

Форум
риск-менеджеров

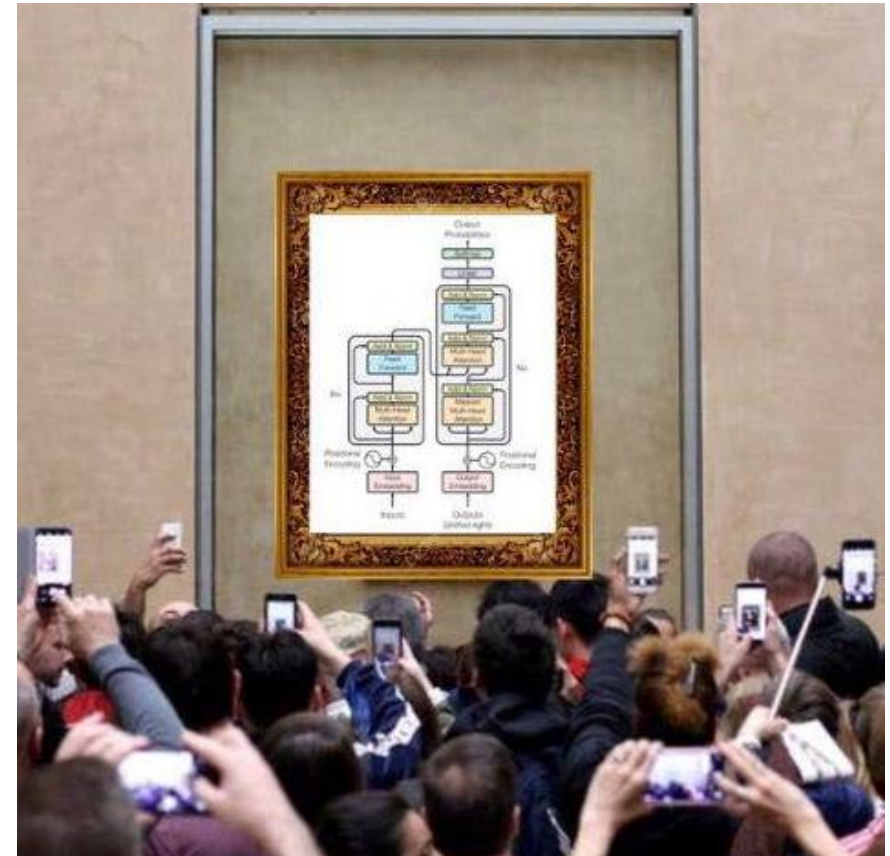
Сергей Афанасьев

КБ «Ренессанс Кредит»

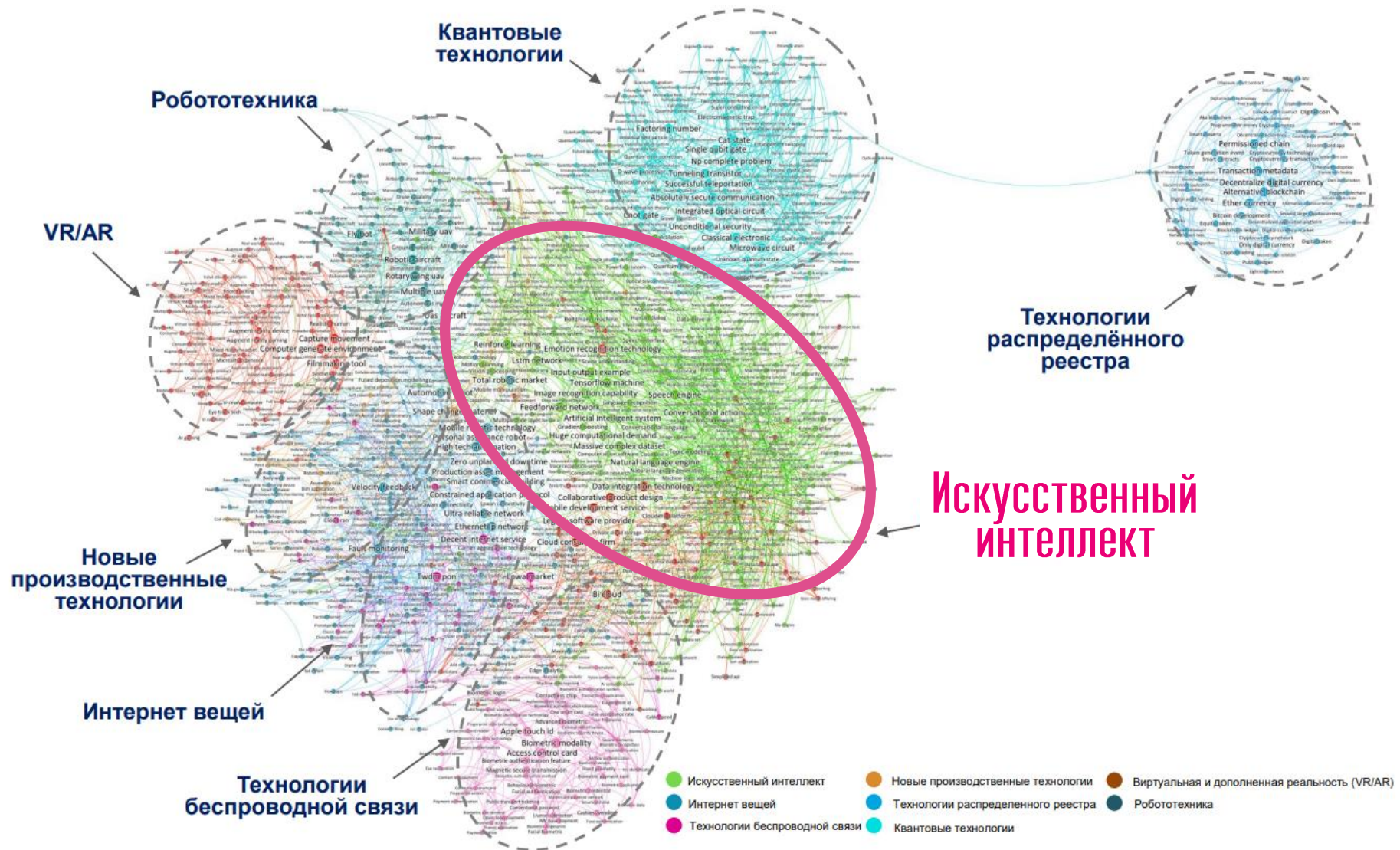


Natural Language Processing в финансовой
организации – опыт банка «Ренессанс Кредит»

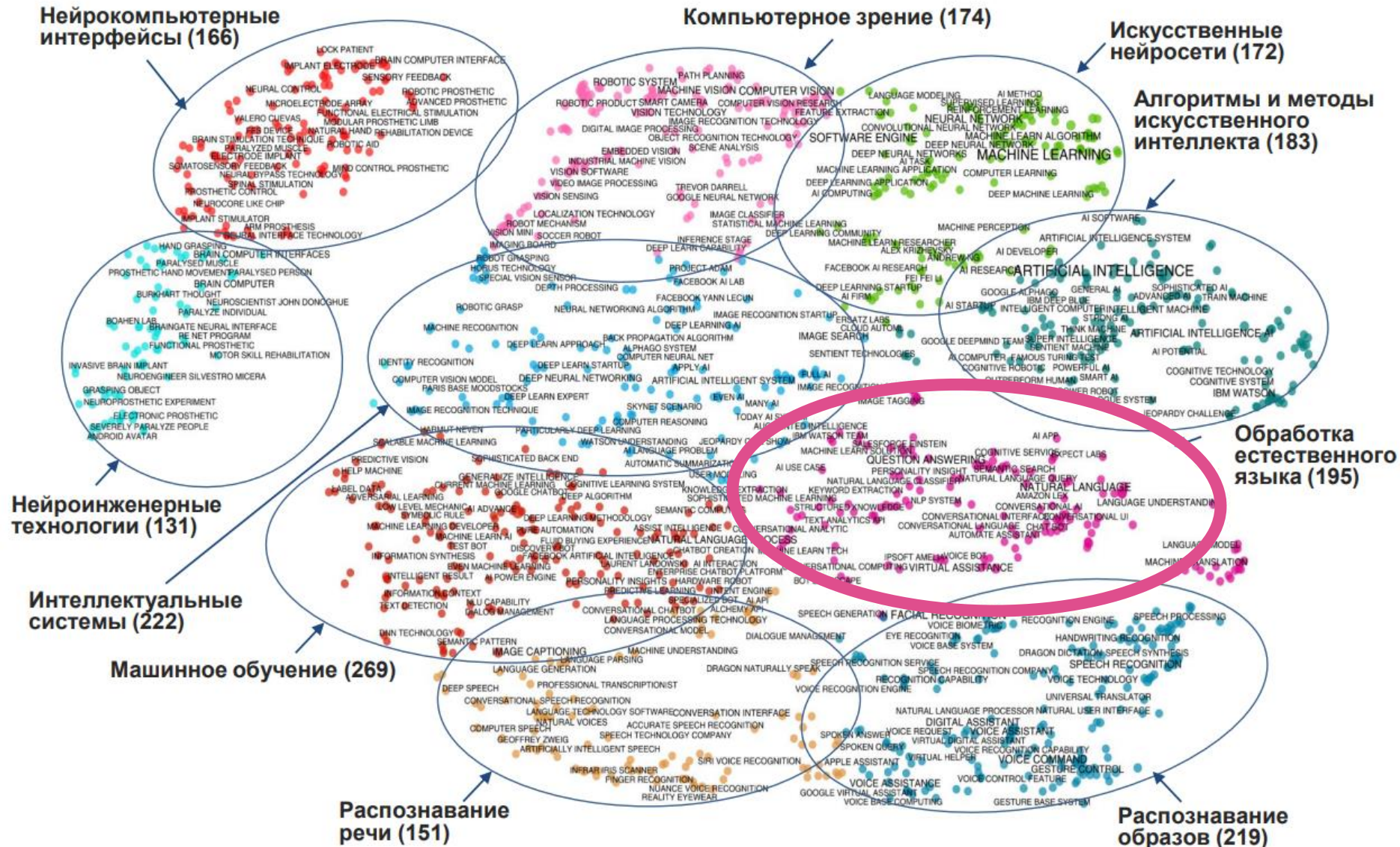
- Как отличить Machine Learning от AI?
- Если написано на Python – это Machine Learning, а если на PowerPoint, то AI



Искусственный интеллект – центральное звено высоких технологий



Тематика ИИ разнообразна – почти 1900 направлений



Вклад стран в глобальную повестку AI исследований



Источник: Система интеллектуального анализа больших данных iFORA (правообладатель - ИСИЭЗ НИУ ВШЭ)

Банк «Ренессанс Кредит»

75 млн
заявок



15 млн
клиентов



9 млн
кредитов



Data Science Team

Risk

CRM

Collection

Antifraud

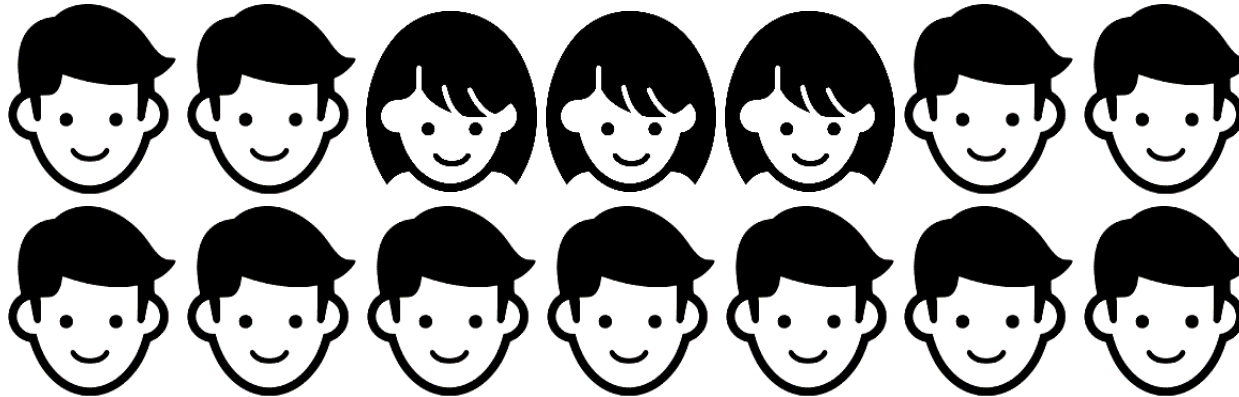
Finance

Marketing

...

Audit

Validation



CRO

Risk
Technologies

Data Base

Architecture

Infrastructure

Направления моделирования

Классика



Risks (scoring)



CRM



Collection



Antifraud



Разработка/поддержка мониторингов моделей

Регуляторка



Разработка моделей IRB/IFRS9



Внедрение моделей IRB в продакшн



Разработка мониторингов



Валидация/аудит моделей IRB/IFRS-9



Подготовка документов для подачи заявки в ЦБ

R&D



NLP (Natural Language Processing)

- Транскрибация звонков
- Текстовая аналитика
- Сервисные задачи для подразделений Банка



Python-Pipeline

- Имплементация новых подходов в Python-Pipeline



Research

- Тестирование внешних источников данных
- Feature Engineering
- ML-Research

NLP задачи для подразделений банка

Обогащение моделей



Collection

+3% GINI

Обогащение моделей взыскания текстовыми данными



CRM

+2% GINI

Обогащение CRM-моделей текстовыми данными



Risks

+? GINI

Обогащение скоринговых карт текстовыми данными

Сервисные NLP-задачи



Customer Service

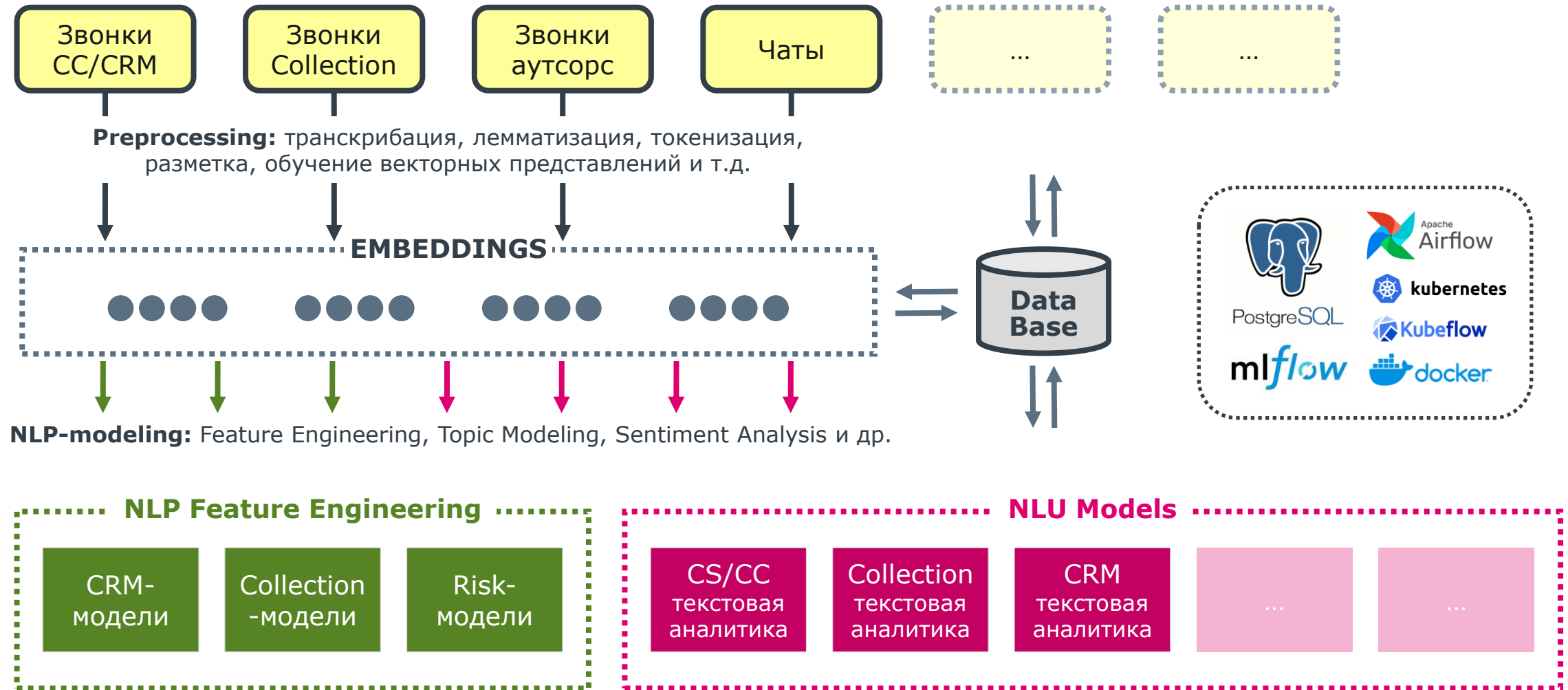
- Автоматизация ответов должникам
- Маршрутизация запросов в Mail Stream
- Выявление дополнительных тематик для чат-бота
- Анализ телефонных разговоров с клиентами



Others

- CRM
- Collection
- Antifraud, AML, IT-Security
- Operational Risk
- ...

Архитектура DL/NLP-платформы



Кейс 1: Бренды покинувшие РФ



Задача

Найти клиентов, работающих в ушедших из РФ компаниях

Проблема

- Список ушедших из РФ компаний опубликован на sravni.ru в английской литературе
- В клиентских анкетах работодатель указан на русском языке в произвольном формате

Реализация

2 недели, фонетическая модель

Кейс 1: Бренды покинувшие РФ

Список компаний

Фонетический модуль



Свои правила на фонетику

Расстояние Левенштейна



400+ брендов

Продукты	54
Одежда	43
Косметика	33
Спорт	26
Автомобили	24
Энергетика/ промышленность	20
Техника	14
ПО	11

[Docs](#) » [transliterate](#) [Edit on GitHub](#)

transliterate

Bi-directional transliterator for Python. Transliterates (unicode) strings according to the rules specified in the language packs (source script <-> target script).

[pypi](#) [v1.10.2](#)
[python](#) [2.7](#) | [3.4](#) | [3.5](#) | [3.6](#)
[build](#) [passing](#)
[license](#) [GPL-2.0-only OR LGPL-2.1-or-later](#)

coverage 92%

Comes with language packs for the following languages (listed in alphabetical order):

- Armenian
- Bulgarian (beta)
- Georgian
- Greek
- Macedonian (alpha)
- Mongolian (alpha)
- Russian
- Serbian (alpha)
- Ukrainian (beta)

There are also a number of useful tools included, such as:

- Simple lorem ipsum generator, which allows lorem ipsum generation in the language chosen.
- Language detection for the text (if appropriate language pack is available).
- Slugify function for non-latin texts.

Prerequisites

- Python >=2.7, >=3.4, PyPy

Installation

Install with latest stable version from PyPI.

```
pip install transliterate
```

or install the latest stable version from BitBucket:

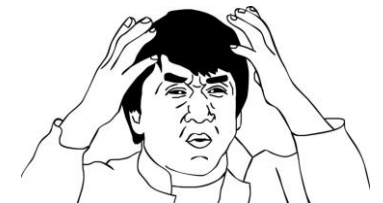
```
pip install https://bitbucket.org/barseghyanartur/transliterate/get/stable.tar.gz
```

or install the latest stable version from GitHub:

[illegible]

$$\text{lev}(a, b) = \begin{cases} |a| & \text{if } |b| = 0, \\ |b| & \text{if } |a| = 0, \\ \text{lev}(\text{tail}(a), \text{tail}(b)) & \text{if } a[0] = b[0], \\ 1 + \min \begin{cases} \text{lev}(\text{tail}(a), b) \\ \text{lev}(a, \text{tail}(b)) \\ \text{lev}(\text{tail}(a), \text{tail}(b)) \end{cases} & \text{otherwise,} \end{cases}$$

	L	e	v	e	n	s	h	t	e	i	n	
L	0	0.5	1	1.5	2	2.5	3	3.5	4	4.5	5	5.5
0.5	0	0.5	1	1.5	2	2.5	3	3.5	4	4.5	5	5.5
1	0.5	0	0.5	1	1.5	2	2.5	3	3.5	4	4.5	5
1.5	1	0.5	0	0.5	1	1.5	2	2.5	3	3.5	4	4.5
2	1.5	1	0.5	0	0.5	1	1.5	2	2.5	3	3.5	4
2.5	2	1.5	1	0.5	0	0.5	1	1.5	2	2.5	3	3.5
3	2.5	2	1.5	1	0.5	0	0.5	1	1.5	2	2.5	3
3.5	3	2.5	2	1.5	1	0.5	0	0.5	1	1.5	2	2.5
4	3.5	3	2.5	2	1.5	1	0.5	0	0.5	1	1.5	2
4.5	4	3.5	3	2.5	2	1.5	1	0.5	0	0.5	1	1.5
5	4.5	4	3.5	3	2.5	2	1.5	1	0.5	0	0.5	1
5.5	5	4.5	4	3.5	3	2.5	2	1.5	1	0.5	0	0.5



Кейс 2: События операционного риска



Задача

Выявлять события операционного риска из бухгалтерских проводок

Проблема

События операционного риска выявляются рисковиком с помощью грубых правил – рутинная работа, которую хочется автоматизировать

Реализация

1,5 месяца, NLP-классификатор

Кейс 2: События операционного риска

Бухгалтерские проводки

NLP-preprocessing

Decision Tree Classifier

Классификация проводок

DESCRIPTION	RISK_FLAG	RISK_CASE
Сторно списанной задолженности по штрафам на просроченный основной долг по кредитному договору [REDACTED] от 28/10/2013, [REDACTED]	0	NaN
Сторно списанной задолженности по штрафам на просроченный основной долг по кредитному договору [REDACTED] от 28/10/2013, [REDACTED]	1	1.0

DESCRIPTION	RISK_FLAG	RISK_CASE
Возврат стоимости Сервис-пакет "Финансовая защита" (26/07/2019) согласно с.3 от 29/08/2019, договор [REDACTED]	0	NaN
Возврат стоимости Сервис-пакет "Финансовая защита" (24/12/2020) согласно с.3 от 15/01/2021, договор [REDACTED]	1	1.0
Возврат стоимости Сервис-пакет "Финансовая защита" (17/09/2020) согласно с.3 от 06/10/2020, договор [REDACTED]	0	NaN
Возврат стоимости Сервис-пакет "Финансовая защита" (11/01/2021) согласно с.3 от 08/04/2021, договор [REDACTED]	1	1.0
Возврат стоимости Сервис-пакет "Финансовая защита" (11/07/2019) согласно с.3 от 16/01/2020, договор [REDACTED]	1	1.0

TF-IDF

Word	TF		IDF	TF*IDF	
	A	B		A	B
The	1/7	1/7	$\log(2/2) = 0$	0	0
Car	1/7	0	$\log(2/1) = 0.3$	0.043	0
Truck	0	1/7	$\log(2/1) = 0.3$	0	0.043
Is	1/7	1/7	$\log(2/2) = 0$	0	0
Driven	1/7	1/7	$\log(2/2) = 0$	0	0
On	1/7	1/7	$\log(2/2) = 0$	0	0
The	1/7	1/7	$\log(2/2) = 0$	0	0
Road	1/7	0	$\log(2/1) = 0.3$	0.043	0
Highway	0	1/7	$\log(2/1) = 0.3$	0	0.043

1. Удаление нетекстовых символов, перевод в нижний регистр;
2. Лемматизация;
3. Удаление имен, названий географических субъектов, месяцев, стоп-слов;
4. Добавление токенов начала и конца предложения;
5. TF-IDF

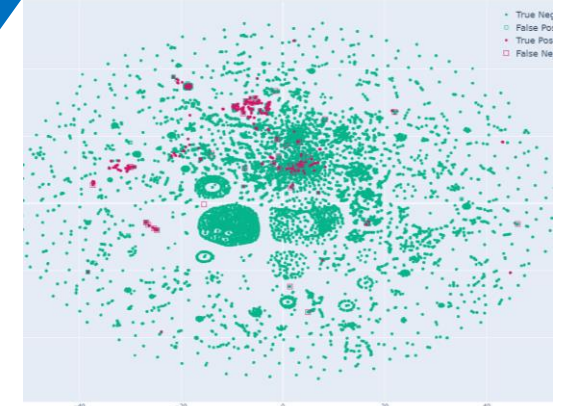
```

|---- возврат <= 2.31
|---- стorno <= 4.02
|---- осн <= 3.46
|---- слуховой <= 3.50
|---- затрата <= 4.19
|---- суа <= 3.66
|---- возмещение <= 3.64
|---- основание <= 3.44
|---- штраф <= 3.97
|---- дело <= 4.05
|---- bos исправительный <= 3.60
|---- truncated branch of depth 31
|---- bos исправительный > 3.60
|---- truncated branch of depth 28
|---- дело <= 4.05
|---- взыскатель <= 3.94
|---- truncated branch of depth 11
|---- взыскатель > 3.94
|---- truncated branch of depth 3
|---- штраф <= 3.97
|---- задолженность <= 1.23
|---- bos noрaжение <= 1.99
|---- truncated branch of depth 31
|---- bos noрaжение > 1.99
|---- class: 0.0
|---- задолженность > 1.23
|---- class: 0.0
|---- основание <= 3.44
|---- клиент <= 2.33
|---- obligation <= 1.57
|---- резерв <= 1.91
|---- truncated branch of depth 81
|---- резерв > 1.91
|---- class: 0.0
|---- obligation <= 1.57
|---- class: 0.0
|---- клиент > 2.33
|---- class: 0.0
|---- возмещение > 3.64
|---- договор <= 0.81
|---- основание <= 3.44
|---- стоимость <= 3.25
|---- согласно <= 2.51
|---- truncated branch of depth 4
|---- согласно > 2.51
|---- truncated branch of depth 3
|---- стоимость > 3.25
|---- дог <= 2.16
|---- truncated branch of depth 8
|---- дог > 2.16
|---- truncated branch of depth 2
|---- основание <= 3.44
|---- bos возмещение <= 4.05
|---- class: 0.0
|---- bos возмещение > 4.05
|---- записка <= 3.49
|---- class: 1.0
|---- записка > 3.49
|---- class: 0.0

```

t-SNE

T-SNE plot, test data filtered, normalized description



	ROC AUC	Recall	Precision
Train	99.98%	99.88%	72.98%
Test	99.71%	97.97%	70.82%
OOT	99.13%	94.44%	44.74%

Кейс 3: Раздолжники



Задача

Выявлять клиентские обращения от фирм-раздолжников

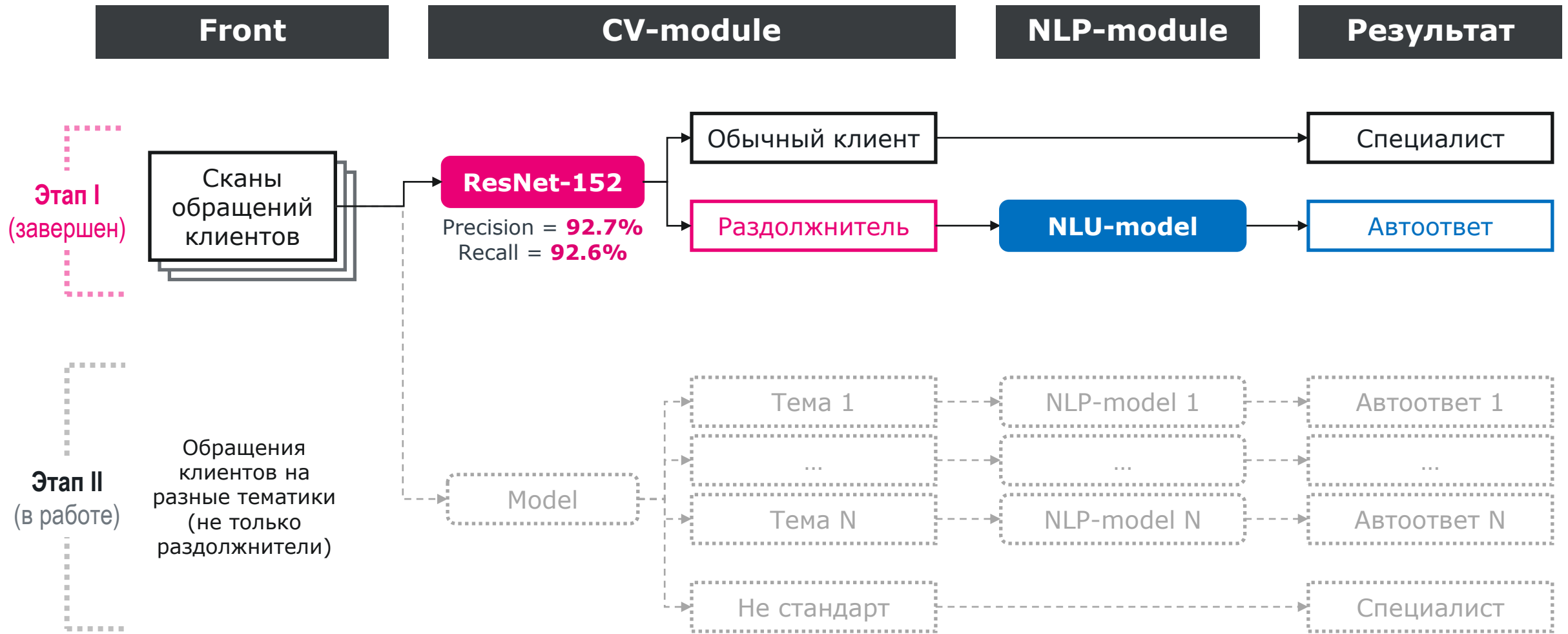
Проблема

Специалистам банка необходимо выявлять письма, подготовленные мошенническими юридическими фирмами – рутинная работа, которую хочется автоматизировать

Реализация

5 месяцев, CV + OCR + NLP

Кейс 3: Раздолжники



Кейс 3: Раздолжнители

NLP-preprocessing

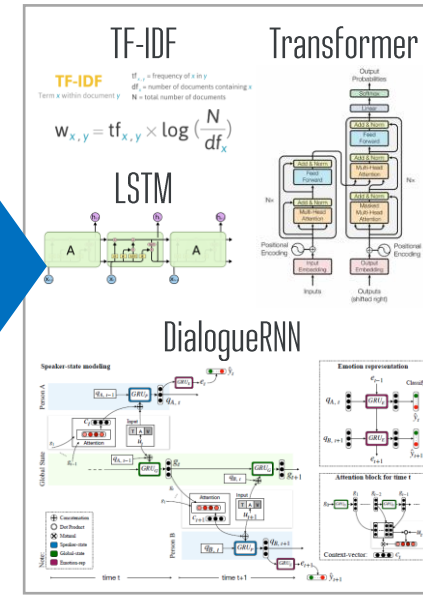
Обращение
клиента (pdf)

Удаление
вотермарок (pdf)

Распознавание и
корректировка
текста (txt)

NLU-модуль
для подготовки
автоответа

Шаблоны
автоответов



Шаблон-1

ИВАН ИВАНОВИЧ, выписка по лицевому счету – документ, в котором отражены поступления и списания денежных средств с Вашего счета за определенный период. Согласно Положению ЦБ РФ 27 февраля 2017 года N 579-П (п.2.1., раздела 2, части III) Выписки из лицевых счетов, рассчитанные с использованием средств вычислительной техники, выдываются клиентам без штампов и подписей работников кредитной организации.

...

Шаблон-2

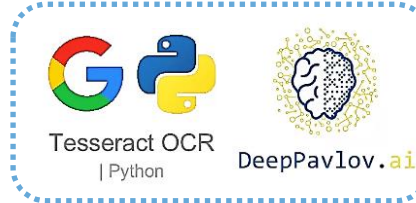
ИВАН ИВАНОВИЧ, Вами был заключен договор с КБ «Ренессанс Кредит» (ООО), подписав который, Вы приняли на себя обязательства по исполнению его надлежащим образом. Пункт 1 ст. 450 ГК РФ устанавливает, что изменение и расторжение договора возможны по соглашению сторон, если иное не предусмотрено настоящим Кодексом, другими законами или договором.

...

Шаблон-N

ИВАН ИВАНОВИЧ, сообщаем, что по состоянию на г. по Вашему кредиту имеется просроченная задолженность. Мы уверены, что Вы ответственно относитесь к своим обязательствам перед Банком, и надеемся, что Вы предпримите все возможные действия с Вашей стороны для погашения просроченной задолженности в кратчайшие сроки.

...



Спасибо за внимание!

Афанасьев Сергей

Вице-президент
Начальник управления
статистического анализа

КБ «Ренессанс Кредит»

safanasev@rencredit.ru



"Théâtre D'opéra Spatial", Jason Allen & Midjourney, 2022