Инструменты и методы текстовой аналитики

На пальцах

Алексей Пятов // Руководитель группы текстовой аналитики и управления данными
Константин Дудников // Эксперт по текстовой аналитике

210527

План

1. Регулярные выражения
2. Тематическое моделирование
3. Классификация текстов
4. Практикум №1
5. Извлечение сущностей и фактов
6. Бизнес: анализ юридических документов
7. Практикум №2

Регулярные выражения

Регулярные выражения

Регулярное выражение – это шаблон, по которому мы ищем информацию в тексте

Примеры задач:

Найти в тексте ИНН и ОГРН юрлица

Найти в тексте название компании

Убрать все имена из корпуса в 9 млн текстов

Регулярные выражения

Найти в тексте ИНН и ОГРН юрлица

ИНН – последовательность из 12 цифр для физлица, и из 10 цифр – для юрлица.
ОГРН – последовательность из 13 цифр.

Шаблон для ИНН: \b\d{10}\b или \b[0-9]{10}\b
Шаблон для ОГРН: \b\d{13}\b или \b[0-9]{13}\b

\b – граница слова
\d – любая цифра
[0-9] – любое значение от 0 до 9
{n} – n повторений

Примеры контекстов, где такого шаблона недостаточно:
«Номер телефона для связи: +7 9991234567»
«По условиям госконтракта № 1000000763972 …»

Регулярные выражения

Найти в тексте ИНН и ОГРН юрлица

Какие контексты необходимо учесть:
ИНН 1234567890
ИНН: 1234567890
ИНН – 1234567890

Шаблон для ИНН: (?<=инн:)\d{10}\b|(?<=инн [–\-])\d{10}\b|(?<=инн)\d{10}\b
Шаблон для ОГРН: (?<=огрн:)\d{13}\b|(?<=огрн [–\-])\d{13}\b|(?<=огрн)\d{13}\b

Регулярные выражения

Найти в тексте ИНН и ОГРН юрлица

Какие контексты необходимо учесть:
ИНН 1234567890
ИНН: 1234567890
ИНН – 1234567890

Шаблон для ИНН: (?<=инн:)\d{10}\b|(?<=инн [–\-])\d{10}\b|(?<=инн)\d{10}\b
Шаблон для ОГРН: (?<=огрн:)\d{13}\b|(?<=огрн [–\-])\d{13}\b|(?<=огрн)\d{13}\b

Что тут происходит?

Регулярные выражения

Найти в тексте ИНН и ОГРН юрлица

Шаблон для ИНН: (?<=инн:)\d{10}\b|(?<=инн [–\-])\d{10}\b|(?<=инн)\d{10}\b
Шаблон для ОГРН: (?<=огрн:)\d{13}\b|(?<=огрн [–\-])\d{13}\b|(?<=огрн)\d{13}\b

| – оператор ИЛИ.
(?<=)\d{10} – ищем 10 цифр только если перед ними есть выражение, указанное в скобках (positive lookbehind).
\- – ищем в тексте дефис. Если без слэша, то дефис внутри квадратных скобкок – это специальный символ.
[–\-] – ищем на выбор тире или дефис.

Документация python по библиотеке re
Регулярные выражения, пособие для новичков часть 1, часть 2
Онлайн-редактор и проверка регулярных выражений

Регулярные выражения

Пример кода

```
In [1]: import re

In [2]: innR = re.compile(r'(?<=инн: )\d{10}\b|(?<=инн [-\-] )\d{10}\b|(?<=инн )\d{10}\b')

In [3]: corpus = [
            '1. Реквизиты компании: ИНН - 1234567890.',
            '2. ООО "Ромашка", инн: 0987654321.',
            '3. Иванов И.И., ИНН 123456789012, тел. +7 9991234567'
        ]

In [4]: inns = []
        for text in corpus:
            inn = re.search(innR, text.lower())
            if inn:
                inns.append(inn[0])
            else:
                inns.append(False)

        inns

Out[4]: ['1234567890', '0987654321', False]
```

Тематическое моделирование

Тематическое моделирование

На что в основном жалуются пользователи?

Тематическое моделирование

Что такое текст и что такое тема?

Тематическое моделирование

Что такое текст и что такое тема?

Текст – набор слов.
Тема – набор ключевых слов.

Тематическое моделирование

Что такое текст и что такое тема?

Текст – набор слов.
Тема – набор ключевых слов.

1. Британская полиция знает о местонахождении основателя WikiLeaks.
2. В суде США начинается процесс против россиянина, рассылавшего спам.
3. Церемонию вручения Нобелевской премии мира бойкотируют 19 стран.
4. В Великобритании арестован основатель сайта Wikileaks Джулиан Ассандж.
5. Украина игнорирует церемонию вручения Нобелевской премии.
6. Шведский суд отказался рассматривать апелляцию основателя Wikileaks.
7. НАТО и США разработали планы обороны стран Балтии против России.
8. Полиция Великобритании нашла основателя WikiLeaks, но не арестовала.
9. В Стокгольме и Осло сегодня состоится вручение Нобелевских премий.

Тематическое моделирование

Подготовка корпуса

Токенизация – делаем из строки набор слов.

1. [британская], [полиция], [знает], [о], [местонахождении], [основателя], [wikileaks]
2. [в], [суде], [сша], [начинается], [процесс], [против], [россиянина], [рассылавшего], [спам]
3. [церемонию], [вручения], [нобелевской], [премии], [мира], [бойкотируют], [19], [стран]
4. [в], [великобритании], [арестован], [основатель], [сайта], [wikileaks], [джулиан], [ассандж]
5. [украина], [игнорирует], [церемонию], [вручения], [нобелевской], [премии]
6. [шведский], [суд], [отказался], [рассматривать], [апелляцию], [основателя], [wikileaks]
7. [нато], [и], [сша], [разработали], [планы], [обороны], [стран], [балтии], [против], [россии]
8. [полиция], [великобритании], [нашла], [основателя], [wikileaks], [но], [не], [арестовала]
9. [в], [стокгольме], [и], [осло], [сегодня], [состоится], [вручение], [нобелевских], [премий]

Тематическое моделирование

Подготовка корпуса

Лемматизация – делаем количество слов в корпусе меньше.

1. [британская], [полиция], [знает], [о], [местонахождении], [основателя], [wikileaks]
2. [в], [суде], [сша], [начинается], [процесс], [против], [россиянина], [рассылавшего], [спам]
3. [церемонию], [вручения], [нобелевской], [премии], [мира], [бойкотируют], [19], [стран]
4. [в], [великобритании], [арестован], [основатель], [сайта], [wikileaks], [джулиан], [ассандж]
5. [украина], [игнорирует], [церемонию], [вручения], [нобелевской], [премии]
6. [шведский], [суд], [отказался], [рассматривать], [апелляцию], [основателя], [wikileaks]
7. [нато], [и], [сша], [разработали], [планы], [обороны], [стран], [балтии], [против], [россии]
8. [полиция], [великобритании], [нашла], [основателя], [wikileaks], [но], [не], [арестовала]
9. [в], [стокгольме], [и], [осло], [сегодня], [состоится], [вручение], [нобелевских], [премий]

Тематическое моделирование

Подготовка корпуса

Лемматизация – делаем количество слов в корпусе меньше.

1. [британский], [полиция], [знать], [о], [местонахождение], [основатель], [wikileaks]
2. [в], [суд], [сша], [начинаться], [процесс], [против], [россиянин], [рассылавший], [спам]
3. [церемония], [вручение], [нобелевский], [премия], [мир], [бойкотировать], [19], [страна]
4. [в], [великобритания], [арестованный], [основатель], [сайт], [wikileaks], [джулиан], [ассандж]
5. [украина], [игнорировать], [церемония], [вручение], [нобелевский], [премия]
6. [шведский], [суд], [отказаться], [рассматривать], [апелляция], [основатель], [wikileaks]
7. [нато], [и], [сша], [разработать], [план], [оборона], [страна], [балтия], [против], [россия]
8. [полиция], [великобритания], [найти], [основатель], [wikileaks], [но], [не], [арестовать]
9. [в], [стокгольм], [и], [осло], [сегодня], [состояться], [вручение], [нобелевский], [премия]

Тематическое моделирование

Подготовка корпуса

Удаление стоп-слов – никакого смысла не несут.

1. [британский], [полиция], [знать], [о], [местонахождение], [основатель], [wikileaks]
2. [в], [суд], [сша], [начинаться], [процесс], [против], [россиянин], [рассылавший], [спам]
3. [церемония], [вручение], [нобелевский], [премия], [мир], [бойкотировать], [19], [страна]
4. [в], [великобритания], [арестованный], [основатель], [сайт], [wikileaks], [джулиан], [ассандж]
5. [украина], [игнорировать], [церемония], [вручение], [нобелевский], [премия]
6. [шведский], [суд], [отказаться], [рассматривать], [апелляция], [основатель], [wikileaks]
7. [нато], [и], [сша], [разработать], [план], [оборона], [страна], [балтия], [против], [россия]
8. [полиция], [великобритания], [найти], [основатель], [wikileaks], [но], [не], [арестовать]
9. [в], [стокгольм], [и], [осло], [сегодня], [состояться], [вручение], [нобелевский], [премия]

Тематическое моделирование

Подготовка корпуса

Удаление стоп-слов – никакого смысла не несут.

1. [британский], [полиция], [знать], [местонахождение], [основатель], [wikileaks]
2. [суд], [сша], [начинаться], [процесс], [россиянин], [рассылавший], [спам]
3. [церемония], [вручение], [нобелевский], [премия], [мир], [бойкотировать], [страна]
4. [великобритания], [арестованный], [основатель], [сайт], [wikileaks], [джулиан], [ассандж]
5. [украина], [игнорировать], [церемония], [вручение], [нобелевский], [премия]
6. [шведский], [суд], [отказаться], [рассматривать], [апелляция], [основатель], [wikileaks]
7. [нато], [сша], [разработать], [план], [оборона], [страна], [балтия], [россия]
8. [полиция], [великобритания], [найти], [основатель], [wikileaks], [арестовать]
9. [стокгольм], [осло], [состояться], [вручение], [нобелевский], [премия]

Тематическое моделирование

Подготовка корпуса

Удаление редких слов – либо опечатки, либо не помогают описать тему.

1. [британский], [полиция], [знать], [местонахождение], [основатель], [wikileaks]
2. [суд], [сша], [начинаться], [процесс], [россиянин], [рассылавший], [спам]
3. [церемония], [вручение], [нобелевский], [премия], [мир], [бойкотировать], [страна]
4. [великобритания], [арестованный], [основатель], [сайт], [wikileaks], [джулиан], [ассандж]
5. [украина], [игнорировать], [церемония], [вручение], [нобелевский], [премия]
6. [шведский], [суд], [отказаться], [рассматривать], [апелляция], [основатель], [wikileaks]
7. [нато], [сша], [разработать], [план], [оборона], [страна], [балтия], [россия]
8. [полиция], [великобритания], [найти], [основатель], [wikileaks], [арестовать]
9. [стокгольм], [осло], [состояться], [вручение], [нобелевский], [премия]

Тематическое моделирование

Латентно-семантический анализ

Строим терм-документную матрицу

	d1	d2	d3	d4	d5	d6	d7	d8	d9
апелляция	0	0	0	0	0	1	0	0	0
арестованный	0	0	0	1	0	0	0	0	0
арестовать	0	0	0	0	0	0	0	1	0
балтия	0	0	0	0	0	0	1	0	0
бойкотировать	0	0	1	0	0	0	0	0	0
британский	1	0	0	0	0	0	0	0	0
великобритания	0	0	0	1	0	0	0	1	0
вручение	0	0	1	0	1	0	0	0	1
…									

Тематическое моделирование

Латентно-семантический анализ

Взвешиваем термы:

- tf-idf
- Mutual information (MI)
- Entropy

	d1	d2	d3	d4	d5	d6	d7	d8	d9
апелляция	0	0	0	0	0	0,46	0	0	0
арестованный	0	0	0	0,18	0	0	0	0	0
арестовать	0	0	0	0	0	0	0	0,61	0
балтия	0	0	0	0	0	0	0,38	0	0
бойкотировать	0	0	0,37	0	0	0	0	0	0
британский	0,19	0	0	0	0	0	0	0	0
великобритания	0	0	0	0,18	0	0	0	0,23	0
вручение	0	0	0,26	0	0,38	0	0	0	0,52
…									

Тематическое моделирование

Латентно-семантический анализ

Выбираем количество тем и выполняем сингулярное разложение матрицы.

апелляция	0,57	-0,01	0,01
арестованный	0,34	0	0,07
арестовать	0,34	0	0,01
балтия	0	0,12	-0,62
бойкотировать	0	0,52	-0,21
британский	0,57	-0,01	-0,18
великобритания	0,31	0,05	0,15
вручение	0,02	0,67	0,09
…			

матрица U

3,41	0	0
0	3,3	0
0	0	2,27

матрица Σ

d1	d2	d3	d4	d5	d6	d7	d8	d9
0,63	0,05	0,01	0,54	0	0,47	0,01	0,63	0
0	0,02	0,65	-0,01	0,59	0	0,09	-0,01	0,48
0,03	-0,7	-0,04	0,06	0,1	-0,16	-0,67	0,09	0,09

матрица V^T

Тематическое моделирование

Латентно-семантический анализ

Интерпретируем результаты

	Темы		
	1	2	3
апелляция	0,57	-0,01	0,01
арестованный	0,34	0	0,07
арестовать	0,34	0	0,01
балтия	0	0,12	-0,62
бойкотировать	0	0,52	-0,21
британский	0,57	-0,01	-0,18
великобритания	0,31	0,05	0,15
вручение	0,02	0,67	0,09
…			

матрица U

3,41	0	0
0	3,3	0
0	0	2,27

матрица Σ

d1	d2	d3	d4	d5	d6	d7	d8	d9	
0,63	0,05	0,01	0,54	0	0,47	0,01	0,63	0	Тема 1
0	0,02	0,65	-0,01	0,59	0	0,09	-0,01	0,48	Тема 2
0,03	-0,7	-0,04	0,06	0,1	-0,16	-0,67	0,09	0,09	Тема 3

матрица V^T

Тематическое моделирование

Латентно-семантический анализ

Проверяем

	d1	d2	d3	d4	d5	d6	d7	d8	d9
апелляция,арестованный, арестовать, британский, великобритания	1	0	0	1	0	1	0	1	0
бойкотировать, вручение	0	0	1	0	1	0	0	0	1
балтия, США	0	1	0	0	0	0	1	0	0

1. Британская полиция знает о местонахождении основателя WikiLeaks.
2. В суде США начинается процесс против россиянина, рассылавшего спам.
3. Церемонию вручения Нобелевской премии мира бойкотируют 19 стран.
4. В Великобритании арестован основатель сайта Wikileaks Джулиан Ассандж.
5. Украина игнорирует церемонию вручения Нобелевской премии.
6. Шведский суд отказался рассматривать апелляцию основателя Wikileaks.
7. НАТО и США разработали планы обороны стран Балтии против России.
8. Полиция Великобритании нашла основателя WikiLeaks, но не арестовала.
9. В Стокгольме и Осло сегодня состоится вручение Нобелевских премий.

Тематическое моделирование

Латентное размещение Дирихле

А что если распределение тем по документам и слов по темам смоделировать распределением Дирихле?..

Получится смоделировать много тем с условием: в каждом документе не более 3-5 тем.

- Немного о работе LDA на пальцах
- Чуть больше и с примерами кода – здесь есть ссылки на подробные статьи с формулами

Тематическое моделирование

Библиотеки

gensim – topic modelling for humans.
sklearn – topic modelling with LDA.

Классификация текстов

Классификация текстов

Анализ тональности
Классификация названий медицинских услуг
Оценка свободных ответов учеников

Классификация текстов

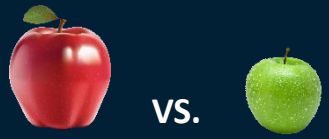
Анализ тональности
Классификация названий медицинских услуг
Оценка свободных ответов учеников

Можно обучать классификаторы, использовать правила или совмещать.

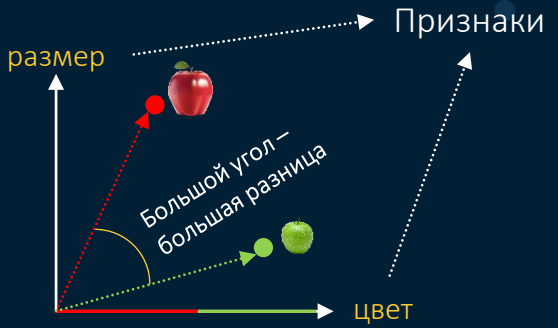
Классификация текстов

С помощью машинного обучения

Как сравнить 2 яблока?



С помощью векторов



Как сравнить тексты?

1 балл

Текст 1
Подъем аэростата прекратится, когда архимедова сила станет меньше силы тяжести

Текст 2
Когда подъемная сила будет равна силе тяжести шара

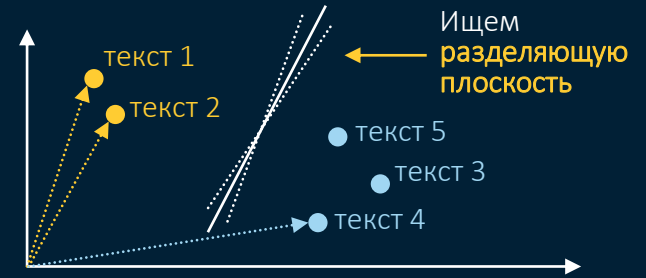
0 баллов

Текст 3
Когда закончутся баласты

Текст 4
Если в шаре не будет газа

Текст 5
Когда воздух будет выше 100 градусов

С помощью векторов



Признаки неинтерпретируемые.
Обычно 200-700

Классификация текстов

С помощью машинного обучения

Верхнеуровневое описание подхода:

Текст → Вектор → Классификатор → Категория

Классификация текстов

С помощью машинного обучения

Верхнеуровневое описание подхода:

Текст → Вектор → Классификатор → Категория

Как получить вектор текста?

Классификация текстов

Вектора текстов с помощью SVD

апелляция	0,57	-0,01	0,01
арестованный	0,34	0	0,07
арестовать	0,34	0	0,01
балтия	0	0,12	-0,62
бойкотировать	0	0,52	-0,21
британский	0,57	-0,01	-0,18
великобритания	0,31	0,05	0,15
вручение	0,02	0,67	0,09
…			

матрица U

3,41	0	0
0	3,3	0
0	0	2,27

матрица Σ

d1	d2	d3	d4	d5	d6	d7	d8	d9
0,63	0,05	0,01	0,54	0	0,47	0,01	0,63	0
0	0,02	0,65	-0,01	0,59	0	0,09	-0,01	0,48
0,03	-0,7	-0,04	0,06	0,1	-0,16	-0,67	0,09	0,09

матрица V^T

Классификация текстов

Вектора текстов с помощью SVD

Берем матрицу V

d1	d2	d3	d4	d5	d6	d7	d8	d9
0,63	0,05	0,01	0,54	0	0,47	0,01	0,63	0
0	0,02	0,65	-0,01	0,59	0	0,09	-0,01	0,48
0,03	-0,7	-0,04	0,06	0,1	-0,16	-0,67	0,09	0,09

матрица V^T

Классификация текстов

Вектора текстов с помощью SVD

Берем матрицу V

d1	0,63	0	0,03
d2	0,05	0,02	-0,7
d3	0,01	0,65	-0,04
d4	0,54	-0,01	0,06
d5	0	0,59	0,1
d6	0,47	0	-0,16
d7	0,01	0,09	-0,67
d8	0,63	-0,01	0,09
d9	0	0,48	0,09

матрица V

Классификация текстов

Вектора текстов с помощью SVD

Берем матрицу V и 300 измерений

d1	0,63	0	0,03	…	…	…	…	…
d2	0,05	0,02	-0,7	…	…	…	…	…
d3	0,01	0,65	-0,04	…	…	…	…	…
d4	0,54	-0,01	0,06	…	…	…	…	…
d5	0	0,59	0,1	…	…	…	…	…
d6	0,47	0	-0,16	…	…	…	…	…
d7	0,01	0,09	-0,67	…	…	…	…	…
d8	0,63	-0,01	0,09	…	…	…	…	…
d9	0	0,48	0,09	…	…	…	…	…

вектор текста

матрица V

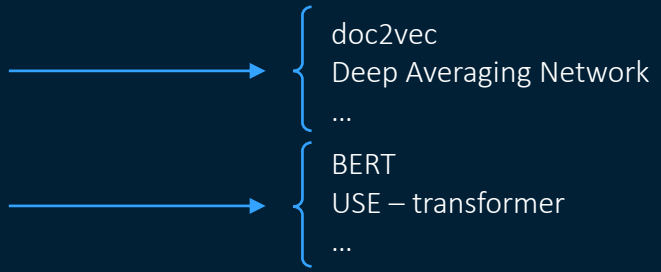
Классификация текстов

Как еще получают вектора для текстов

"I am a sentence for which I would like to get its embedding"

Виды магии:

- Bag of Words – не учитывает порядок слов в предложении (различные виды усреднения). Относительно быстро учится и получает вектора.
- Синтаксическая. Требует большой обучающей выборки, много времени на обучение, работает только с небольшими текстами. Зато качество state-of-the-art.



Классификация текстов

Universal Sentence Encoder

```
In [1]: import tensorflow_hub as hub
        import numpy as np
        import tensorflow_text

In [2]: embed = hub.load("https://tfhub.dev/google/universal-sentence-encoder-multilingual/3")

In [3]: sentence = 'Хочу получить вектор этого предложения.'

In [4]: result = embed(sentence)

In [5]: result

Out[5]: <tf.Tensor: shape=(1, 512), dtype=float32, numpy=
        array([[-0.03499177, -0.00310049,  0.00212201, -0.01303443, -0.11130663,
                -0.005326  , -0.02134681,  0.02525296, -0.07166617, -0.02084302,
                 0.0230692 , -0.02330862,  0.07504725,  0.06589342, -0.04158042,
                 0.06805082,  0.05940025, -0.01126915, -0.0086282 ,  0.04886374,
                -0.0514462 , -0.00843818,  0.06717453,  0.07215355, -0.06209831,
                -0.05632949,  0.07022848, -0.01598333,  0.09836854, -0.08614098,
                 0.0044453 , -0.00754533,  0.02111935, -0.03690185, -0.01194055,
                -0.0728237 , -0.04036488, -0.06009685, -0.05511959, -0.04890082,
                 0.0063697 ,  0.03101335, -0.08288708,  0.0157161 , -0.03562398,
```

Классификация текстов

С помощью машинного обучения

Верхнеуровневое описание подхода:

Текст → Вектор → Классификатор → Категория

```
In [ ]: svm = SVC(kernel='poly', gamma='scale', probability=True)

In [ ]: %%time
        svm.fit(train_embeddings, train_s[class_var])

In [ ]: classesSVM = svm.predict(test_embeddings)
        scoresSVM = svm.predict_proba(test_embeddings)
```

Начинайте с линейных
Naïve Bayes
Support Vector Machine
Linear Regression

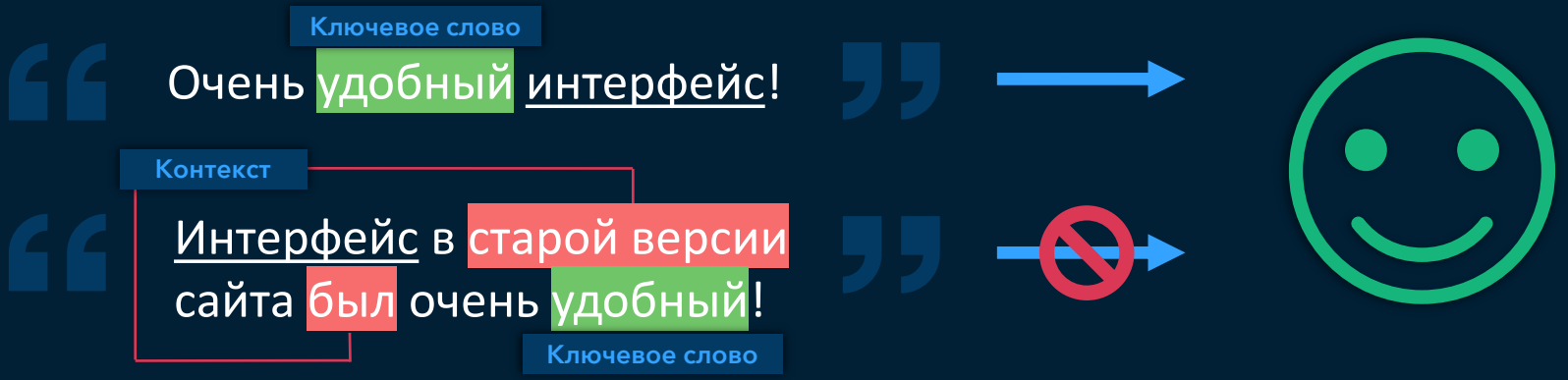
Чтобы повысить точность – увеличивайте порог уверенности модели, что текст принадлежит данной категории (cutoff):

- Класс 0: 51%, класс 1: 49%. – Ненадежный результат.
- Класс 0: 80%, класс 1: 20%. – Вероятность ошибки меньше.

Классификация текстов

С помощью правил

Правила – не просто ключевые слова



Классификация текстов

С помощью правил

SpaCy поддерживает правила для извлечения паттернов.

Поверх извлеченных паттернов можно настроить правила вида «если – то».

Если в документе присутствует паттерн «удобный интерфейс» и не присутствует паттерн «в старой версии был удобный интерфейс», то присвоить тексту категорию Positive.

```
import spacy
from spacy.matcher import Matcher

nlp = spacy.load("en_core_web_sm")
matcher = Matcher(nlp.vocab)
# Add match ID "HelloWorld" with no callback and one pattern
pattern = [{"LOWER": "hello"}, {"IS_PUNCT": True}, {"LOWER": "world"}]
matcher.add("HelloWorld", None, pattern)

doc = nlp("Hello, world! Hello world!")
```

Классификация текстов

Гибридный подход

Обучаем классификатор и повышаем cutoff.
Для сложных случаев пишем правила.

Используем классификатор для категоризации.
Используем правила для извлечения сущностей.

Практикум по тематическому моделированию и классификации

Извлечение именованных сущностей и фактов

Извлечение именованных сущностей

Примеры задач

Найти в корпусе текстов все имена
Найти все адреса электронной почты
Найти мнения пользователей о новом интерфейсе
Извлечь сумму дивидендов по разным типам акций
Извлечь срок действия договора

Извлечение именованных сущностей

Являюсь клиентом уже 11 лет, раньше все устраивало с мобильной связью, но последний год то и дело звоню в поддержку и ругаюсь. У меня подключен тариф Выгодный с без лимитным интернетом. Интернет еле работает везде, периодически не удается дозвониться другим абонентам. Самое что интересное, вместо 850р я получаю каждый месяц счета от 1700р до 2500, заграницу я не звоню, подключены две услуги, общей стоимость 160р, но это никак не выходит 2000 в месяц!!!! Открыла детализацию и практически все услуги 0р., за исключением 2 операций в роуминге (максимум +300р), но это опять же не тот счет, за который я плачу! Окончательной каплей была моя поездка в Америку, когда телефон работал первые два дня, а потом есть сеть, но не могу никому набрать, но самое главное не доходят смс. Из Москвы звонили в поддержку, т.к. мой телефон даже туда не соединял, мое маме сказали, что возможно это из-за подключенной услуги антиАОН, они мне ее отключили, телефон и правда начал дозваниваться, но смс так и не приходили, что было огромной проблемой, мы не могли купить билеты на самолет, покрыть кредитную карту и т.д., пришлось пользоваться услугами родителей из Москвы. Я даже не знаю, что бы мы делали если бы никто не мог нам помочь финансово!! Смс начали снова приходить опять в последние 2 дня отпуска. Вернулась в Москву, а тут интернет снова не работает, оператор говорит лишь перезагрузите айфон. Вчера пошла и перевелась на другого провайдера, надоело бороться со связью. Что касается ТВ: одна приставка работает хорошо, вторая вечно просит перезагрузить, не работает, запрашивает какой то пароль, в службе поддержки пинают от оператора к операторы, по две-три недели не приходит сотрудник. Домашний интернет тоже периодически не работает, но с ним дела обстоят намного лучше, я бы даже сказала неплохо, относительно всех остальных услуг. Желаю Оператору настроить свои проблемы со связью и интернетом, в противном случае потеряете всех клиентов!

Жалоба на оператора сотовой связи, размещенная в публичном доступе на сайте banki.ru

Проблема:
невозможно выделить значимую информацию из «простыни» текста

Извлечение именованных сущностей

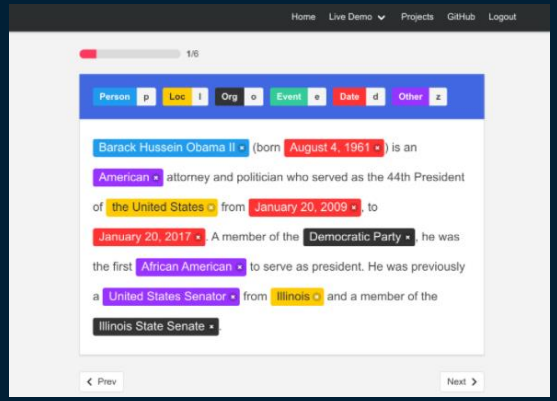
Являюсь клиентом уже 11 лет, раньше все устраивало с мобильной связью, но последний год то и дело звоню в поддержку и ругаюсь. У меня подключен тариф Выгодный с без лимитным интернетом. Интернет еле работает везде, периодически не удается дозвониться другим абонентам. Самое что интересное, вместо 850р я получаю каждый месяц счета от 1700р до 2500, заграницу я не звоню, подключены две услуги, общей стоимость 160р, но это никак не выходит 2000 в месяц!!!! Открыла детализацию и практически все услуги 0р., за исключением 2 операций в роуминге (максимум +300р), но это опять же не тот счет, за который я плачу! Окончательной каплей была моя поездка в Америку, когда телефон работал первые два дня, а потом есть сеть, но не могу никому набрать, но самое главное не доходят смс. Из Москвы звонили в поддержку, т.к. мой телефон даже туда не соединял, мое маме сказали, что возможно это из-за подключенной услуги антиАОН, они мне ее отключили, телефон и правда начал дозваниваться, но смс так и не приходили, что было огромной проблемой, мы не могли купить билеты на самолет, покрыть кредитную карту и т.д., пришлось пользоваться услугами родителей из Москвы. Я даже не знаю, что бы мы делали если бы никто не мог нам помочь финансово!! Смс начали снова приходить опять в последние 2 дня отпуска. Вернулась в Москву, а тут интернет снова не работает, оператор говорит лишь перезагрузите айфон. Вчера пошла и перевелась на другого провайдера, надоело бороться со связью. Что касается ТВ: одна приставка работает хорошо, вторая вечно просит перезагрузить, не работает, запрашивает какой то пароль, в службе поддержки пинают от оператора к операторы, по две-три недели не приходит сотрудник. Домашний интернет тоже периодически не работает, но с ним дела обстоят намного лучше, я бы даже сказала неплохо, относительно всех остальных услуг. Желаю Оператору настроить свои проблемы со связью и интернетом, в противном случае потеряете всех клиентов!

- Интернет: негатив
- Связь: негатив
- Общая негативная оценка
- Контакт-центр: проблема
- Брак продукта
- Сотрудники: негатив
- Обращение к детализации
- Коммуникация: КЦ
- Смена оператора
- Давний клиент

Извлечение именованных сущностей

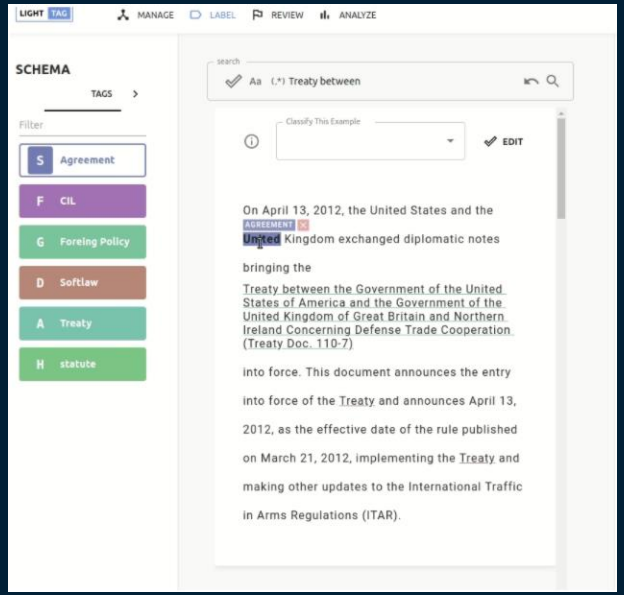
Инструменты разметки

[Doccano]



Бесплатный, не поддерживает кириллицу

[LightTag]



Платный, поддерживает кириллицу

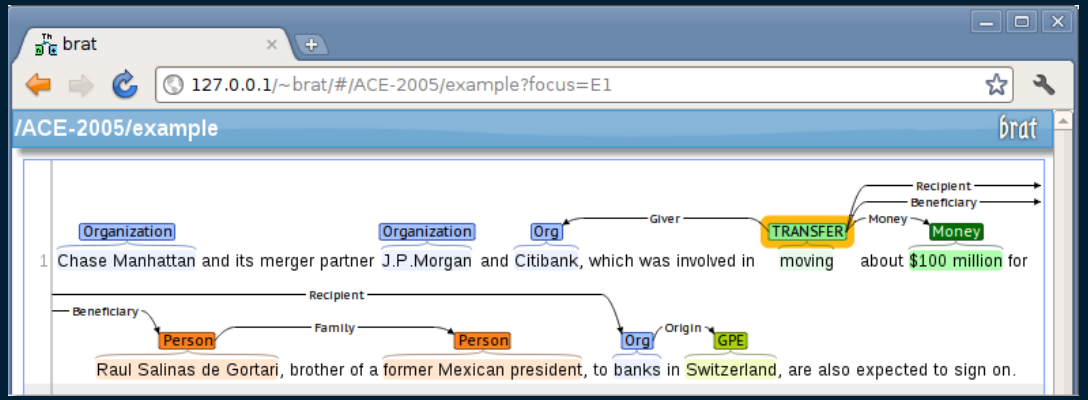
Под капотом у обоих json:

```
{
 “id”: 1,
 “text”: ”On April 13, 2012 …”,
 “labels”: [
  [0,16,”Date”],
  [25,34,”Location”]
 ]
}
```

Извлечение именованных сущностей

Инструменты разметки

Brat бесплатный, поддерживает кириллицу.



Examples of annotation for an entity (T1), an event trigger (T2), an event (E1) and a relation (R1) are shown in the following.

```
T1      Organization 0 4                    Sony
T2      MERGE-ORG 14 27                     joint venture
T3      Organization 33 41                  Ericsson
E1      MERGE-ORG:T2 Org1:T1 Org2:T3
T4      Country 75 81                       Sweden
R1      Origin Arg1:T3 Arg2:T4
```

← Под капотом свой формат файлов .ann

Извлечение именованных сущностей

Предобученные модели

DeepPavlov

3 предобученные модели с разной архитектурой:

- ner_rus_bert – state-of-the-art, BERT с опциональным CRF-слоем. F1 = 98,1.
- ner_rus – Bi-LSTM + CRF. F1 = 95,1. Зато быстрее и легче остальных.
- ner_collection3_m1 – LSTM. F1 = 97,8.

SlovNet

Извлекает стандартные сущности: PER, LOC, ORG.
Качество на 1-2% хуже чем ner_rus_bert, но весит в 60 раз меньше и работает быстрее.
Обучена на корпусе новостей, поэтому в другом домене работает хуже.

Извлечение именованных сущностей

Подход на правилах

SpaCy = паттерны либо прямо в коде, либо отдельными файлами. Библиотека Python.

```
import spacy
from spacy.matcher import Matcher

nlp = spacy.load("en_core_web_sm")
matcher = Matcher(nlp.vocab)
# Add match ID "HelloWorld" with no callback and one pattern
pattern = [{"LOWER": "hello"}, {"IS_PUNCT": True}, {"LOWER": "world"}]
matcher.add("HelloWorld", None, pattern)

doc = nlp("Hello, world! Hello world!")
```

Извлечение именованных сущностей

Подход на правилах

Tomita Parser = словари и грамматики отдельными файлами + консоль.
Написан на C++.

```
#encoding "utf8"

StreetW -> 'проспект' | 'проезд' | 'улица' | 'шоссе';
StreetSokr -> 'пр' | 'просп' | 'пр-д' | 'ул' | 'ш';

StreetDescr -> StreetW | StreetSokr;

StreetNameNoun -> (Adj<gnc-agr[1]>) Word<gnc-agr[1],rt> (Word<gram="род">);

StreetNameAdj -> Adj<h-reg1> Adj*;

Street -> StreetDescr interp (Address.Descr) StreetNameNoun<gram="род", h-reg1> interp
 (Address.StreetName);
Street -> StreetDescr interp (Address.Descr) StreetNameNoun<gram="им", h-reg1> interp
 (Address.StreetName);
```

Извлечение именованных сущностей

Подход на правилах

Yargy-parser = опенсорсная Томита на Python.

Томита-парсер	Yargy
Разрабатывался много лет внутри Яндекса	Open source, разрабатывается сообществом
10 000+ строк кода на C++	1000+ на Python
CLI	Python-библиотека
Protobuf + конфигурационные файлы	Python DSL
Нет готовых правил	Natasha — готовые правила для извлечения имён, дат, адресов и других сущностей
Медленный	Очень медленный

```
GEO = rule(
    and_(
        gram('ADJF'),  # так помечается прилагательное, остальные пометки описаны в
                       # http://pymorphy2.readthedocs.io/en/latest/user/grammemes.html
        is_capitalized()
    ),
    gram('ADJF').optional().repeatable(),
    dictionary({
        'федерация',
        'республика'
    })
)
```

Анализ юридических документов

Сколько договоров заключает компания?

Крупная промышленная компания X занимается производством удобрений. Удобрения получают из полезных ископаемых. Полезные ископаемые добывают на руднике.

Рудником владеет компания Y. Она добывает руду. Руда – это смесь обычных камней и полезных. Руду надо обогатить – избавиться от обычных камней и оставить только полезные. Руду обогащают на фабрике.

У компании Z есть обогатительная фабрика. Но руду из рудника на фабрику надо как-то доставить. Этим занимается транспортная компания N.

Почему нельзя сразу добывать полезные камни →

Закрытая разработка

Приземистый погрузчик, который работает в шахте

Машина, которая режет гору

В шахте никто не будет ковыряться и отбирать только полезные камни.
Там буквально режут гору и вывозят всё сразу, чтобы не мешало.

Открытая разработка

БелАЗ – самосвал для работы в карьере

Карьерные экскаваторы размером с БелАЗ

Чтобы добыть в карьере полезные камни, породу взрывают. Получается куча полезных и обычных камней. В карьере работают огромные машины. При таких масштабах никто не будет ковыряться в куче камней и отбирать только полезные.

Сколько договоров заключает компания?

Крупная промышленная компания X занимается производством удобрений. Удобрения получают из полезных ископаемых. Полезные ископаемые добывают на руднике.

Рудником владеет компания Y. Она добывает руду. Руда – это смесь обычных камней и полезных. Руду надо обогатить – избавиться от обычных камней и оставить только полезные. Руду обогащают на фабрике.

У компании Z есть обогатительная фабрика. Но руду из рудника на фабрику надо как-то доставить. Этим занимается транспортная компания N.

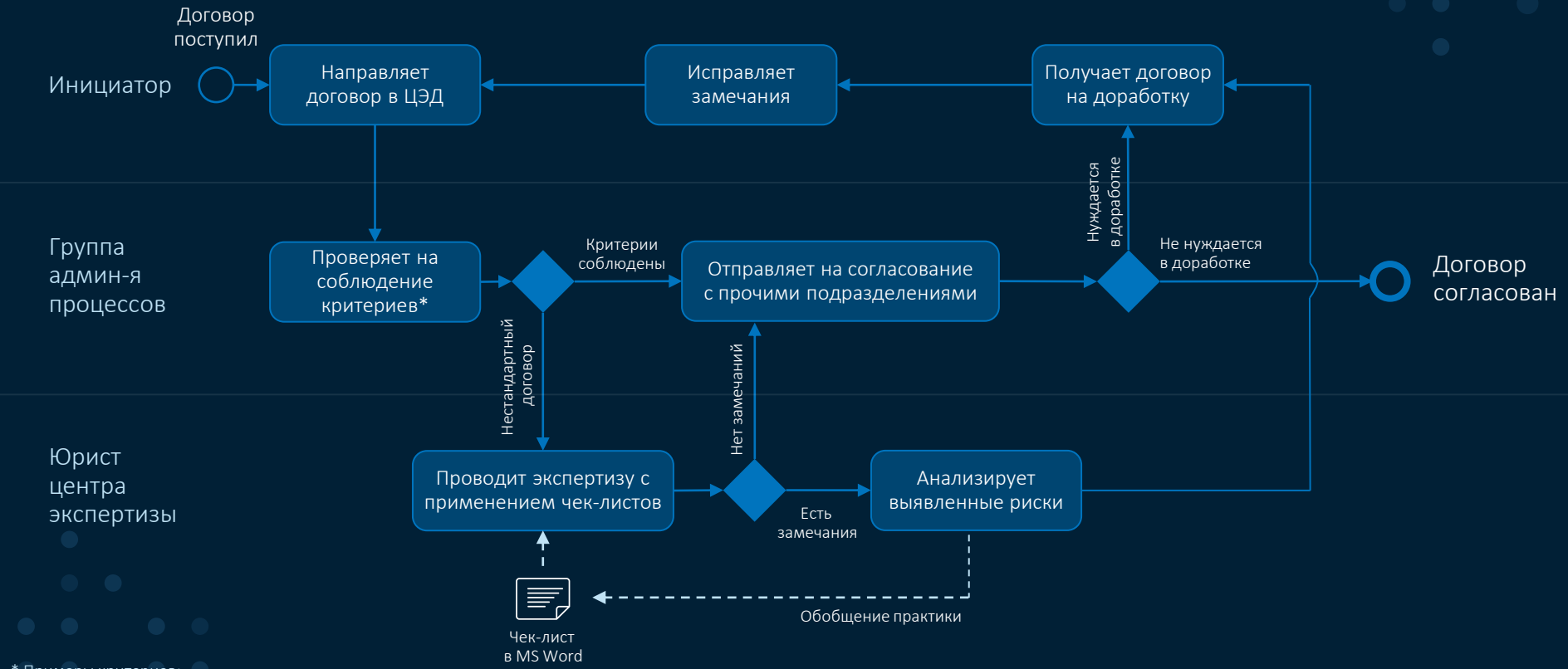
Чтобы компания X могла производить удобрения, ей нужно заключить договор на покупку руды с компанией Y, договор на обогащение руды с компанией Z и договор на перевозку сырой и обогащенной руды с компанией N. Потом компания X заключает договоры на поставку готовых удобрений покупателям.

И это еще далеко не все договоры.

Каждый договор проверяют юристы.

А мог бы искусственный интеллект

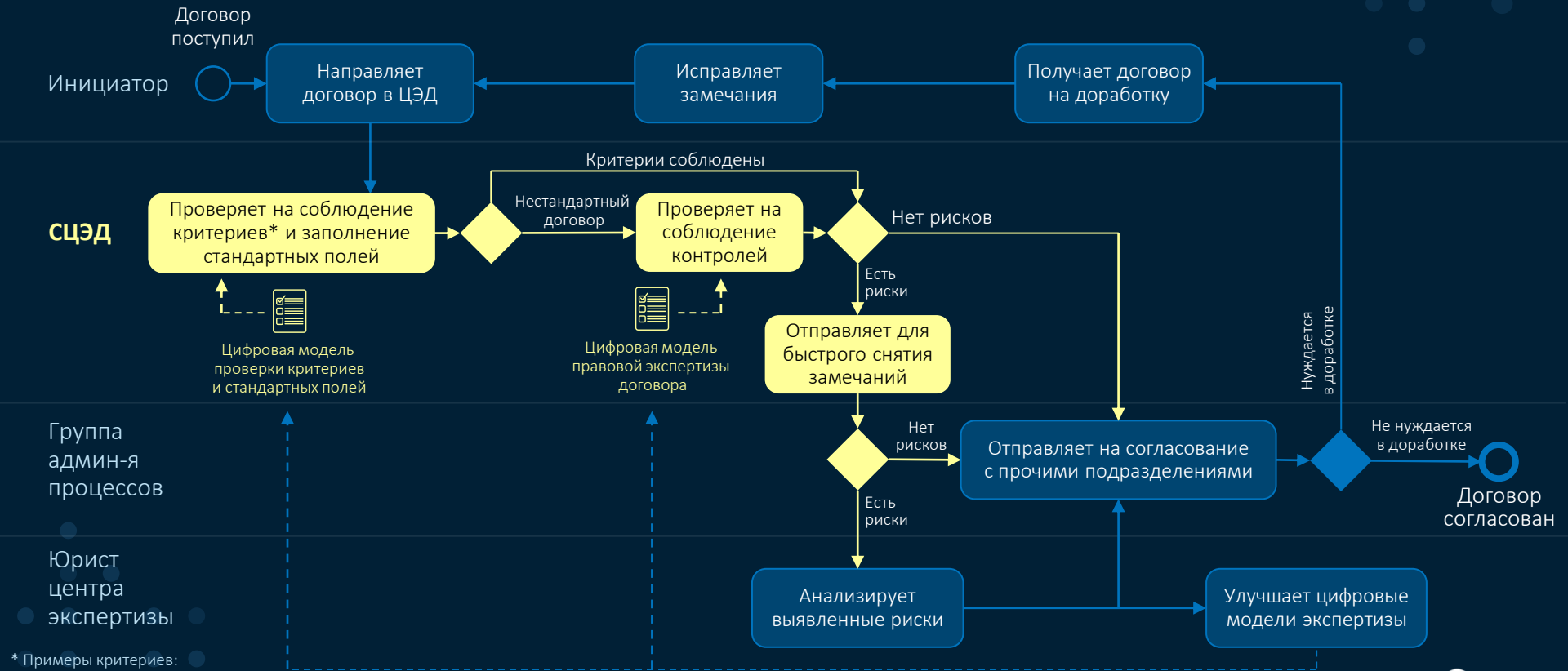
Текущий процесс работы центра экспертизы договоров



* Примеры критериев:

- Договор оформлен по единой форме
- Сумма договора не превышает X млн. руб.
- Контрагент – из группы компаний

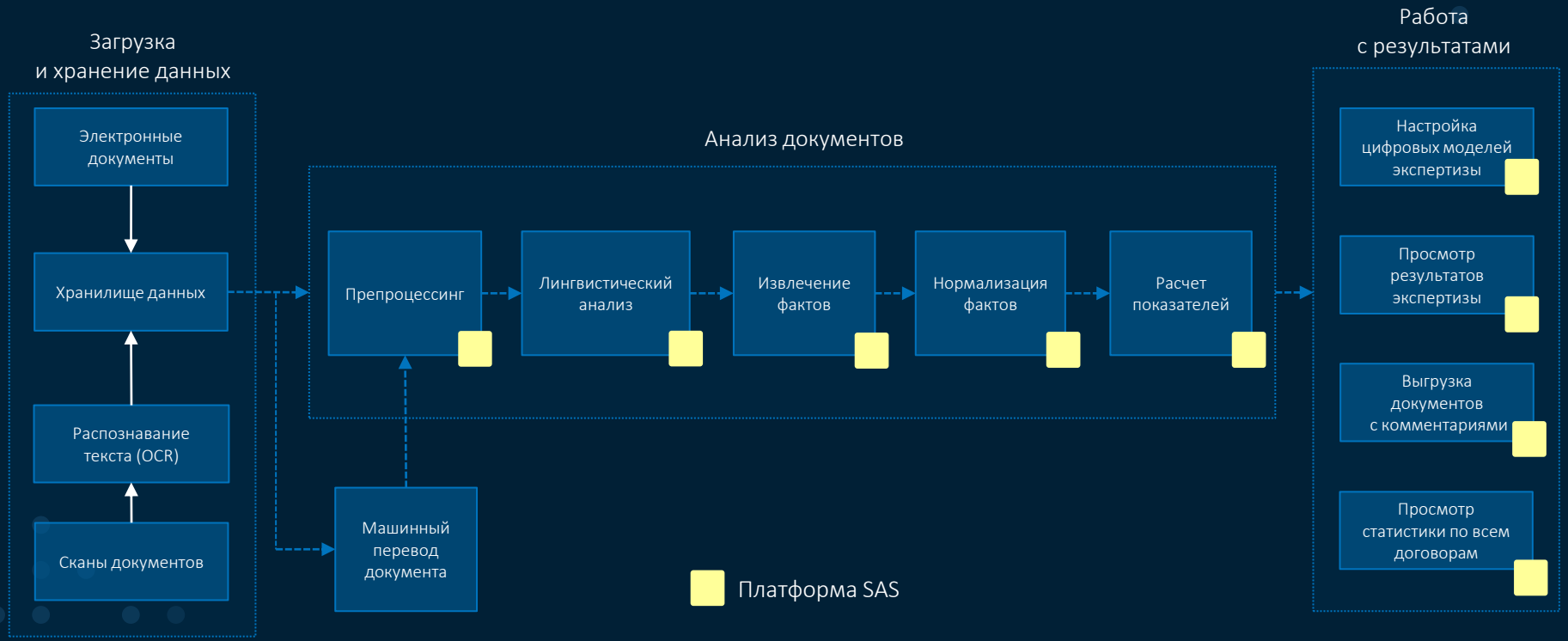
Целевой автоматизированный процесс работы ЦЭД



* Примеры критериев:

- Договор оформлен по единой форме
- Сумма договора не превышает X млн. руб.
- Контрагент – из группы компаний

Предлагаемая концептуальная архитектура решения



Система цифровой экспертизы договоров на основе SAS позволит

- Сконцентрироваться на стандартизации методологии правовой экспертизы вместо рутинных проверок
- Сократить трудозатраты на экспертизу договоров
- Ускорить процесс экспертизы и согласования договоров
- Снизить влияние человеческого фактора на результат экспертизы
- Выявлять типовые риски для улучшения конструктора договоров

— Система автоматически проверяет каждый договор и выявляет риски

— Экспертиза с участием ЮП нужна только при наличии рисков

— Система проверяет каждый договор за несколько секунд

— Договоры проверяет беспристрастный, неустающий искусственный интеллект

— Система хранит и обобщает результаты экспертизы всех договоров

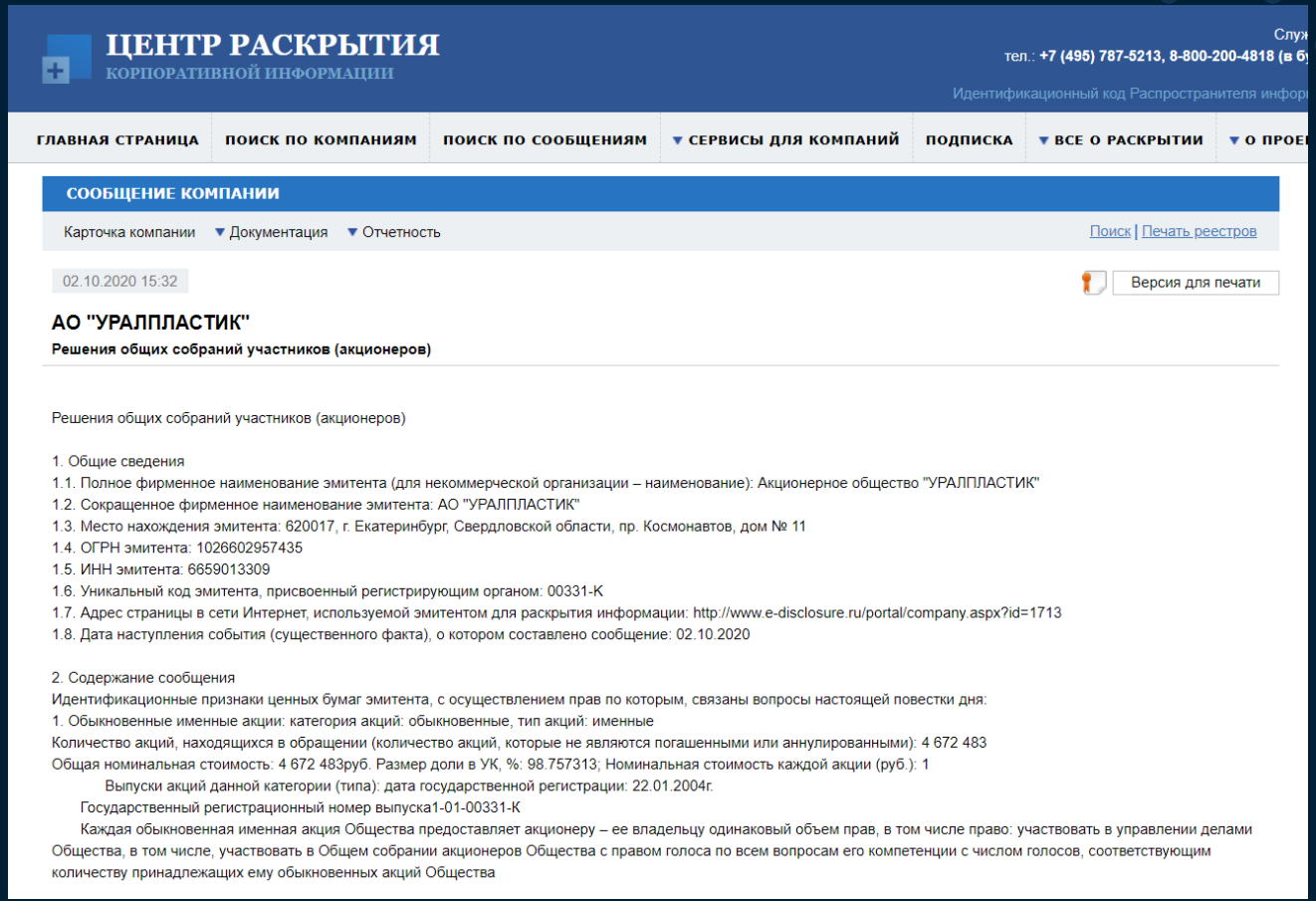
Анализ протоколов собраний акционеров

Что это такое?

Примерно раз в квартал акционеры крупных компаний собираются и решают вопросы:

- как распределить прибыль компании
- кого избрать в совет директоров
- кого избрать в ревизионную комиссию
- кого назначить аудитором
- утвердить ли бухгалтерскую и финансовую отчетность
- и др.

Свои решения они оформляют протоколом и публикуют в открытом доступе. Так обязывает закон.



ЦЕНТР РАСКРЫТИЯ
КОРПОРАТИВНОЙ ИНФОРМАЦИИ

Слу
тел.: +7 (495) 787-5213, 8-800-200-4818 (в б
Идентификационный код Распространителя инфор

ГЛАВНАЯ СТРАНИЦА | ПОИСК ПО КОМПАНИЯМ | ПОИСК ПО СООБЩЕНИЯМ | ▼ СЕРВИСЫ ДЛЯ КОМПАНИЙ | ПОДПИСКА | ▼ ВСЕ О РАСКРЫТИИ | ▼ О ПРОЕ

СООБЩЕНИЕ КОМПАНИИ

Карточка компании ▼ Документация ▼ Отчетность Поиск | Печать реестров

02.10.2020 15:32

Версия для печати

АО "УРАЛПЛАСТИК"

Решения общих собраний участников (акционеров)

Решения общих собраний участников (акционеров)

1. Общие сведения
1.1. Полное фирменное наименование эмитента (для некоммерческой организации – наименование): Акционерное общество "УРАЛПЛАСТИК"
1.2. Сокращенное фирменное наименование эмитента: АО "УРАЛПЛАСТИК"
1.3. Место нахождения эмитента: 620017, г. Екатеринбург, Свердловской области, пр. Космонавтов, дом № 11
1.4. ОГРН эмитента: 1026602957435
1.5. ИНН эмитента: 6659013309
1.6. Уникальный код эмитента, присвоенный регистрирующим органом: 00331-K
1.7. Адрес страницы в сети Интернет, используемой эмитентом для раскрытия информации: http://www.e-disclosure.ru/portal/company.aspx?id=1713
1.8. Дата наступления события (существенного факта), о котором составлено сообщение: 02.10.2020

2. Содержание сообщения
Идентификационные признаки ценных бумаг эмитента, с осуществлением прав по которым, связаны вопросы настоящей повестки дня:
1. Обыкновенные именные акции: категория акций: обыкновенные, тип акций: именные
Количество акций, находящихся в обращении (количество акций, которые не являются погашенными или аннулированными): 4 672 483
Общая номинальная стоимость: 4 672 483руб. Размер доли в УК, %: 98.757313; Номинальная стоимость каждой акции (руб.): 1
Выпуски акций данной категории (типа): дата государственной регистрации: 22.01.2004г.
Государственный регистрационный номер выпуска1-01-00331-K
Каждая обыкновенная именная акция Общества предоставляет акционеру – ее владельцу одинаковый объем прав, в том числе право: участвовать в управлении делами Общества, в том числе, участвовать в Общем собрании акционеров Общества с правом голоса по всем вопросам его компетенции с числом голосов, соответствующим количеству принадлежащих ему обыкновенных акций Общества

Кому это может быть интересно?

Аналитическим и рейтинговым агентствам

По информации из протоколов, отчетов, эмиссионных документов и др. они делают выводы о стабильности компаний, ищут аффилированных лиц, присваивают кредитные рейтинги.

Юридическим департаментам крупных банков

Когда компания подает заявку на кредит, юристы банка проверяют в том числе протоколы собраний акционеров. Они ищут там аффилированных лиц, решения об одобрении крупных сделок, условия этих сделок и др.

Практикум по извлечению именованных сущностей и фактов

Вопросы можете задавать в Telegram:

@pyatov – Алексей Пятов
@kondud – Константин Дудников

sas.com