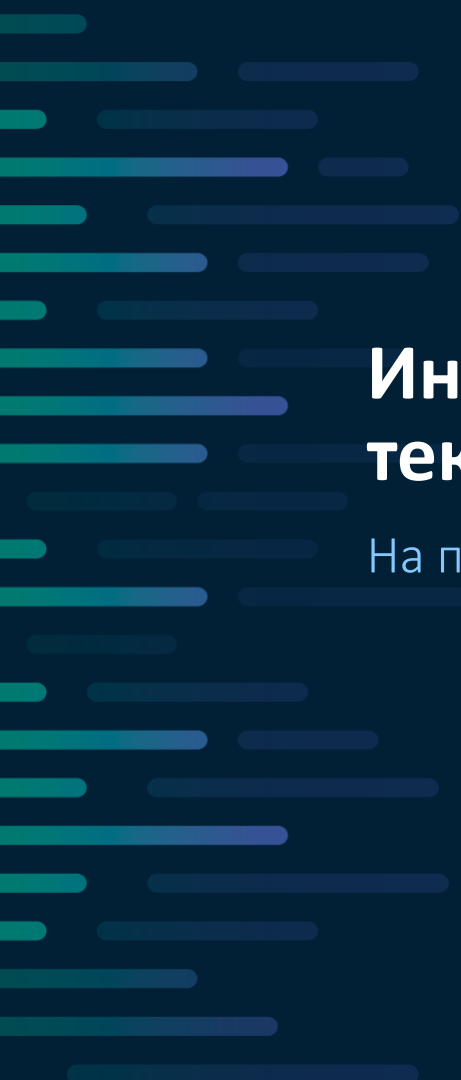


Текстовая аналитика в бизнесе

На платформе SAS и не только

Алексей Пятов // Руководитель практики управления данными SAS Russia
Константин Дудников // Руководитель группы текстовой аналитики SAS Russia

211125

A series of horizontal bars of varying lengths and colors (teal, blue, and dark blue) are arranged in a staggered, overlapping pattern on the left side of the slide.

Инструменты и методы текстовой аналитики

На пальцах

План

1. Обзор свободного ПО
2. Практикум по использованию свободного ПО
3. Обзор enterprise-решений
4. Демонстрация работы с платформой SAS Viya

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 скобок – 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 термины:

- 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
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

Тема 1

Тема 2

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.

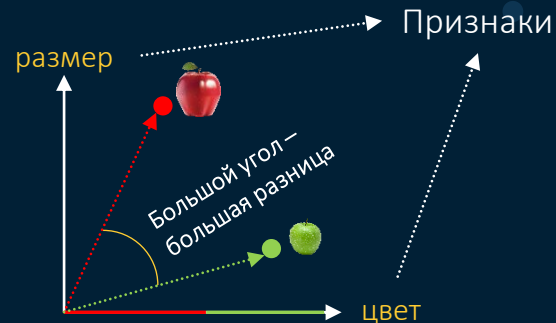
Классификация текстов

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 закончатся баласты

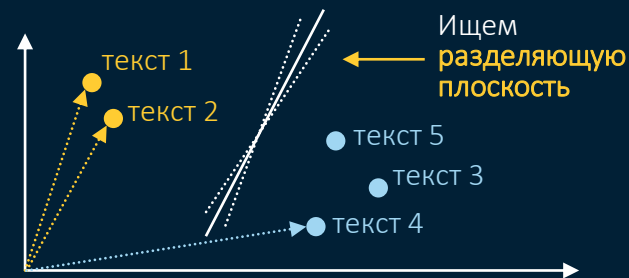
Текст 4

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

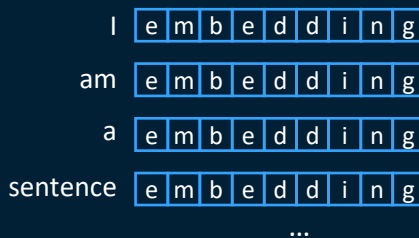
• • •

матрица V

Классификация текстов

Как еще получают вектора для текстов

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



Магия



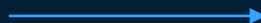
[0,7764657876, -0,09763543, ...]
n dimensions

Виды магии:

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



{ doc2vec
Deep Averaging Network
...



{ BERT
USE – transformer
...

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,
         0.02306092, -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

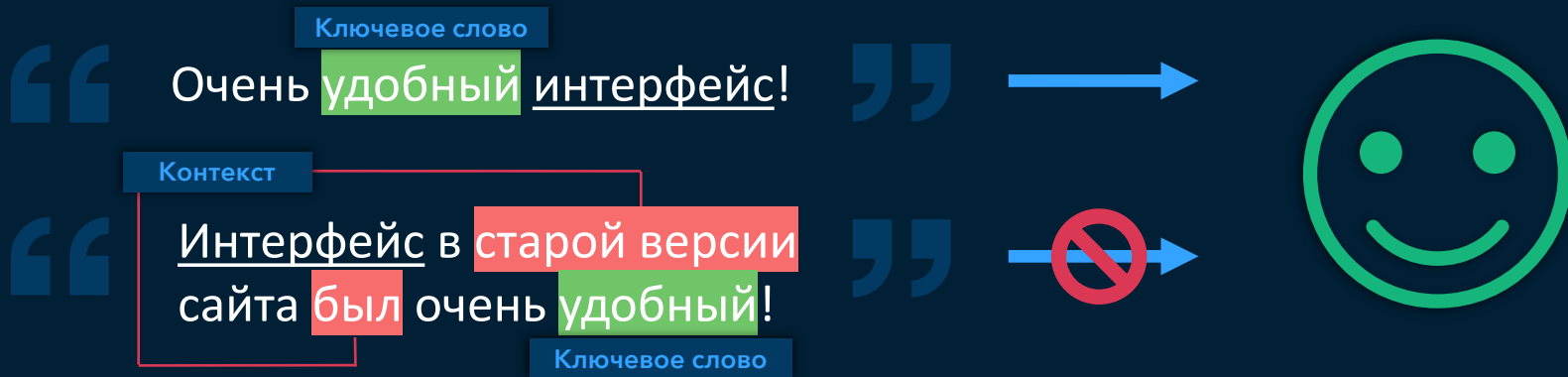
Чтобы повысить точность — увеличивайте порог уверенности модели, что текст принадлежит данной категории (cutoff):

- 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.

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, за границу я не звоню, подключены две услуги, общей стоимостью 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 оператору, 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, границу я не звоню, подключены две услуги, общей стоимостью 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

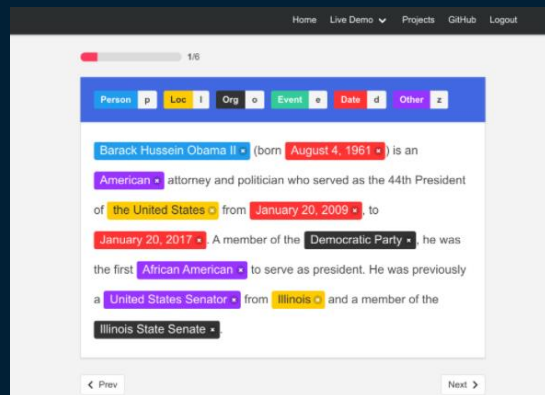
Смена оператора

Давний клиент

Извлечение именованных сущностей

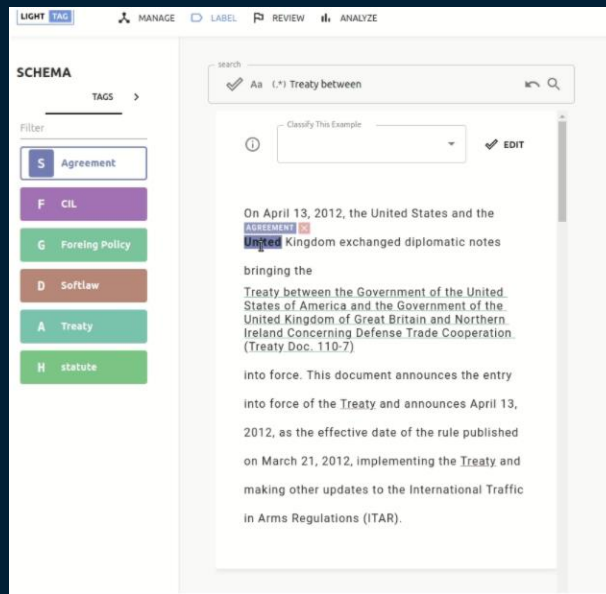
Инструменты разметки

[Doccano](#)



Бесплатный, не
поддерживает кириллицу

[LightTag](#)



Платный, поддерживает кириллицу

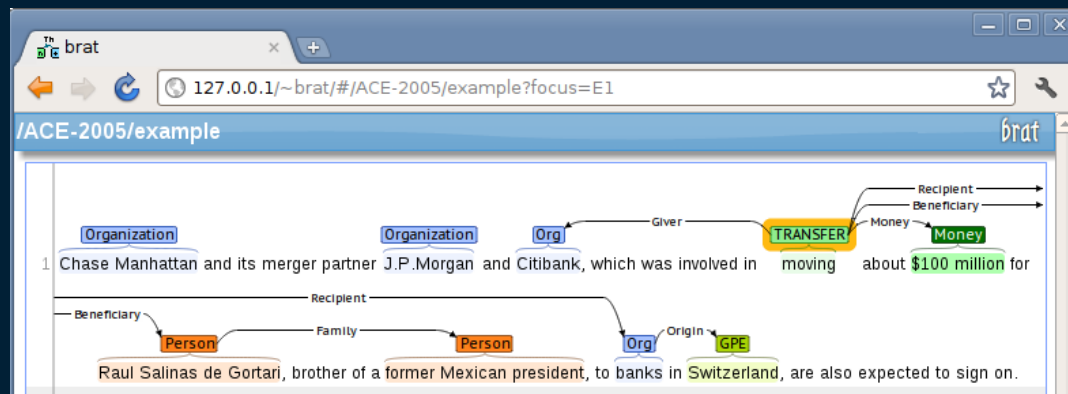
Под капотом у обоих json:

```
{
  "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++.

```
1 #encoding "utf8"
2
3 StreetW -> 'проспект' | 'проезд' | 'улица' | 'шоссе';
4 StreetSokr -> 'пр' | 'просп' | 'пр-д' | 'ул' | 'ш';
5
6 StreetDescr -> StreetW | StreetSokr;
7
8 StreetNameNoun -> (Adj<gnc-agr[1]>) Word<gnc-agr[1],rt> (Word<gram="род">);
9
10 StreetNameAdj -> Adj<h-reg1> Adj*;
11
12 Street -> StreetDescr interp (Address.Descr) StreetNameNoun<gram="род", h-reg1> interp
13   · (Address.StreetName);
14 Street -> StreetDescr interp (Address.Descr) StreetNameNoun<gram="им", h-reg1> interp
15   · (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'
    })
)
```

Open-source практикум

На рисунках представлены изображения различных явлений.



Движение автомобиля
по закруглённой дороге



Излучение света раскалённым
металлом



Скисание молока
в тёплой комнате



Появление радуги после дождя

Какое из представленных явлений является химическим? Свой ответ поясните.

Задача: проверять ответы на подобные вопросы автоматически.

Что делать?

Зайти на <https://colab.research.google.com/>

(можно просто в поисковике ввести “google colab”)

Выбрать вкладку GitHub

Ввести url: <https://github.com/krinistopen/hse-workshop> и нажать Enter

Выбрать появившийся ноутбук

Enterprise-решения

Какие бывают

On-premise приложения

On-premise платформы

Облачные платформы

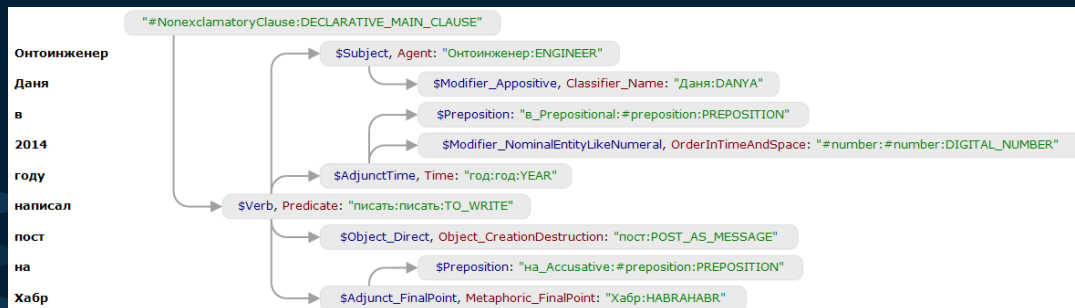
On-premise приложения

ABBYY Compeno

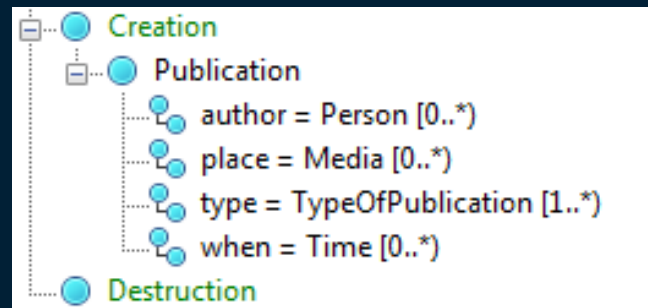
Решает задачу извлечения сущностей и фактов:

- для каждого предложения строит дерево разбора
- по правилам ищет в дереве сущности, например персоны
- по шаблонам из онтологии собирает факты с аргументами, например факт публикации – это использование глаголов «написать», «опубликовать» и т.п. и упоминания, кто, где, когда и что опубликовал

Правила и онтологии создают сотрудники ABBYY, клиенты к ним доступа не имеют.



Семантико-синтаксическое дерево предложения



Часть онтологии

On-premise платформы

[Orange](#), [RapidMiner](#), [Knime](#) и др.

Платформа – единое место для работы с данными, создания решений, дэшбордов. Аналитические задачи решаются заранее разработанными модулями, можно настраивать параметры.

On-premise платформы

Orange

The screenshot displays the Orange Data Mining software interface with a workflow for topic modeling and word cloud generation. The workflow consists of three widgets: Twitter, Preprocess Text, and Topic Modelling, followed by Word Cloud.

Twitter Widget: The Twitter API Key is entered. The Query is set to "Slovenia Germany". The Search by dropdown is set to "Content". The Date range is from "2016-09-20" to "2016-09-30". The Language is set to "Any". The Max tweets is set to "100". The Text includes checkboxes for "Content" and "Author Description".

Topic Modelling Widget: The widget is configured for Latent Semantic Indexing (LSI) with 10 topics. The Topic keywords list shows the following topics and their associated keywords:

- 1 germany, merkel, hillary, usa, last, tesla, week, via, cas, ruined
- 2 bank, deutsche, shares, u, issue, slide, admits, perception, via, https://t.co/8ehvce
- 3 merkel, germany, hillary, germany's, leader, clinton, wants, bringing, muslims, an
- 4 de, whatsapp, datos, https://t.co/0whitqym, usuarios, detener, utilización, order
- 5 cas, whales, dead, parts, sperm, stomachs, full, plastic, found, via
- 6 usa, und, schaff, aschen, der, ihr, ab, mit, kema, ungläublich
- 7 last, win, slovenia, republic, qualifiers, brdo, peppard, panel, month, lee
- 8 germany's, muslims, role, bringing, open, realize, hillary's, like, vetted, borders
- 9 wins, get, uss, nazi, doesn't, effeminate, extremely, certain, trump, less
- 10 economy, post, db, collapse, unlike, insures, etc, petrodollar, renewable, could

Word Cloud Widget: The widget is configured for 0 words in a topic and 100 documents with 611 words. The Cloud preferences are set to "Color words". The Words title is set to "Regenerate word cloud". The Words & weights table shows the following words and their weights:

Weight	Word
66	germany
16	merkel
7	bank
7	u
6	shares
6	deutsche
6	via
5	issue
5	perception
5	last
5	admits
4	hillary
4	usa

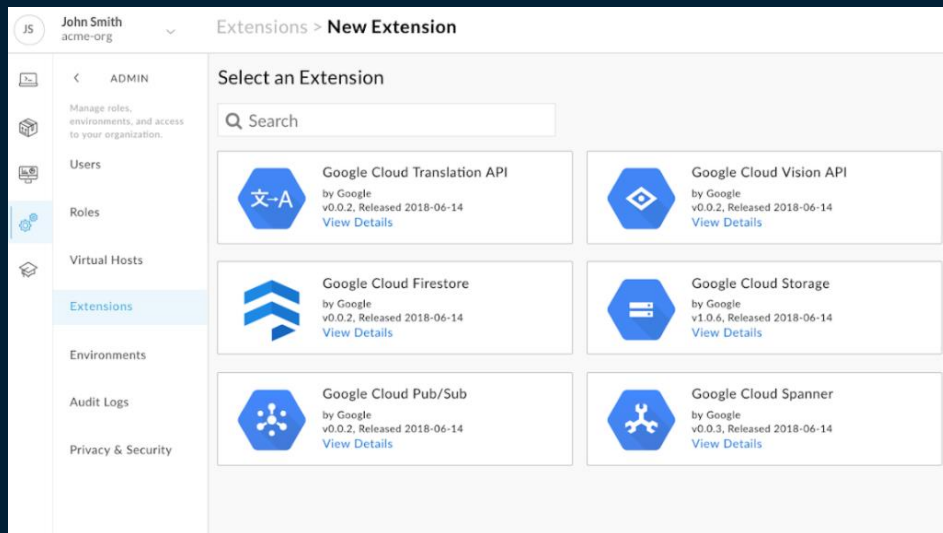
The Word Cloud widget displays a word cloud of the extracted words, with the most prominent words being "germany", "merkel", "hillary", "usa", "last", "tesla", "week", "via", "cas", "ruined", "bank", "deutsche", "shares", "u", "issue", "slide", "admits", "perception", "via", "https://t.co/8ehvce", "merkel", "germany", "hillary", "germany's", "leader", "clinton", "wants", "bringing", "muslims", "an", "de", "whatsapp", "datos", "https://t.co/0whitqym", "usuarios", "detener", "utilización", "order", "cas", "whales", "dead", "parts", "sperm", "stomachs", "full", "plastic", "found", "via", "usa", "und", "schaff", "aschen", "der", "ihr", "ab", "mit", "kema", "ungläublich", "last", "win", "slovenia", "republic", "qualifiers", "brdo", "peppard", "panel", "month", "lee", "germany's", "muslims", "role", "bringing", "open", "realize", "hillary's", "like", "vetted", "borders", "wins", "get", "uss", "nazi", "doesn't", "effeminate", "extremely", "certain", "trump", "less", "economy", "post", "db", "collapse", "unlike", "insures", "etc", "petrodollar", "renewable", "could".

On-premise платформы



Text Mining Series
Twitter Data Analysis

Облачные платформы



Интерфейс личного кабинета Google Cloud.
Выбор подключаемых модулей

Google Cloud, Microsoft Azure, Mail.ru Cloud Solutions, Yandex Cloud и др.

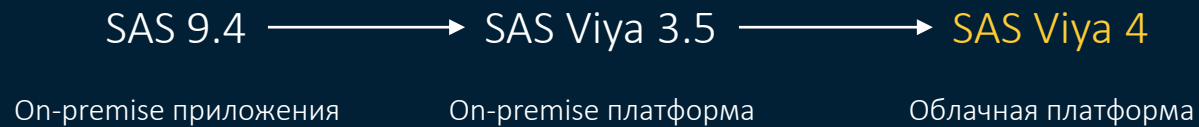
Извлечение сущностей для русского языка есть только у Google.

То же, что и on-premise платформы, только в облаке.

Преимущества:

- не надо покупать и содержать серверное оборудование
- легко подключать необходимые компоненты и отключать ненужные: базы данных, дополнительные ядра, ПО для машинного обучения, веб-серверы и др.
- можно платить только тогда, когда пользуешься платформой

ПО SAS



Демонстрация SAS Viya

Вопросы можете задавать в Telegram:

[@pyatov](#) – Алексей Пятов

[@kondud](#) – Константин Дудников

sas.com

