Факультет экономических наук | Москва, 2022

Автоматизация работы с документами: извлечение сущностей и фактов из сообщений о раскрытии

Микаилова Сабина, Хасянова Альфия,
Коробкина Елизавета, Кобцева Анастасия

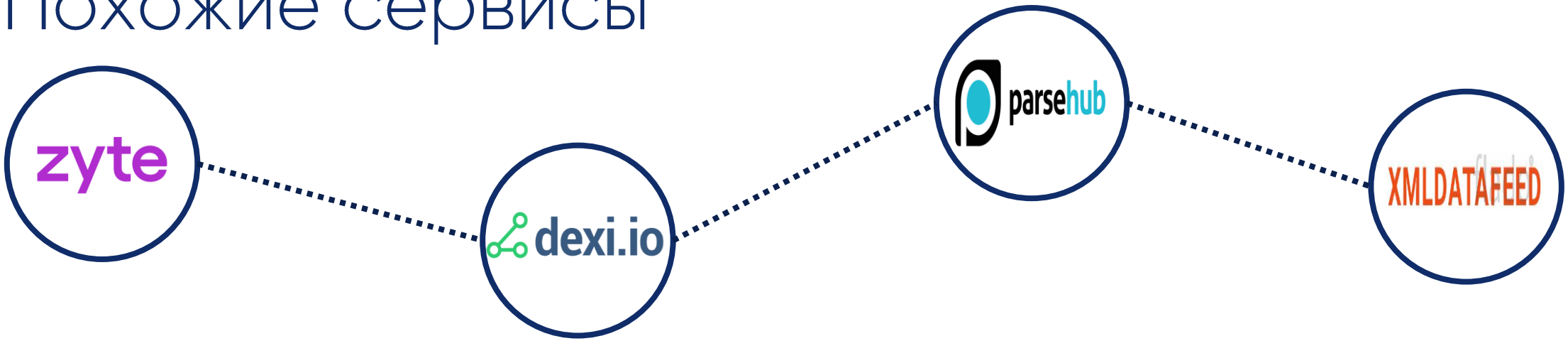
Цели создания сервиса

- Упрощение обработки больших слабо структурированных текстов сообщений о раскрытии
- Автоматизация единовременной обработки больших массивов данных с сообщениями о раскрытии
- Создание узконаправленного инструмента для анализа экономического положения контрагентов
- Сбор данных о контрагентах из надежного источника для принятия управленческих решений и статистики

Решаемые задачи

- Единовременный парсинг большого количества (1200 уникальных записей) сообщений о раскрытии в обход AJAX-скриптов
- Извлечение требуемой информации из решений общих собраний участников (акционеров)
- Скорость обработки данных и экспорт результатов в универсальном формате для пост-обработки (моделями машинного обучения)

Похожие сервисы



- Данные сервисы могут испытывать трудности с AJAX-скриптами и POST-запросами в обход скриптов
- Получаемый результат не структурирован в связи с отсутствием специально разработанного инструмента для e-disclosure.ru
- Специфичный характер информации в сообщениях о раскрытии не позволяет кастомизировать выходной поток информации

Этапы работы сервиса

1. Получение ссылок на тексты решений собраний акционеров POST-запросом
2. Сбор текстов сообщений о раскрытии и предобработка
3. Извлечение сущностей и фактов из текстов
4. Экспорт данных для последующей пост-обработки

Получение ссылок на тексты сообщений

- Было сделано:
  - Подбор параметров POST-запроса для единовременного получения 1200 карточек сообщений о раскрытии
  - Парсинг карточек с выделением даты регистрации сообщения, наименования эмитента и ссылки на текст сообщения
- Было получено:
  - Сводная таблица: каждой строке соответствует одно сообщение о раскрытии с извлеченными атрибутами

Сбор текстов сообщений о раскрытии

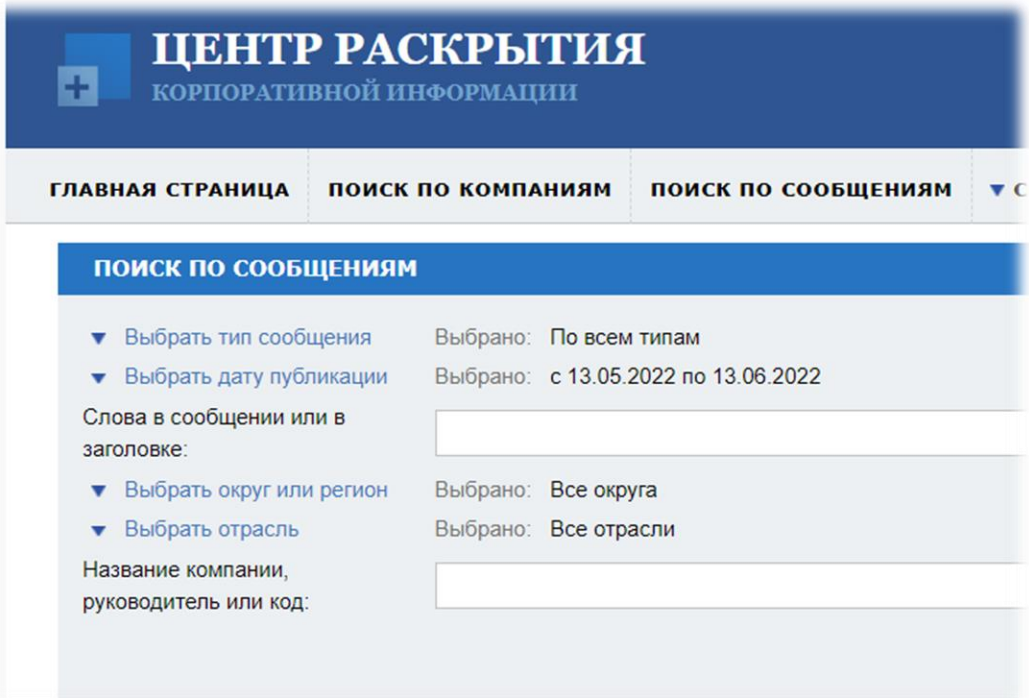
- Было сделано:
  - По каждой ссылке был получен текст сообщения о раскрытии
  - Фильтрация решений общих собраний акционеров (участников)
  - Токенизация текстов с очисткой от неинформативных токенов
- Было получено:
  - Сводная таблица дополнена предобработанными потенциально информативными и отфильтрованными токенами

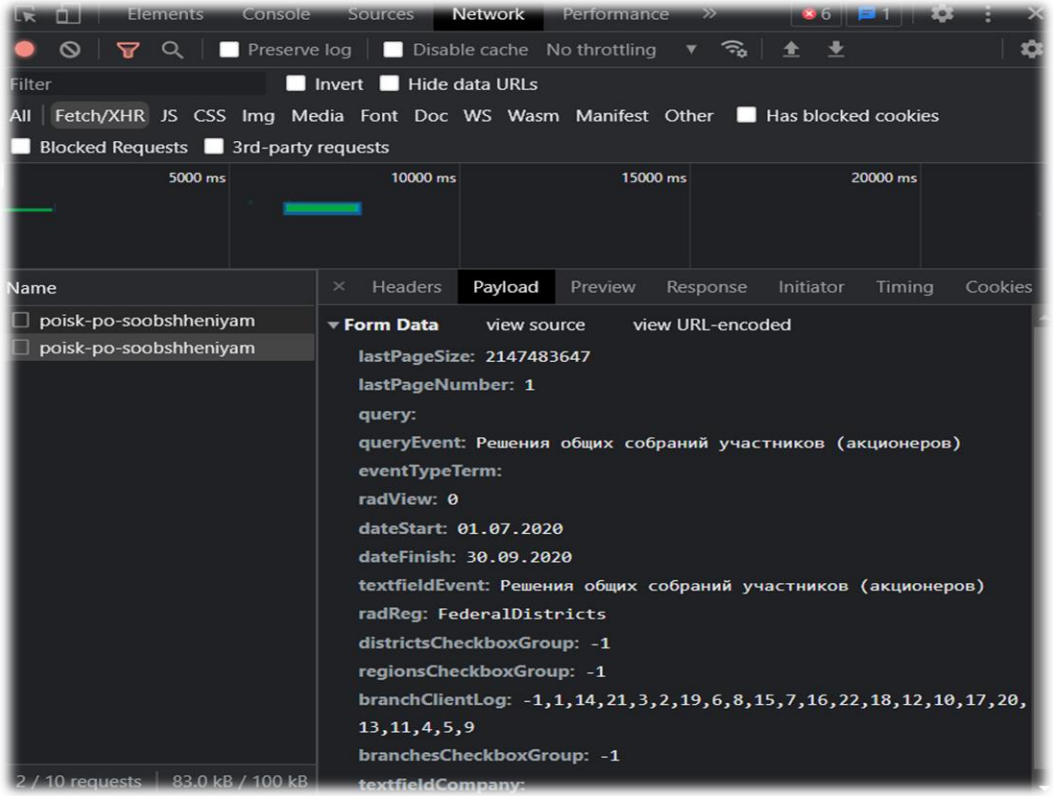
Извлечение сущностей и фактов из текстов

- Было сделано:
  - Для каждой сущности подобраны токены потенциально содержащие информацию
  - Каждый полученный токен обработан набором подобранных по данным эвристик
  - Извлеченные сущности и факты записаны в карточку сообщения
- Было получено:
  - Набор карточек: каждая соответствует одному сообщению и содержит информацию о выделенных сущностях и фактах

Экспорт данных

- Было сделано:
  - Карточки преобразованы в сводную таблицу: каждому столбцу соответствует конкретная сущность или факт
  - Полученная таблица объединена с ранее созданной сводной таблицей сообщений с сохранением порядка
  - Пост-обработка итоговой таблицы и экспорт в универсальные форматы: xlsx и csv
- Было получено:
  - Excel-таблица (датасет) для анализа





0 Интерфейс платформы

1 Параметры POST-запроса

```
<html><body><div class="infoblock">
<span style="color: red">Всего найдено сообщений: 2896, отображено сообщений: 1200.
an>
</div>
<div id="cont_wrap">
<table class="live noBorderTbl" style="width:100%">
<tr>
<td style="width:108px">30.09.2020 19:58</td>
<td>
<a href="https://e-disclosure.ru/portal/company.aspx?id=1929" target="_blank">ПАО "
<a href="https://e-disclosure.ru/portal/event.aspx?EventId=pjhgajsVBEC-CA6UsF3t5VQ-
Dz9%2bDx8u3o6u7iICjg6vbo7u3l8O7iKQ%3d%3d" style="color: red" target="_blank"><span
l">общих</span> <span class="hl">собраний</span> <span class="hl">участников</span>
<span class="graytext">ИНТЕРФАКС</span>
</td>
</tr>
<tr>
<td style="width:108px">30.09.2020 19:54</td>
<td>
```

```
{'registration_date': '30.09.2020 19:58',
 'entity': 'ПАО "СПБ Банк"',
 'entity_page_link': 'https://e-disclosure.ru/portal/company.aspx?id=1929',
 'event': 'Решения общих собраний участников (акционеров)',
 'event_page_link': 'https://e-disclosure.ru/portal/event.aspx?EventId=pjhg
403o6SDz9%2bDx8u3o6u7iICjg6vbo7u3l8O7iKQ%3d%3d'}
```

1 HTML-код ответа на запрос

1 Парсинг поисковой выдачи

	registration_date	entity	entity_page_link	event	event_page_link
0	30.09.2020 19:58	ПАО "СПБ Банк"	https://e-disclosure.ru/portal/company.aspx?id...	Решения общих собраний участников (акционеров)	https://e-disclosure.ru/portal/event.aspx?Even...
1	30.09.2020 19:54	ЗАО Агрофирма "Нива"	https://e-disclosure.ru/portal/company.aspx?id...	Решения общих собраний участников (акционеров)	https://e-disclosure.ru/portal/event.aspx?Even...
2	30.09.2020 19:46	ОАО "ДОРСТРОЙ"	https://e-disclosure.ru/portal/company.aspx?id...	Решения общих собраний участников (акционеров)	https://e-disclosure.ru/portal/event.aspx?Even...
3	30.09.2020 19:24	ООО КСН «Структурные инвестиции 1»	https://e-disclosure.ru/portal/company.aspx?id...	Решения общих собраний участников (акционеров)	https://e-disclosure.ru/portal/event.aspx?Even...
4	30.09.2020 19:12	ОАО "КАНАТ"	https://e-disclosure.ru/portal/company.aspx?id...	Решения общих собраний участников (акционеров)	https://e-disclosure.ru/portal/event.aspx?Even...
...	...	...	...	...	...
1195	25.08.2020 08:07	ООО «Магистраль двух столиц»	https://e-disclosure.ru/portal/company.aspx?id...	Решения общих собраний участников (акционеров)	https://e-disclosure.ru/portal/event.aspx?Even...
1196	25.08.2020 08:05	ПАО "СПБ Банк"	https://e-disclosure.ru/portal/company.aspx?id...	Решения совета директоров (наблюдательного сов...	https://e-disclosure.ru/portal/event.aspx?Even...
1197	24.08.2020 19:04	ОАО "ДОРСТРОЙ"	https://e-disclosure.ru/portal/company.aspx?id...	Решения совета директоров (наблюдательного сов...	https://e-disclosure.ru/portal/event.aspx?Even...
1198	24.08.2020 19:04	ОАО "ДОРСТРОЙ"	https://e-disclosure.ru/portal/company.aspx?id...	Созыв общего собрания участников (акционеров)	https://e-disclosure.ru/portal/event.aspx?Even...
1199	24.08.2020 18:40	АО "ИСТОК"	https://e-disclosure.ru/portal/company.aspx?id...	Решения общих собраний участников (акционеров)	https://e-disclosure.ru/portal/event.aspx?Even...

1200 rows × 5 columns

	registration_date	entity	entity_page_link	event	event_page_link
0	30.09.2020 19:58	ПАО "СПБ Банк"	https://e-disclosure.ru/portal/company.aspx?id...	Решения общих собраний участников (акционеров)	https://e-disclosure.ru/portal/event.aspx?Even...
1	30.09.2020 19:54	ЗАО Агрофирма "Нива"	https://e-disclosure.ru/portal/company.aspx?id...	Решения общих собраний участников (акционеров)	https://e-disclosure.ru/portal/event.aspx?Even...
2	30.09.2020 19:46	ОАО "ДОРСТРОЙ"	https://e-disclosure.ru/portal/company.aspx?id...	Решения общих собраний участников (акционеров)	https://e-disclosure.ru/portal/event.aspx?Even...
3	30.09.2020 19:24	ООО КСН «Структурные инвестиции 1»	https://e-disclosure.ru/portal/company.aspx?id...	Решения общих собраний участников (акционеров)	https://e-disclosure.ru/portal/event.aspx?Even...
4	30.09.2020 19:12	ОАО "КАНАТ"	https://e-disclosure.ru/portal/company.aspx?id...	Решения общих собраний участников (акционеров)	https://e-disclosure.ru/portal/event.aspx?Even...
...	...	...	...	...	...
1184	25.08.2020 09:34	ПАО "ЧИФ Союзинвест"	https://e-disclosure.ru/portal/company.aspx?id...	Решения общих собраний участников (акционеров)	https://e-disclosure.ru/portal/event.aspx?Even...
1185	25.08.2020 09:34	ПАО "ЧИФ Союзинвест"	https://e-disclosure.ru/portal/company.aspx?id...	Решения общих собраний участников (акционеров)	https://e-disclosure.ru/portal/event.aspx?Even...
1191	25.08.2020 08:39	АО "ЛЗЭП"	https://e-disclosure.ru/portal/company.aspx?id...	Решения общих собраний участников (акционеров)	https://e-disclosure.ru/portal/event.aspx?Even...
1195	25.08.2020 08:07	ООО «Магистраль двух столиц»	https://e-disclosure.ru/portal/company.aspx?id...	Решения общих собраний участников (акционеров)	https://e-disclosure.ru/portal/event.aspx?Even...
1199	24.08.2020 18:40	АО "ИСТОК"	https://e-disclosure.ru/portal/company.aspx?id...	Решения общих собраний участников (акционеров)	https://e-disclosure.ru/portal/event.aspx?Even...

531 rows × 5 columns

1 Таблица сообщений о раскрытии

2 Фильтрация сообщений

Сообщение о существенном факте о проведении общего собрания участников (акционеров) эмитента и о принятых им решениях

1. Общие сведения
1.1. Полное фирменное наименование эмитента (для некоммерческой организации – наименование) Публичное акционерное общество «Бест Эффортс Банк»
1.2. Сокращенное фирменное наименование эмитента ПАО «Бест Эффортс Банк»
1.3. Место нахождения эмитента Российская Федерация, город Москва
1.4. ОГРН эмитента 1037700041323
1.5. ИНН эмитента 7831000034
1.6. Уникальный код эмитента, присвоенный регистрирующим органом 0435
1.7. Адрес страницы в сети Интернет, используемой эмитентом для раскрытия информации http://www.e-disclosure.ru/portal/company.aspx?id=1929
http://besteffortsbank.ru/

1.8. Дата наступления события (существенного факта), о котором составлено сообщение (если применимо) 30.09.2010

2. Содержание сообщения
2.1. Вид общего собрания акционеров эмитента: годовое.
2.2. Форма проведения общего собрания акционеров эмитента: заочное голосование.
2.3. Дата проведения общего собрания:
2.3.1. Дата окончания приема бюллетеней: 29 сентября 2020 года.
2.3.2. Почтовый адрес, по которому направлялись заполненные бюллетени для голосования: Российская Федерация, 127006, г. Москва, ул. Долгоруковская, д.38, стр.1.
2.4. Кворум общего собрания акционеров эмитента:
Число голосов, которыми обладали лица, включенные в список лиц, имеющих право на участие в общем собрании - 56 490 000 голосов.
Число голосов, которыми обладали лица, принявшие участие в собрании – 56 208 742 голоса, что составляет 99,5021% от общего количества голосов, которыми обладали лица, включенные в список лиц, имеющих право на участие в общем собрании. Кворум имеется.
2.5. Повестка дня общего собрания акционеров эмитента:
1) Об утверждении годовой бухгалтерской (финансовой) отчетности ПАО «Бест Эффортс Банк» за 2019 год.
2) О распределении прибыли (в том числе выплате дивидендов) и убытков ПАО «Бест Эффортс Банк».
3) Об утверждении годового отчета ПАО «Бест Эффортс Банк» за 2019 год.
4) Об определении количественного состава Совета директоров ПАО «Бест Эффортс Банк».
5) Об избрании Совета директоров ПАО «Бест Эффортс Банк».
6) Об избрании Ревизионной комиссии ПАО «Бест Эффортс Банк».
7) Об утверждении аудитора для осуществления проверки финансово-хозяйственной деятельности ПАО «Бест Эффортс Банк» на 2020

```
['Сообщение о существенном факте о проведении общего собрания участников (акционеров) эмитента и о принятых им решениях',
 '1. Общие сведения',
 '1.1. Полное фирменное наименование эмитента (для некоммерческой организации – наименование)\tПубличное акционерное общество «Бест Эффортс Банк»',
 '1.2. Сокращенное фирменное наименование эмитента\tПАО «Бест Эффортс Банк»',
 '1.3. Место нахождения эмитента\tРоссийская Федерация, город Москва',
 '1.4. ОГРН эмитента\t1037700041323',
 '1.5. ИНН эмитента\t7831000034',
 '1.6. Уникальный код эмитента, присвоенный регистрирующим органом\t0435',
 '1.7. Адрес страницы в сети Интернет, используемой эмитентом для раскрытия информации\thttp://www.e-disclosure.ru/portal/company.aspx?id=1929 ',
 'http://besteffortsbank.ru/',
 '1.8. Дата наступления события (существенного факта), о котором составлено сообщение (если применимо)\t30.09.2010',
 '2. Содержание сообщения ',
 '2.1. Вид общего собрания акционеров эмитента: годовое.',
 '2.2. Форма проведения общего собрания акционеров эмитента: заочное голосование.',
 '2.3. Дата проведения общего собрания: ',
 '2.3.1. Дата окончания приема бюллетеней: 29 сентября 2020 года.',
 '2.3.2. Почтовый адрес, по которому направлялись заполненные бюллетени для голосования: Российская Федерация, 127006, г. Москва, ул. Долгоруковская, д.38, стр.1.',
 '2.4. Кворум общего собрания акционеров эмитента: ',
 'Число голосов, которыми обладали лица, включенные в список лиц, имеющих право на участие в общем собрании - 56 490 000 голосов.',
 'Число голосов, которыми обладали лица, принявшие участие в собрании – 56 208 742 голоса, что составляет 99,5021% от общего количества голосов, которыми обладали лица, включенные в список лиц, имеющих право на участие в общем собрании. Кворум имеется.',
 '2.5. Повестка дня общего собрания акционеров эмитента:',
 '1)\tОб утверждении годовой бухгалтерской (финансовой) отчетности ПАО «Бест Эффортс Банк» за 2019 год.',
 '2)\tО распределении прибыли (в том числе выплате дивидендов) и убытков ПАО «Бест Эффортс Банк».',
 '3)\tОб утверждении годового отчета ПАО «Бест Эффортс Банк» за 2019 год.',
 '4)\tОб определении количественного состава Совета директоров ПАО «Бест Эффортс Банк».',
 '5)\tОб избрании Совета директоров ПАО «Бест Эффортс Банк».',
 '6)\tОб избрании Ревизионной комиссии ПАО «Бест Эффортс Банк».',
 '7)\tОб утверждении аудитора для осуществления проверки финансово-хозяйственной деятельности ПАО «Бест Эффортс Банк» на 2020 год.',
 '2.6. Результаты голосования по вопросам повестки дня общего собрания акционеров эмитента, по которым имелся кворум, и формули
```

2 Получение текстов сообщений

2 Токенизация и предобработка

{'полное наименование': 'Публичное акционерное общество «Бест Эффортс Банк»',
'сокращенное наименование': 'ПАО «Бест Эффортс Банк»',
'адрес': 'Российская Федерация, город Москва',
'ИНН': '7831000034',
'ОГРН': '1037700041323',
'дата собрания': '30.09.2010',
'форма собрания': 'заочное голосование',
'наименование аудитора': 'Утвердить Общество с ограниченной ответственностью «Моор Стивенс» ',
'ИНН аудитора': 'не указан',
'тип отчетности': 'не указан',
'состав совета директоров': 'Жизненко Олег Михайлович, Бурдонова Марина Павловна, Горюнов Роман
евич, Ионова Ирина Борисовна, Соколов Кирилл Юрьевич, Старовойтова Ольга Владимировна',
'выплата дивидендов': 'принято решение не выплачивать дивиденды'}

	дата регистрации сообщения	текст сообщения	полное наименование	сокращенное наименование	адрес	ИНН	ОГРН	дата собрания	форма собрания	наименование аудитора
0	30.09.2020 19:58	Сообщение о существенном факте о проведении об...	Публичное акционерное общество «Бест Эффортс Б...	ПАО «Бест Эффортс Банк»	Российская Федерация, город Москва	7831000034	1037700041323	30.09.2010	заочное голосование	Утвердить Общество с ограниченной ответственно...
1	30.09.2020 19:54	Решения общих собраний участников (акционеров)...	Закрытое акционерное общество Агрофирма "Нива"	ЗАО Агрофирма "Нива"	140090, Московская область, г. Дзержинский, ул...	5027028404	1035010951722	30.09.2020	заочное голосование	Общество с ограниченной ответственностью «Конс...
2	30.09.2020 19:46	Решения общих собраний участников (акционеров)...	Открытое акционерное общество по строительству...	ОАО "ДОРСТРОЙ"	347800, Ростовская область, г. Каменск-Шахтинс...	6147002495	1026102107008	30.09.2020	совместное присутствие	Общество с ограниченной ответственностью «Аудит»
3	30.09.2020 19:24	Решения общих собраний участников (акционеров)...	Общество с ограниченной ответственностью «Комп...	ООО КСН «Структурные инвестиции 1»	125171, Российская Федерация, г. Москва, Ленин...	7743928024	1147746610725	30.09.2020	совместное присутствие	вопрос не поднимался
4	30.09.2020 19:12	Решения общих собраний участников (акционеров)...	Открытое акционерное общество "КАНАТ"	ОАО "КАНАТ"	Россия, Санкт-Петербург, 197110, Петровский пр...	7813054069	1027806857693	28.09.2020	заочное голосование	ООО «Агентство «Бизнес-Проект»

3 Обработка токенов и создание карточек

4 Сводная таблица сущностей и фактов

Направления возможного развития

Количественное развитие	Качественное развитие
Расширение охвата анализируемых типов сообщений о раскрытии (раскрытия годовой отчетности, финансовой отчетности)	Реализация более сложных эвристик, инвариантных к изменениям в данных
Расширение охвата сервисов с увеличением количества собираемых данных (например сбор данных из Банка данных ФССП)	Аналитика полученных результатов методами математической статистики и машинного обучения
Увеличение количества используемых платформ (например создание Telegram-бота с доступом к интерфейсу сервиса)	Добавление асинхронности и многопоточности для увеличения скорости работы с данными
Парсинг всей доступной базы сообщений о раскрытии с автоматизацией добора данных по мере их появления	Централизованное хранение в распределенной системе

СПАСИБО ЗА ВНИМАНИЕ!