

В рамках курса «Анализ данных в бизнесе»
Проект на тему

«Анализ событий для системы раннего предупреждения»

Выполнили:
Артем Буров
Эльдар Дамиров
Мария Поклонская

БИЗНЕС-ЗАДАЧА

Создать сервис, который по запросу
о любой компании выдает
структурированный анализ самой важной
информации об этой компании
из разных источников

ЦЕЛИ И ЗАДАЧИ ПРОЕКТА

- 1 Выбрать компанию и собрать о ней данные из открытых источников
- 2 Выделить типовые события, категории новостей.
- 3 Экспертно определить значимость каждого события и разработать систему классификации новостей.
- 4 Разработать систему автоматического определения субъекта и объекта события.
- 5 Разработать систему извлечения значимых фактов из новостей.

АНАЛОГИЧНЫЕ ПРОДУКТЫ



Агрегатор новостей с базовой информации событиях или персонах и сопутствующими справочными материалами, включая статистические данные и аналитику, ссылки на профессиональные исследования, статьи и обзоры

«газпром» 3 542

Компании От 18.05.2022 – До

ООО "Газпром энергохолдинг" ПАО "ГАЗПРОМ НЕФТЬ" ПАО "ГАЗПРОМ" ООО "Газпром межрегионгаз" ООО "Газпром экспорт"

Победители конкурса профмастерства в «Газпромнефть-Хантос» будут защищать честь предприятия на корпорат...
6 минут назад

Промышленность
Германия объяснила прекращение подачи российского газа во Францию
7 минут назад
Компании • География 59 сек.

Новак прогнозировал проблемы в Европе осенью-зимой из-за ситуации с газом
12 минут назад

АНАЛОГИЧНЫЕ ПРОДУКТЫ



Программно-аппаратный комплекс мониторинга и анализа данных с аналитическим центром изучения социальных медиа. Закрытый доступ к аналитике

Виды анализа



Поток текстовых публикаций



Поток изображений



Данные из профилей аккаунтов



Информация о действиях аккаунта (лайки, репосты, неявные связи и т. д.)



Характер развития инфоповода



Медийная сила аккаунта

Источники данных

Платформа хранит данные из всех социальных сервисов и СМИ русскоязычного Интернета

АНАЛОГИЧНЫЕ ПРОДУКТЫ



Сервис для поиска нужной информации, отсортированной по авторитетности СМИ. Также определяет риск-факторы. Закрытый доступ

Мониторинг СМИ и коммуникация с журналистами

Весь инструментарий для PR:

- база СМИ;
- ежедневный мониторинг;
- пресс-клиплинги по инфоповодам;
- медиа-аналитика;
- оперативные кризисные оповещения.

Проверка контрагентов и установление деловой репутации

- поиск упоминаний контрагентов в медиа по риск-факторам;
- риски по портфелю;
- расследования;
- экспресс-аудит репутации;
- справки для регламентов по должностной осмотрительности.

ПРОТОТИП



| все материалы проекта можно найти на гитхабе

ЭТАП 1: СБОР И ПАРСИНГ ДАННЫХ

| ПАРСИНГ ДАННЫХ ПРИ ПОМОЩИ БИБЛИОТЕКИ *SELENIUM*

- Программа в автоматическом режиме открывает страницу с новостями
- В базу данных копирует заголовки и тексты новостей (в два различных столбца)
- Для получения данных со всех источников - достаточно один раз запустить программу, она выдаст по каждому источнику таблицу с данными
- Чтобы получить новые данные - удалите старые таблицы и перезапустите программу
- Процесс парсинга занимает около 4 минут

ЭТАП 1: СБОР И ПАРСИНГ ДАННЫХ

| ПАРСИНГ ДАННЫХ ПРИ ПОМОЩИ БИБЛИОТЕКИ *SELENIUM*

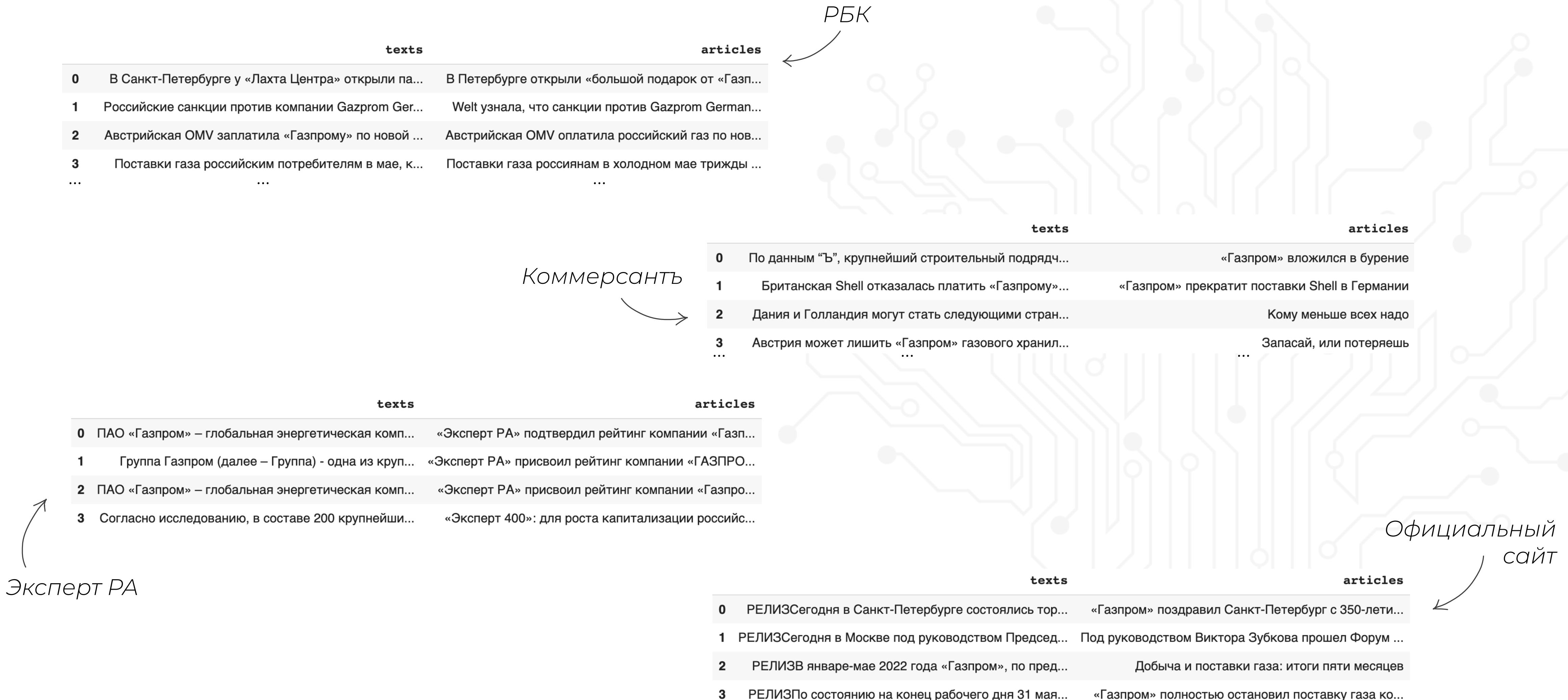
РБК: 40 последних новостей

Газпром: все события за последний месяц

Коммерсант: 40 последних новостей

Эксперт РА: 4 последних обновления рейтинга

ЭТАП 1: РЕЗУЛЬТАТЫ РАБОТЫ



ЭТАП 2: КАТЕГОРИИ СОБЫТИЙ

- **добыча газа** «Газпром» планирует в 2022 году снизить добычу газа на 4%
- **поставки газа** «Газпром» подтвердил прекращение поставок газа в Финляндию
- **биржа** Спекулянты отыгрались на «Газпроме»
- **инвестиции** Для роста капитализации российских компаний необходимы сильные внутренние инвесторы
- **рейтинг** «Эксперт РА» подтвердил рейтинг компании «Газпром» на уровне А++.GQ
- **санкции** Россия ввела санкции против бывших «дочек» «Газпрома»
- **события** «Газпром» поздравил Санкт-Петербург с 350-летием со дня рождения Петра I
- **совет директоров** Шрёдер отклонил предложение войти в совет директоров «Газпрома»

ЭТАП 2: КАТЕГОРИИ СОБЫТИЙ

РАЗДЕЛЕНИЕ НОВОСТЕЙ НА КАТЕГОРИИ ПРИ ПОМОЩИ TF-IDF ВЕКТОРОВ ТЕКСТОВ И АЛГОРИТМА КЛАСТЕРИЗАЦИИ K-MEANS

Номер кластера: 0
['рейтинг' , 'рейтинг']

 истинная разметка (категории) новостей, которые попали в кластеры

Номер кластера: 3
['санкции', 'совет директоров', 'совет директоров', 'совет директоров', 'события']

ЭТАП 2: КАТЕГОРИИ СОБЫТИЙ

РАЗДЕЛЕНИЕ НОВОСТЕЙ НА КАТЕГОРИИ ПРИ ПОМОЩИ TF-IDF ВЕКТОРОВ ТЕКСТОВ И АЛГОРИТМА КЛАСТЕРИЗАЦИИ K-MEANS

Номер кластера: 0

['рейтинг' , 'рейтинг']

рейтинг

Номер кластера: 1 **ПОСТАВКИ ГАЗА**

['события', 'поставки газа', 'санкции', 'санкции', 'санкции', 'поставки газа', 'оплата газа', 'поставки газа', 'поставки газа', 'поставки газа', 'поставки газа', 'санкции', 'поставки газа']

Номер кластера: 2 **финансы**

Номер кластера: 3 **события**

['санкции', 'совет директоров', 'совет директоров', 'совет директоров', 'события']

события

Номер кластера: 4 санкции

ЭТАП 2: КАТЕГОРИИ СОБЫТИЙ

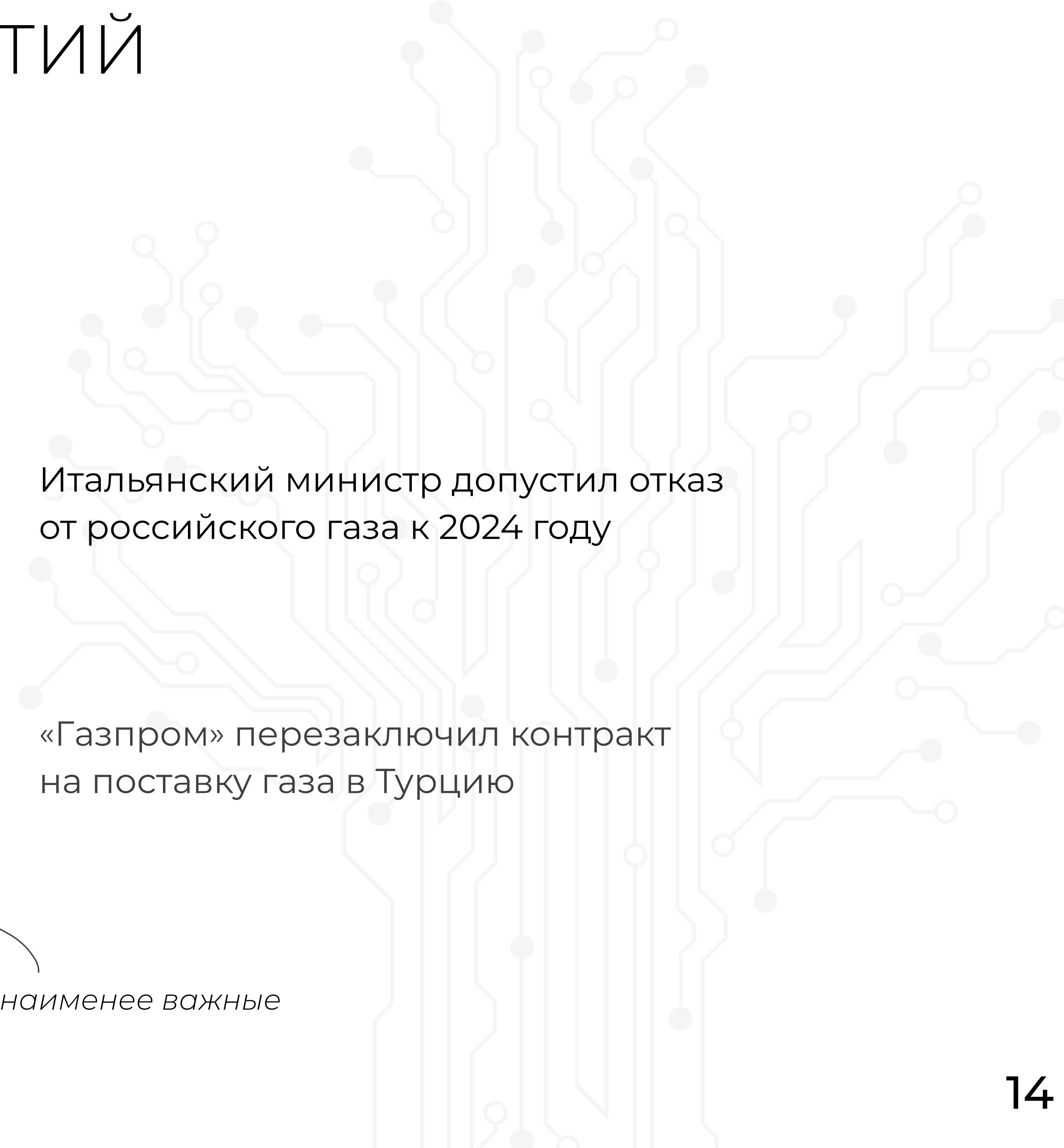
РАЗДЕЛЕНИЕ НОВОСТЕЙ НА КАТЕГОРИИ ПРИ ПОМОЩИ TF-IDF ВЕКТОРОВ ТЕКСТОВ И АЛГОРИТМА КЛАСТЕРИЗАЦИИ K-MEANS

	Article	Cluster	True category
0	«Эксперт 400»: для роста капитализациироссийс...	финансы	инвестиции
1	«Эксперт РА» присвоил рейтинг компании «Газпро...	рейтинг	рейтинг
2	«Эксперт РА» присвоил рейтинг компании «ГАЗПРО...	финансы	рейтинг
3	Европа «перебрала» с запасами	финансы	добыча газа
4	«Газпром» восьмой день не бронирует мощности г...	финансы	поставки газа
...

ВЫВОДЫ

- Алгоритм работает близко к истинной разметке и правильно разделил большую часть новостей по кластерам
- Неправильно разделились новости, находившиеся на стыке категорий, особенно это заметно в путанице между категориями "санкции" и "поставки газа"
- Второй кластер по сути можно объединить как "финансы". Если вчитаться в соответствующие новости, то можно заметить, что они относятся к бирже, инвестициям, оплате газа: всему, что можно отнести к финансовой части.

ЭТАП 2: ЗНАЧИМОСТЬ СОБЫТИЙ

- 
- наиболее важные
- 1 «Газпром»: экспорт газа в Китай с начала года вырос почти на 60%
- 2 Итальянский министр допустил отказ от российского газа к 2024 году
- 3 «Газпром» собирает газ севера Ямала
- 4 «Газпром» перезаключил контракт на поставку газа в Турцию
- 5 В Петербурге открыли «большой подарок от «Газпрома»
- наименее важные
- 1 «Газпром»: экспорт газа в Китай с начала года вырос почти на 60%
 - 2 Итальянский министр допустил отказ от российского газа к 2024 году
 - 3 «Газпром» собирает газ севера Ямала
 - 4 «Газпром» перезаключил контракт на поставку газа в Турцию
 - 5 В Петербурге открыли «большой подарок от «Газпрома»

ЭТАП 2: РЕЗУЛЬТАТЫ РАБОТЫ

	Источник	Дата	Значимость	Категория	Заголовок	Новость
0	Эксперт РА	21.10.2014	4	инвестиции	«Эксперт 400»: для роста капитализации российс...	Согласно исследованию, в составе 200 крупнейши...
1	Эксперт РА	19.02.21	3	рейтинг	«Эксперт РА» присвоил рейтинг компании «Газпро...	ПАО «Газпром» – глобальная энергетическая комп...
2	Эксперт РА	18.06.21	3	рейтинг	«Эксперт РА» присвоил рейтинг компании «ГАЗПРО...	Группа Газпром (далее – Группа) - одна из круп...
3	Коммерсант	27.12.21	4	добыча газа	Европа «перебрала» с запасами	Последнюю неделю европейский спотовый рынок бо...
4	Коммерсант	28.12.21	5	поставки газа	«Газпром» восьмой день не бронирует мощности г...	«Газпром» на дополнительной ночной сессии 28 д...
...
84	РБК	01.06.22	2	поставки газа	«Газпром» подтвердил остановку поставок датско...	«Газпром» полностью остановил поставку газа да...
85	РБК	01.06.22	3	поставки газа	В Молдавии сообщили о снижении цены на российс...	Закупочная цена российского природного газа дл...
86	Коммерсант	06.06.22	4	биржа	«Газпром» вложился в бурение	По данным “Ъ”, крупнейший строительный подрядч...
87	Газпром	09.06.22	5	события	«Газпром» поздравил Санкт-Петербург с 350-лети...	РЕЛИЗСегодня в Санкт-Петербурге состоялись тор...
88	Газпром	09.06.22	5	события	Под руководством Виктора Зубкова прошел Форум ...	РЕЛИЗСегодня в Москве под руководством Председ...

ЭТАП 3: СУММАРИЗАЦИЯ ИНФОРМАЦИИ

| ИНИЦИАЛИЗАЦИЯ *TRANSFORMER-BASED PRE-TRAIN* МОДЕЛИ СУММАРИЗАЦИИ, АДАПТИРОВАННОЙ НА КОРПУСЕ ТЕКСТОВ НОВОСТЕЙ.

«Газстройпром» профинансировал покупку третьей стороной «Газпром бурения» почти на 68 млрд руб. По данным «Ъ», крупнейший строительный подрядчик «Газпрома» помог расплатиться за актив с Аркадием Ротенбергом. Смена владельцев потребовалась после того, как бизнесмены попали под жесткие санкции США, тогда как их компании получали подряды на крупнейшие стройки «Газпрома», в том числе в Европе.



СУММАРИЗАЦИЯ ТЕКСТА ИЗ ОДНОЙ НОВОСТИ

ЭТАП 3: СУММАРИЗАЦИЯ ИНФОРМАЦИИ

ИНИЦИАЛИЗАЦИЯ TRANSFORMER-BASED PRE-TRAIN МОДЕЛИ СУММАРИЗАЦИИ, АДАПТИРОВАННОЙ НА КОРПУСЕ ТЕКСТОВ НОВОСТЕЙ.

«Газстройпром» профинансирует покупку третьей стороной «Газпром бурения» почти на 68 млрд руб. Смена владельцев потребовалась после того, как бизнесмены попали под жесткие санкции США, тогда как их компании получали подряды на крупнейшие стройки «Газпрома», в том числе в Европе.

«Газпром» не сможет использовать трубопровод «Ямал—Европа» для прокачки газа в Европу из-за проблем с поставками газа через Украину. На фоне стремительного ослабления курса рубля инвесторы не исключают срабатывания большого количества стоп-заявок. Глава «Роснефти» Игорь Сечин предлагает установить для монополии минимальный и максимальный объем продаж газа на бирже СПБМТСБ.

«Газпром» прекратил поставки газа реверсом в Польшу, несмотря на то, что Германия прекратила физический реверс по Ямалу—Европе. Это произошло из-за появившихся заявок потребителей Италии и Франции, которые обеспечиваются по польскому коридору. Эксперты полагают, что раскол на европейском рынке газа «уже обозначился» и будет нарастать по мере поляризации позиций.

Российский «Газпром» готовится выйти из капитала немецкой компании Gazprom Germania, которая с 1 апреля перестала контролироваться российским «Газпромом». Ранее сообщалось, что власти Германии анализируют возможность национализации Gazprom Germania для того, чтобы предотвратить риски массового нарушения энергоснабжения в случае возникновения трудностей у этих компаний.

«Газпром» не стал бронировать дополнительные твердые мощности для транзита газа через Украину и Польшу на второй и третий кварталы 2022 года. Накануне зимы объемы закачки в европейские ПХГ находились на критически низких уровнях, сейчас этот показатель опустился до 36,2%. В «Роснефти» считают, что это «дополнительный вклад в рост цен».

«Газпром» подал иск в суд на польскую госкомпанию PGNiG SA и их совместное предприятие Europol Gaz с требованием о признании права на дивиденды и участие в управлении СП. Польские СМИ сообщали, что компания блокирует попытки «Газпрома» получить дивиденды Europol Gaz на том основании, что это позволит российской компании финансировать строительство «Северного потока-2».

ЭТАП 3: СУММАРИЗАЦИЯ ИНФОРМАЦИИ

ИНИЦИАЛИЗАЦИЯ TRANSFORMER-BASED PRE-TRAIN МОДЕЛИ СУММАРИЗАЦИИ, АДАПТИРОВАННОЙ НА КОРПУСЕ ТЕКСТОВ НОВОСТЕЙ.

В Санкт-Петербурге открыли памятник «Петр I, спасающий утопающих близ Лахты» по мотивам памятника Леопольда Бернштама. Ранее Россия объявила о новом механизме оплаты газа, из-за которого страна лишилась доступа примерно к половине золотовалютных резервов. Ранее Shell заявила о постепенном отказе от закупок российской нефти, а также нефтепродуктов, трубопроводного газа и СПГ еще в начале марта.

«Газпром» приостановил поставки газа в Молдавию из-за неуплаты за апрель. Накануне компания GasTerra объявила, что не будет выполнять требования «Газпрома» о переходе на новую схему оплаты топлива. Эксперты считают, что это может привести к нарушению введенных Евросоюзом санкций, а также повлечь за собой другие финансовые и операционные риски.

Поставки российского газа в Молдавию могут быть прекращены из-за того, что компания GasTerra не будет платить за топливо в рублях. Ранее экс-канцлер Германии Герхард Шрёдер отклонил предложение войти в совет директоров «Газпрома».

В совет директоров «Газпрома» вошли министр сельского хозяйства Дмитрий Патрушев, министр энергетики Николай Шульгинов, председатель правления компании Алексей Миллер, глава Газпромбанка Андрей Акимов, спецпредставитель президента Виктор Зубков, министр промышленности и торговли Денис Мантуров, ректор РГУ им. Губкина Виктор Мартынов и ректор РАНХиГС Владимир May.

«Газпром экспорт» намерен защищать свои интересы в арбитраже с Финляндией, которая не приняла новые российские условия оплаты газа и «передала контракт в арбитраж». Ранее президент Владимир Путин обязал недружественные страны платить за российский газ в рублях, чтобы избежать заморозки и тех денег, что будут получены за его продажу.

В СМИ появилась информация о возможном прекращении поставок газа Финляндии в ответ на ее намерение вступить в НАТО. По словам пресс-секретаря президента России Дмитрия Пескова, эти сообщения — «утка», поскольку «Газпром» неоднократно демонстрировал свою надежность как компания, которая поставляет энергоресурсы в Европу. Ранее сообщалось, что поставки российского газа могут прекратиться в период с 21 по 23 мая, после того как в Хельсинки отказались оплачивать топливо в рублях.

ЭТАП 3: СУММАРИЗАЦИЯ ИНФОРМАЦИИ

ИНИЦИАЛИЗАЦИЯ TRANSFORMER-BASED PRE-TRAIN МОДЕЛИ СУММАРИЗАЦИИ, АДАПТИРОВАННОЙ НА КОРПУСЕ ТЕКСТОВ НОВОСТЕЙ.

В Москве состоялся Форум «Технологическая независимость сегодня — мировое лидерство завтра». На мероприятии приняли участие представители Министерства промышленности и торговли РФ, ПАО «Газпром» и других энергетических компаний, руководители промышленных предприятий — членов Ассоциации производителей оборудования «Новые технологии газовой отрасли».

«Газпром» в январе-мае добыл меньше газа, чем в прошлом году, на 4,8%. Поставки компании из газотранспортной системы на внутренний рынок находятся на уровне прошлого года. На этом фоне поставки газа российским потребителям по Единой системе газоснабжения 22 раза достигали рекордных суточных показателей за 10 лет.

«Газпром экспорт» не получил платеж за поставки газа в апреле от компании Shell Energy Europe Limited в соответствии с Указом Президента РФ №172 от 31.03.2022 г. Платежи за газ, поставляемый с 1 апреля, должны осуществляться в рублях с использованием новых реквизитов, о чем контрагенты были своевременно проинформированы.

«Газпром экспорт» не получил платеж за поставки газа в апреле от компании Orsted Salg & Service в соответствии с Указом Президента РФ №172 от 31.03.2022 г. Платежи за газ, поставляемый с 1 апреля, должны осуществляться в рублях с использованием новых реквизитов, о чем контрагенты были своевременно проинформированы.

анализ на основе официального сайта

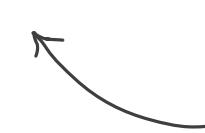
ЭТАП 3: СУММАРИЗАЦИЯ ИНФОРМАЦИИ

ИНИЦИАЛИЗАЦИЯ TRANSFORMER-BASED PRE-TRAIN МОДЕЛИ СУММАРИЗАЦИИ, АДАПТИРОВАННОЙ НА КОРПУСЕ ТЕКСТОВ НОВОСТЕЙ.

Агентство Эксперт РА опубликовало результаты внутреннего аудита ПАО «Газпром» за 2018 год. Эксперты отмечают, что в компании сохраняется высокая степень защиты прав собственников и остальных стейххолдеров, в том числе в АО «СОГАЗ». При этом сдерживающее влияние на уровень рейтинга оказывает невысокая доля независимых директоров – 27,3% (3 независимых директора из 11). Агентство отмечает появление в компании новой должности - старший независимый директор в рамках развития взаимодействия между независимыми директорами, акционерами и Председателем Совета директоров.

Группа Газпром — одна из крупнейших в мире энергетических компаний, крупнейший мировой производитель газа. Головной компанией Группы является ПАО «Газпром» и является одним из крупнейших налогоплательщиков Российской Федерации. Агентство оценивает устойчивость к внешним шокам отрасли Группы как умеренную. Это обусловлено относительно стабильным спросом на углеводороды, низкой себестоимостью добычи, вертикально-интегрированной бизнес-моделью Компании и спецификой ценообразования при реализации газа, обеспечивающей частичное хеджирование ценовой волатильности.

Агентство Эксперт РА оценивает уровень корпоративной этики в «Газпроме» в соответствии с международными стандартами. При этом агентство отмечает, что давление на рейтинговую оценку оказывает отсутствие в Совете директоров комитета по аудиту, который возглавляет независимый член Совета директоров, в отличие от комитета по назначениям и вознаграждениям. При этом на оценку положительно влияет наличие страхования ответственности членов Совета директоров и должностных лиц с высоким размером страховой суммы в АО «СОГАЗ» (ruAAA, Эксперт РА).



анализ на основе сайта Эксперт РА

ЭТАП 3: СУММАРИЗАЦИЯ ИНФОРМАЦИИ

| СРАВНИМ С ИНФОРМАЦИЕЙ, ПОЛУЧАЕМОЙ ИЗ ЗАГОЛОВКОВ

ЗАГОЛОВКИ

Кому меньше всех надо	Запасай, или потеряешь
«Газпром» прекратит поставки Shell в Германии	«Газпром» увеличил поставки газа в Европу
«Газпром» вложился в бурение	Газу перекрывают маршруты
Спекулянты отыгрались на «Газпроме»	

ВЫВОДЫ

- СМИ часто выбирают громкие заголовки, не несущие смысловой нагрузки
- СМИ нередко выбирают заголовки, связанные с новостью лишь косвенно
- Часто, даже при условии корреляции заголовка с реальной повесткой, никакой полезной информации извлечь из него не получится, непонятен контекст новости
- Часть заголовков дублируют друг друга (в нашем решении дублирование сведено к минимуму)

ЭТАП 4: СЕМАНТИЧЕСКАЯ ОКРАСКА НОВОСТЕЙ

| ДЕМОНСТРАЦИЯ РАБОТЫ МОДЕЛИ ОЦЕНКИ СЕМАНТИЧЕСКОЙ ОКРАСКИ ТЕКСТОВ

```
print(predict_sentiment('у Газпрома все плохо'))  
[2]  
  
print(predict_sentiment('у Газпрома все хорошо'))  
[1]  
  
print(predict_sentiment('у Газпрома все нормально'))  
[0]  
  
predict_sentiment(news_1_summary)  
array([0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,  
      0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0])
```

| ВЫВОДЫ

- Тексты изданий, выбранных для анализа, написаны сухим и деловым языком, для работы с которым модели необходим fine-tuning (адаптация).

ЭТАП 5: ИЗВЛЕЧЕНИЕ СУЩНОСТЕЙ

ДЛЯ ВЫДЕЛЕНИЯ СУЩНОСТЕЙ В ТЕКСТАХ, ВОСПОЛЬЗУЕМСЯ МОДЕЛЬЮ, РЕШАЮЩУЮ ЗАДАЧУ NER (NAMED ENTITY RECOGNITION)

```
[('пётр', 13),  
 ('газпром', 12),  
 ('м', 7),  
 ('петербург', 5),  
 ('александр', 5),  
 ('пао', 4),  
 ('санкт', 3),  
 ('-', 3),  
 ('алексей', 3),  
 ('миллер', 3)]
```

для данных с официального сайта

```
[('группа', 23),  
 ('газпром', 16),  
 ('совет', 15),  
 ('россия', 8),  
 ('европа', 7),  
 ('ао', 6),  
 ('ра', 6),  
 ('правление', 5),  
 ('мсфо', 5),  
 ('нпф', 5)]
```

для данных с сайта Эксперт РА

ПЕРСПЕКТИВЫ РАЗВИТИЯ СЕРВИСА

- | СОВМЕСТИТЬ В ОДИН АЛГОРИТМ ВСЕ ЭТАПЫ РАБОТЫ
- | УЛУЧШИТЬ ПРОЦЕСС ИЗВЛЕЧЕНИЯ СУБЪЕКТА И ОБЪЕКТА СОБЫТИЯ
- | АДАПТИРОВАТЬ МОДЕЛЬ ПОД ДЕЛОВОЙ ФОРМАТ НОВОСТЕЙ ДЛЯ БОЛЕЕ КОРРЕКТНОЙ ОЦЕНКИ СЕМАНТИЧЕСКОЙ ОКРАСКИ ТЕКСТОВ
- | АВТОМАТИЗИРОВАТЬ СЕРВИС ТАКИМ ОБРАЗОМ, ЧТОБЫ ОН ВЫДАВАЛ ПО ЗАПРОСУ СТРУКТУРИРОВАННЫЙ АНАЛИЗ ДЛЯ ЛЮБОЙ КОМПАНИИ

РАСПРЕДЕЛЕНИЕ РАБОТЫ ПО ПРОЕКТУ

ЭТАП 1: СБОР И ПАРСИНГ ДАННЫХ

Артем Буров

ЭТАП 2: РАСПРЕДЕЛЕНИЕ СОБЫТИЙ ПО
КАТЕГОРИЯМ И ЗНАЧИМОСТИ

Мария Поклонская

ЭТАП 3: СУММАРИЗАЦИЯ ИНФОРМАЦИИ,
СЕМАНТИЧЕСКАЯ ОКРАСКА И ВЫДЕЛЕНИЕ СУЩНОСТЕЙ

Эльдар Дамиров