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

Выполнили:
Даниил Волгин, Иван Фридман

Руководитель:
Константин Дудников

Цели работы

- Построение сервиса для извлечения информации из текстового представления решений общих собраний участников (акционеров)
- Анализ результатов работы построенного сервиса на решениях общих собраний участников за второй квартал 2021 года

Обзор аналогичных сервисов

- **GATE** – бесплатный набор инструментов для многих задач обработки естественного языка, в том числе извлечения информации на различных языках
- **Stanford CoreNLP** – программное средство, которое предназначено для создания приложений анализа текстов на естественном языке
- **Machine Learning for Language Toolkit** – Java библиотека для различных задач обработки естественного языка с использованием методов машинного обучения
- **Томита-парсер** – проект Яндекса, предназначен для извлечения структурированных данных из текста на русском языке

Сбор и парсинг данных

- Получение текстов решений общих собраний акционеров из сервиса https://e-disclosure.ru/ с помощью отправки HTTP запросов
- Парсинг полученных HTML-результатов с помощью python-библиотеки Beautiful Soup
- Используются данные второго квартала 2021 года по Москве
- В итоге имеем таблицу с id и текстами решений

Модель извлечения информации

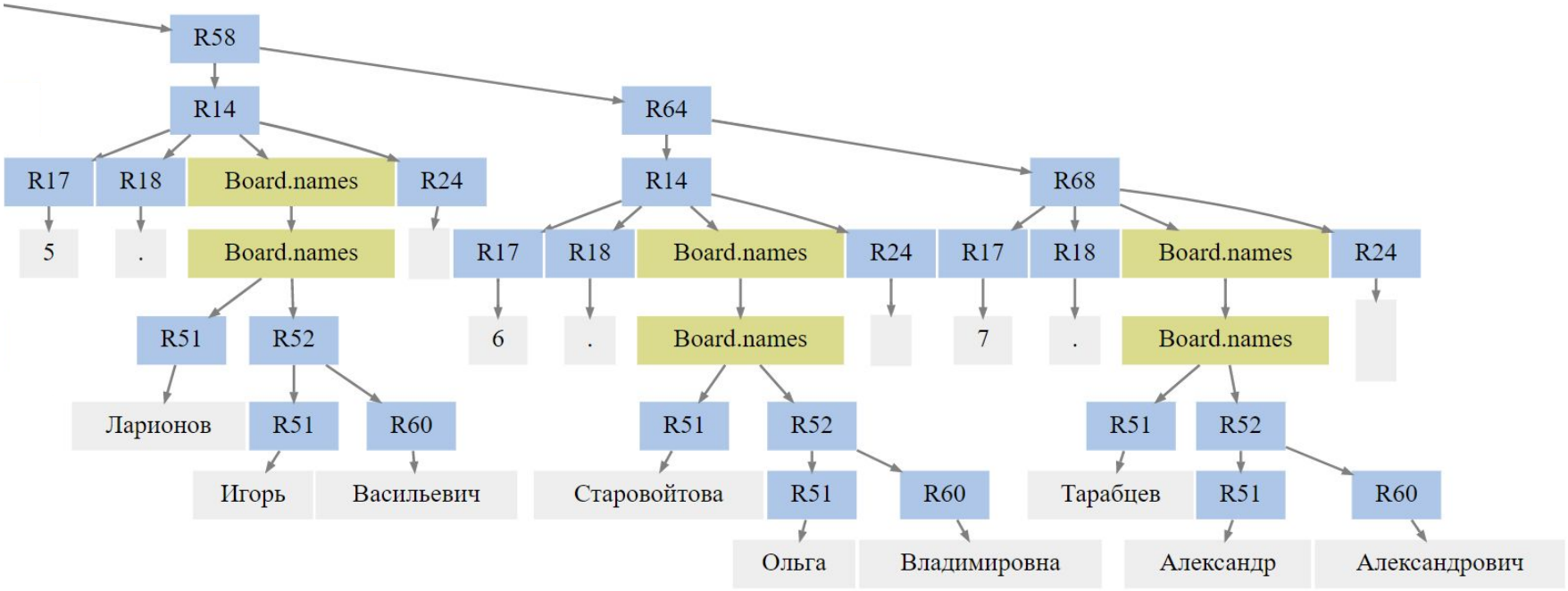
Используется rule-based методы извлечения информации

- Регулярные выражения
- Yargy-парсер, использующий правила и словари, чтобы извлекать структурированную информацию из текстов на русском языке

Выделяемые сущности

- Полное наименование
- Сокращенное наименование
- Адрес
- ИНН
- ОГРН
- Дата собрания
- Форма собрания
- Решение по вопросу выплаты дивидендов
- Состав совета директоров

Пример дерева разбора

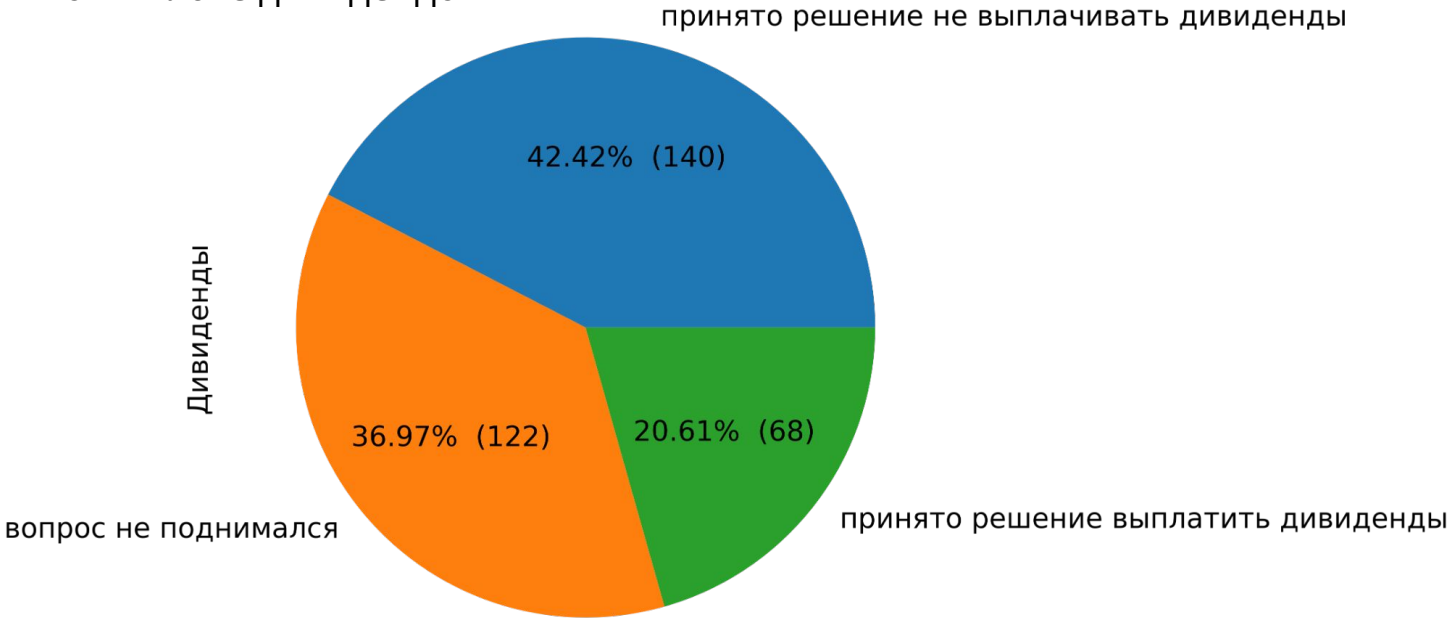


Пример применения решения

content	Полное наименование	Сокращенное наименование	Адрес	ИНН	ОГРН	Дата собрания	Форма собрания	Дивиденды	Совет директоров
Решения общих собраний участников (акционеров)...	Публичное акционерное общество "Аптечная сеть ...	ПАО "Аптечная сеть 36,6"	г. Москва Российская Федерация	7722266450	1027722000239	2021-06-30	заочное голосование	принято решение не выплачивать дивиденды	['Захаров Владимир Эдуардович', 'Кузин Алексан...
Сообщение о существенном факте о проведении об...	Публичное акционерное общество «Бест Эффортс Б...	ПАО «Бест Эффортс Банк»	Российская Федерация, город Москва	7831000034	1037700041323	2021-06-30	собрание	принято решение не выплачивать дивиденды	['Соколов Кирилл Юрьевич', 'Бурдонова Марина П...
Существенные факты, касающиеся событий эмитент...	Акционерное общество "Научно-исследовательский...	АО "Институт Цветметобработка"	119017, г. Москва, Пыжевский пер., д. 5	7706002901	1027700122768	2021-06-30	заочное голосование	принято решение не выплачивать дивиденды	['Дружинина Татьяна Ивановна', 'Райков Юрий Ни...
Сообщение о существенном факте\n«О проведении ...	Публичное акционерное общество «Вымпел-Коммуни...	ПАО «ВымпелКом»	127083, Российская Федерация, г. Москва, ул. 8...	7713076301	1027700166636	2021-06-30	собрание	принято решение выплатить дивиденды	['Махтерем Каан Терзиоглу', 'Серкан Окандан', ...
Решения общих собраний участников (акционеров)...	Публичное акционерное общество Группа компаний...	ПАО ГК «ТНС энерго»	127006, Российская Федерация, г. Москва, Наста...	7705541227	1137746456231	2021-06-30	собрание	принято решение не выплачивать дивиденды	['Тихонова Мария Геннадьевна', 'Стельнова Елен...

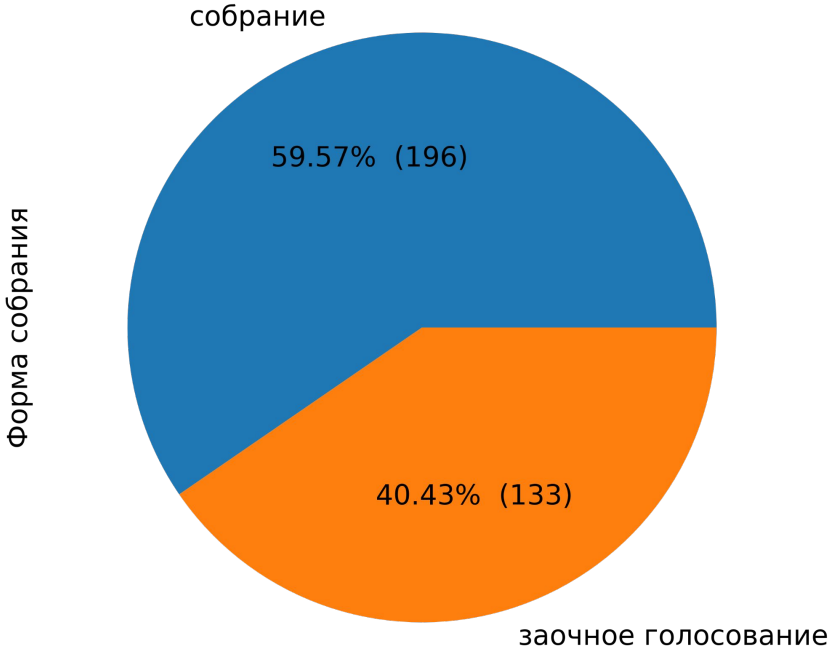
Анализ полученных результатов

Решения по выплате дивидендов



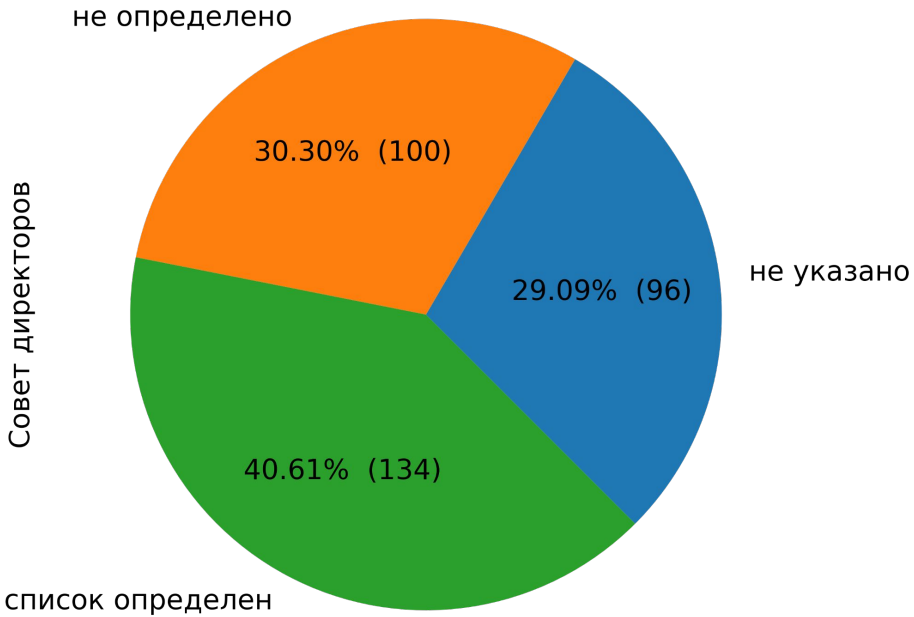
Анализ полученных результатов

Форма собрания

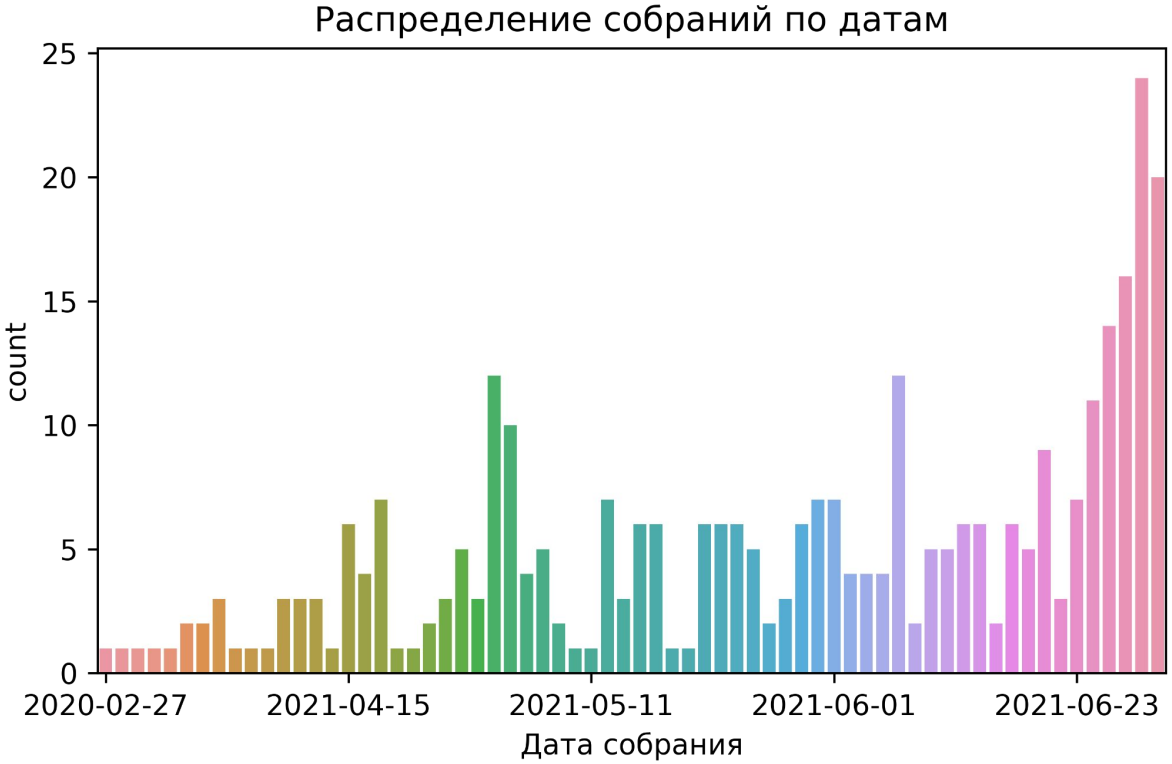


Анализ полученных результатов

Совет директоров



Анализ полученных результатов



Проделанная работа

В ходе работы были достигнуты следующие цели:

- Реализован сервис по выделению структурированной информации из данных в текстовом виде
- Проведено тестирование сервиса на текстах решений собраний акционеров за второй квартал 2021 года по Москве
- Проведен анализ выделенных данных

Распределение задач

Даниил

- Разработана и реализована архитектура сервиса для удобного добавления извлекаемых сущностей и правил для извлечения
- Придуманы и реализованы правила для извлечения части сущностей

Иван

- Придуманы и реализованы правила для извлечения части сущностей
- Проведена валидация и анализ извлеченных данных

Направления развития

- Больше выделяемых сущностей
- Повышение покрытия форматов текстов решений собраний акционеров
- Консольный/графический интерфейс для удобного взаимодействия с сервисом

Ссылки

Реализация

- https://github.com/DanWallgun/disclosure-parsing

Используемые ресурсы

- https://e-disclosure.ru/
- requests, pandas
- Beautiful Soup https://www.crummy.com/software/BeautifulSoup/bs4/
- Yargy-parser https://github.com/natasha/yargy