



Факультет экономических наук

Москва, 2022

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

Коняева Екатерина, Зародова Олеся, Мальков Сергей



ЦЕЛИ И ЗАДАЧИ СЕРВИСА

Цели

- Автоматизация сбора информации из источников сети Интернет
- Эффективная пред- и пост-обработка данных
- Разработка универсальных правил извлечения сущностей
- Оптимальная стратегия извлечения сущностей и фактов

Задачи

- Создание отказоустойчивого алгоритма извлечения данных с e-disclosure.ru
- Разработка минимально необходимых шагов представления текстовых данных
- Поиск оптимальных правил (регулярных выражений) для нахождения паттернов
- Оптимизация по времени и памяти применения алгоритма извлечения данных



ОБЗОР АНАЛОГИЧНЫХ СЕРВИСОВ

○ ScraperAPI

✚ Работа с AJAX

✚ Обход CAPTCHA

— Нет извлечения
сущностей

— Только скрэпинг

○ ParseHub

✚ Работа с GUI

✚ Экспорт в JSON,
CSV, XLSX

— Нет извлечения
сущностей

— Сохранение
«сырых» данных

○ FMiner

✚ Десктоп решение

✚ Многопоточность

— Нет извлечения
сущностей

— Нет пост-обработки
данных

ЭТАП 1: ПАРСИНГ СООБЩЕНИЙ О РАСКРЫТИИ


- Эмуляция браузера через Selenium
- POST-запрос с данными для поиска
- GET-запрос HTML-кода ответа
- Создание карточек из данных поисковой выдачи
- Фильтрация Решений собраний участников (акционеров)

```
cards[0]
```

```
{'event': 'Решения общих собраний участников (акционеров)',  
  'company': 'ПАО "Строймаш"',  
  'event_page': 'https://e-disclosure.ru/portal/event.aspx?EventId=g-AFNEiaBwUutOscxAoe5EQ-B-B&q=00X45e3o%2fyDu4fno9SDx7uHw403o6SDz9%2bDx8u3o6u7iICjg6vbo7u3l807iKQ%3d%3d'}
```

Результат 1-го этапа

ЭТАП 2: ПАРСИНГ ТЕКСТОВ СООБЩЕНИЙ

- 
- ```
graph TD; A[GET-запрос по ссылке сообщения из карточки] --> B[Выявление блока с текстом сообщения]; B --> C[Добавление текста в карточку сообщения];
```
- GET-запрос по ссылке сообщения из карточки
  - Выявление блока с текстом сообщения
  - Добавление текста в карточку сообщения

```
{ 'event': 'Решения общих собраний участников (акционеров)',
 'company': 'ООО «Интертехэлектро- Новая генерация»,
 'event_page': 'https://e-disclosure.ru/portal/event.aspx?EventId=03KS6kNhQuYfwrSKQe5Pig-B-B&q=00X45e30%2fyDu4fno9SDx7uHw403o6S
Dz9%2bDx8u3o6u7iICjg6vbo7u3l807iKQ%3d%3d',
 'text': 'Решения общих собраний участников (акционеров)\r\n\r\n1. Общие сведения\r\n1.1. Полное фирменное наименование эмитент
а (для некоммерческой организации – наименование): Общество с ограниченной ответственностью «Интертехэлектро- Новая генераци
я»\r\n1.2. Сокращенное фирменное наименование эмитента: ООО «Интертехэлектро- Новая генерация»\r\n1.3. Место нахождения эмитен
та: 105062, г.Москва, ул.Чаплыгина, д.11, этаж/пом 2/1\r\n1.4. ОГРН эмитента: 1057749387321\r\n1.5. ИНН эмитента: 7701633050\r
\r\n1.6. Уникальный код эмитента, присвоенный регистрирующим органом: 00171-R\r\n1.7. Адрес страницы в сети Интернет, используемо
й эмитентом для раскрытия информации: http://www.e-disclosure.ru/portal/company.aspx?id=29969; http://www.ite-ng.pw/\r\n1.8. Дат
а наступления события (существенного факта), о котором составлено сообщение: 26.02.2020\r\n\r\n2. Содержание сообщения\r\n2.1.
Вид общего собрания участников: внеочередное\r\n2.2. Форма проведения общего собрания участников: собрание (совместное присутстви
е)\r\n2.3. Дата, место и время проведения общего собрания участников: 26.02.2020, РФ, г. Москва, ул.Чаплыгина, д.11, этаж/пом 2/
1, 13:10\r\n2.4. Кворум: 100%\r\n2.5. Повестка дня общего собрания участников: \r\n1. Определение количественного состава Совета дире
кторов Общества.\r\n2. Избрание членов Совета директоров Общества.\r\n2.6. Результаты голосования и формулировка решений: \r\n3А – 10
0%;\r\n1. Определить количественный состав Совета директоров Общества – 4 (четыре) человека.\r\n2. Избрать Совет директоров Обществ
а в количестве 4 (четырёх) человек в следующем составе:\r\n1). \tБиков Артем Эльбрусович;\r\n2). \tКарапетян Станислав Сейранович;\r
3). \tТерновский Геннадий Семенович;\r\n4). \tКоногоров Дмитрий Николаевич.\r\n2.7. Дата составления и номер протокола общего со
брания участников: 26.02.2020 №6/н\r\n\r\n3. Подпись\r\n3.1. Генеральный директор\r\nс.с. Карапетян\r\n\r\n\r\n3.2. Дата 26.0
2.2020г. '}
```

## Результат 2-го этапа



## ЭТАП 3: ВЫДЕЛЕНИЕ СУЩНОСТЕЙ И ФАКТОВ

Создание регулярных выражений для  
поиска паттернов в текстах

Поиск подстрок текстов сообщений,  
содержащих созданные паттерны

Выделение из подстрок сущностей и  
фактов паттернами регулярных выражений

Фильтрация сущностей и фактов по  
специальным правилам

Обновление карточек сообщений  
выделенными сущностями и фактами

```
{'event': 'Решения общих собраний участников (акционеров)',
'company': 'ООО «Интертехэлектро- Новая генерация»',
'event_page': 'https://e-disclosure.ru/portal/event.aspx?EventId=03KS6kNhQuyfwrSKQe5Pig-B-B&q=00X45e30%2fyDu4fno9SDx7uHw403o6S
Dz9%2bDx8u3o6u7iICjg6vbo7u3l807iKQ%3d%3d',
'text': 'Решения общих собраний участников (акционеров)\r\n\r\n1. Общие сведения\r\n\r\n1.1. Полное фирменное наименование эмитент
а (для некоммерческой организации – наименование): Общество с ограниченной ответственностью «Интертехэлектро- Новая генераци
я»\r\n\r\n1.2. Сокращенное фирменное наименование эмитента: ООО «Интертехэлектро- Новая генерация»\r\n\r\n1.3. Место нахождения эмитен
та: 105062, г.Москва, ул.Чаплыгина, д.11, этаж/пом 2/1\r\n\r\n1.4. ОГРН эмитента: 1057749387321\r\n\r\n1.5. ИНН эмитента: 7701633050\r\n\r\n1.6. Уникальный код эмитента, присвоенный регистрирующим органом: 00171-R\r\n\r\n1.7. Адрес страницы в сети Интернет, используемо
й эмитентом для раскрытия информации: http://www.e-disclosure.ru/portal/company.aspx?id=29969; http://www.ite-ng.pw.ru\r\n\r\n1.8. Дат
а наступления события (существенного факта), о котором составлено сообщение: 26.02.2020\r\n\r\n2. Содержание сообщения\r\n\r\n2.1. Вид
общего собрания участников: внеочередное\r\n\r\n2.2. Форма проведения общего собрания участников: собрание (совместное присутстви
е)\r\n\r\n2.3. Дата, место и время проведения общего собрания участников: 26.02.2020, РФ, г. Москва, ул.Чаплыгина, д.11, этаж/пом 2/
1, 13:10\r\n\r\n2.4. Кворум: 100%\r\n\r\n2.5. Повестка дня общего собрания участников: \r\n\r\n1. Определение количественного состава Совета дире
кторов Общества.\r\n\r\n2. Избрание членов Совета директоров Общества.\r\n\r\n2.6. Результаты голосования и формулировка решений: \r\n3А – 10
0%;\r\n\r\n1. Определить количественный состав Совета директоров Общества – 4 (Четыре) человека.\r\n\r\n2. Избрать Совет директоров Обществ
а в количестве 4 (Четырех) человек в следующем составе:\r\n\r\n1).\tБиков Артем Эльбрусович;\r\n\r\n2).\tКарапетян Станислав Сейранович;\r\n\r\n3).\tТерновский Геннадий Семенович;\r\n\r\n4).\tКоногоров Дмитрий Николаевич.\r\n\r\n2.7. Дата составления и номер протокола общего со
брания участников: 26.02.2020 №6/Н\r\n\r\n\r\n\r\n3. Подпись\r\n\r\n3.1. Генеральный директор\r\n\r\n\r\n\r\n3.2. Дата 26.0
2.2020г.',
'directors_list': 'Биков Артем Эльбрусович, Карапетян Станислав Сейранович, Терновский Геннадий Семенович, Коногоров Дмитрий Н
иколаевич',
'auditor': 'вопрос не поднимался',
'auditor_inn': 'не указан',
'auditor_documents': 'не указан',
'dividends': 'вопрос не поднимался',
'event_date': '26.02.2020',
'event_form': 'собрание (совместное присутствие)',
'place': '105062, г.Москва, ул.Чаплыгина, д.11, этаж/пом 2/1',
'ogrn': '1057749387321',
'inn': '7701633050',
'full_name': 'Общество с ограниченной ответственностью «Интертехэлектро- Новая генерация»',
'short_name': 'ООО «Интертехэлектро- Новая генерация»'}
```

Результат 3-го этапа



## ЭТАП 4: СОЗДАНИЕ СВОДНОЙ ТАБЛИЦЫ

- Создание датафрейма из карточек
- Пост-обработка сводной таблицы
- Экспорт таблицы в формат XLSX
- [https://github.com/ekaterinaHSE/disclosure\\_messages](https://github.com/ekaterinaHSE/disclosure_messages)

| Список совета директоров                          | Наименование аудиторской организации | ИНН аудиторской организации | Тип проверяемой аудитором отчетности  | Решение о выплате дивидендов             | Дата события | Форма события                     | Адрес организации                                                     | ОГРН организации | ИНН организации | Полное наименование организации                             |
|---------------------------------------------------|--------------------------------------|-----------------------------|---------------------------------------|------------------------------------------|--------------|-----------------------------------|-----------------------------------------------------------------------|------------------|-----------------|-------------------------------------------------------------|
| Борисов Александр Борисович, Головачев Виталий... | «Солинг ЛТД»                         | не указан                   | не указан                             | принято решение не выплачивать дивиденды | 31.03.2020   | собрание (совместное присутствие) | 196084, г. Санкт-Петербург, пр. Лиговский, д. ...                     | 1027804884436    | 7810244262      | Публичное акционерное общество "Строитель"                  |
| вопрос не поднимался                              | вопрос не поднимался                 | не указан                   | не указан                             | вопрос не поднимался                     | 31.03.2020   | заочное голосование               | 443048, Самарская область, г. Самара, Красногвардейский район, д. ... | 1176313036712    | 6313553082      | Общество с ограниченной ответственностью "Самарский регион" |
| вопрос не поднимался                              | вопрос не поднимался                 | не указан                   | не указан                             | вопрос не поднимался                     | 31.03.2020   | собрание (совместное присутствие) | 125124, Москва, улица Правды, дом 8, корпус 1                         | 1027700280937    | 7735057951      | Общество с ограниченной ответственностью "Правда"           |
| Андреевский Дмитрий Евгеньевич, Иванченко Макс... | «Поволжье»                           | 632502023                   | не указан                             | принято решение выплатить дивиденды      | 24.03.2020   | собрание (совместное присутствие) | Российская Федерация, Самарская область, г. Сыктывкар                 | 1026303055063    | 6325006694      | Акционерное общество "Самарский регион"                     |
| вопрос не поднимался                              | «Универсалаудит                      | не указан                   | бухгалтерской (финансовой) отчетности | принято решение выплатить дивиденды      | 21.06.2019   | собрание (совместное присутствие) | 142500, Россия, Московская область, г. Павловский Посад               | 1025004643069    | 5035002880      | Акционерное общество "Павловский Посад"                     |

Результат 4-го этапа



# ВЫВОДЫ

- Поиск оптимальных паттернов оптимально осуществлять с помощью регулярных выражений
- Наиболее эффективным является двухуровневый алгоритм: (1) поиск подстрок с требуемой информацией, (2) если найдены, поиск непосредственно сущностей
- При работе с регулярными выражениями не требуется особой пред-обработки данных, паттерны игнорируют специальные символы
- Регулярные выражения не позволяют полностью отказаться от формирования специальных правил семантической фильтрации получившихся сущностей
- Коммерческое использование полученной информации требует проведения глубокой постобработки данных





# РЕКОМЕНДУЕМЫЕ НАПРАВЛЕНИЯ РАЗВИТИЯ

- Увеличение количества выявляемых сущностей и фактов – разложение всего текста сообщения о раскрытии на набор сущностей
- Расширение охвата сервисов для получения разноплановой информации об организациях (реестр ЕГРЮЛ, ЕГРН, ФССП)
- Создание GUI-оболочки для удобного и User-friendly взаимодействия с сервисом
- Оптимизация используемых паттернов для более универсального поиска сущностей без необходимости в специальных правилах фильтрации по семантике
- Интеграция методов машинного обучения в алгоритм поиска сущностей: разметка частей речи, структурирование текстов по текстовым эмбедингам



# ВКЛАД УЧАСТНИКОВ КОМАНДЫ

## ○ Екатерина Коняева:

- парсинг карточек сообщений о раскрытии через Selenium (написание кода)
- создание паттернов поиска списка совета директоров
- написание поискового цикла для списка совета директоров

## ○ Олеся Зародова:

- создание паттернов поиска наименования аудитора, его ИНН и типа отчетности
- создание паттернов поиска даты, формы события и наименований эмитента
- написание поисковых циклов для указанных паттернов

## ○ Сергей Мальков:

- создание паттернов поиска информации о дивидендах, адресе, ИНН и ОГРН
- написание поисковых циклов для данных паттернов
- создание сводной таблицы данных, пост-обработка
- создание презентации

## ○ [https://github.com/ekaterinaHSE/disclosure\\_messages](https://github.com/ekaterinaHSE/disclosure_messages)

**СПАСИБО ЗА ВНИМАНИЕ!**