Информативность (топ тикеров, распределение акций по эшелонам, количество постов с одним тикером, количество постов с несколькими тикерами, «настроение» новостей) 4. Формирование сводного отчета |
|  |  |  |
|  | | |
|  |  |  |
|  | | |
|  | | |
|  | | |

|  |  |  |
| --- | --- | --- |
| **CORPS\_analyze** |  |  |
|  |  |  |
| **Библиотеки** | **Источники данных** | **Функции** |
| pandas | 1. Московская биржа. Листинг ценных бумаг 2. Собственный корпус акций российских компаний 3. Корпус новостей ТГ | 1. Создание корпуса новостей всех каналов за определенный период 2. Обработка правил выбора 3. Группировка по набору правил 4. Формирование сводного количественного отчета по все каналам 5. Формирование отчета по тикерам 6. Формирование отчета с текстом новостей |
|  |  |  |
|  | | |
|  |  |  |
|  | | |
|  |  |  |

|  |  |  |
| --- | --- | --- |
|  | | |
|  |  |  |
|  |  | |  |  | | --- | --- | | TICKERS | NORM\_NAME | | akrn | акрон | | alrs | алроса | | arsa | арсагера | | abio | артген | | acko | аско | | astr | астра | | aflt | аэрофлот | | bisvp | башинформсвязь | | blng | белон | | brzl | бурятзолото | | vgsb,vgsbp | волгоградэнергосбыт | |

**Проект в рамках курса «R и статистика»**

|  |  |
| --- | --- |
|  |  |
|  |

**Итоговая работа**

**«Лингвистические компоненты мониторинга фондового рынка»**

Корпус новостей содержит данные по 58 публичным ТГ - каналам объемом более 133 000 постов (332 Мб).

Настроены процедуры регулярного обновление данных:

1. 3 раза в день новые посты по списку каналов;
2. 1 раз в день актуальные котировки акций;
3. 1 раз в день список эмитентов МосБиржи;
4. 1 раз в неделю значение индекса IMOEX[[5]](#footnote-5).

Для обработки в рамках Проекта взята часть корпуса за период январь 2023 – май 2024 объемом 15 061 постов из 7 каналов с тикерами по 3437 записей (23%).

Ключевые задачи Итогового проекта:

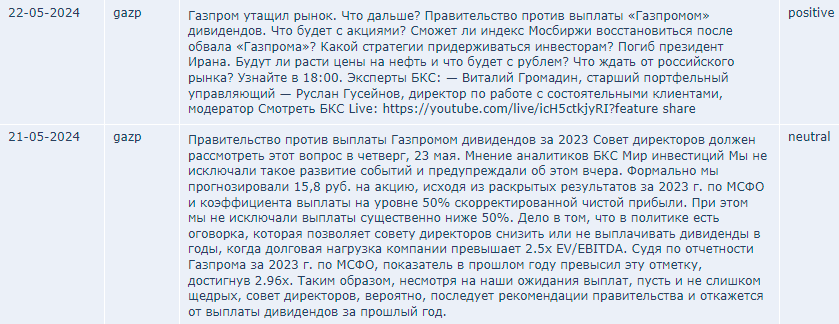
1. Исследование применения модели мультиклассовой классификации для прогнозирования изменения индекса IMOEX с различными типами векторайзеров;
2. Исследование эффективности применения моделей суммаризации текстов;
3. Разработка шаблона веб-приложения для визуализации применяемых методов выявления аномалий.

## Классификация

|  |  |
| --- | --- |
| **Препроцессинг**   * Лемматизация * Расчет изменения цен за 3 дня: вчера, сегодня, завтра. * Стратификация на 3 диапазона (негативный -0,6 нейтральный 0,6 позитивный) для каждого из 3 дней |  |
| **Моделирование**   * Разделение выборки на обучающую и тестовую в соотношении 80/20 * Зависимая переменная – 9 столбцов с индикаторными переменными (get\_dummies) * Векторизация Tf-Idf. Также был эксперимент с n—граммами BOW. |  |

## Анализ тональности

from textblob import TextBlob[[6]](#footnote-6)



## Суммаризация

|  |  |
| --- | --- |
| Summary[[7]](#footnote-7) | Summary\_2[[8]](#footnote-8) |
| model\_name = "IlyaGusev/mbart\_ru\_sum\_gazeta"  tokenizer = MBartTokenizer.from\_pretrained(model\_name) | g = nx.Graph()      pr = nx.pagerank(g) |
|  | |

## Веб-интерфейс

Реализован с помощью библиотеки Streamlit[[9]](#footnote-9).

Предполагает следующие разделы:

1. Общий обзор фондового рынка и суммаризация новостей
2. Мониторинг собственного портфеля акций (динамика цен и новостной фон с сентимент-анализом), «тепловая карта»
3. Статистические показатели корпуса новостей с оценкой достоверности публикуемой информации
4. Выявление схем Pump & Dump
5. Полнотекстовый поиск по корпусу новостей
6. Торговые идеи (моделирование и консеунс-прогнозы)

|  |  |  |
| --- | --- | --- |
|  |  |  |

1. https://www.forbes.ru/finansy-i-investicii/392643-nevidimaya-ruka-rynka-kak-anonimy-v-telegram-nauchilis-dvigat-kotirovki [↑](#footnote-ref-1)
2. https://www.finam.ru/publications/item/durnye-dengi-kak-istoriya-gamestop-popala-na-ekrany-i-chemu-mozhet-nauchit-investorov-20231006-1732/ [↑](#footnote-ref-2)
3. https://www.cbr.ru/inside/inside\_practices/pump\_dump/ [↑](#footnote-ref-3)
4. https://cyberleninka.ru/article/n/vliyanie-publikatsiy-telegram-kanalov-na-dinamiku-aktsiy-rossiyskogo-fondovogo-rynka?ysclid=lryn684fs4339019042 [↑](#footnote-ref-4)
5. https://ru.tradingview.com/symbols/MOEX-IMOEX/ [↑](#footnote-ref-5)
6. https://textblob.readthedocs.io/en/dev/ [↑](#footnote-ref-6)
7. https://huggingface.co/IlyaGusev/mbart\_ru\_sum\_gazeta [↑](#footnote-ref-7)
8. https://networkx.org/documentation/stable/reference/algorithms/generated/networkx.algorithms.link\_analysis.pagerank\_alg.pagerank.html [↑](#footnote-ref-8)
9. https://streamlit.io/ [↑](#footnote-ref-9)