

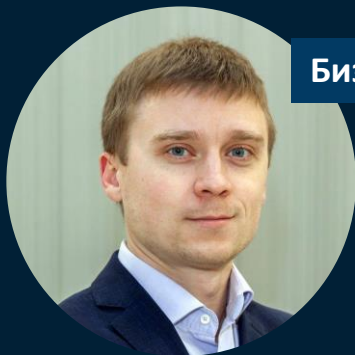
Текстовая аналитика в бизнесе

На платформе SAS и не только

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

211118

Будем знакомы



Бизнес

Алексей Пятов

Data Management &
Text Analytics Team Leader

К.Э.Н.



Техника

Константин Дудников

Text Analytics Consultant

Магистратура ОТиПЛ МГУ

О чем сегодня пойдет речь

- 1. План занятий.
- 2. Тема №1: Какие бизнес-задачи решаются текстовой аналитикой и где она востребована?
- 3. Примеры кейсов.

План занятий: обзорная часть

- **Тема 1. «Бизнес-задачи текстовой аналитики»**

Обзор задач бизнеса, которые решаются с помощью классификации текстов или извлечения фактов из неструктурированных данных (отчетов, пользовательских отзывов, новостей и т.д.). Рассмотрение индустрий, где такие задачи встречаются.

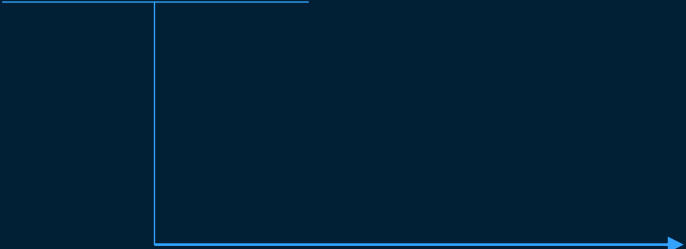
- **Тема 2. «Инструменты и методы текстовой аналитики»**

Обзор методов обработки текста и извлечения знаний из неструктурированных данных. Обзор программных продуктов: свободного ПО, библиотек Python и SAS Viya, которые эти методы реализуют.

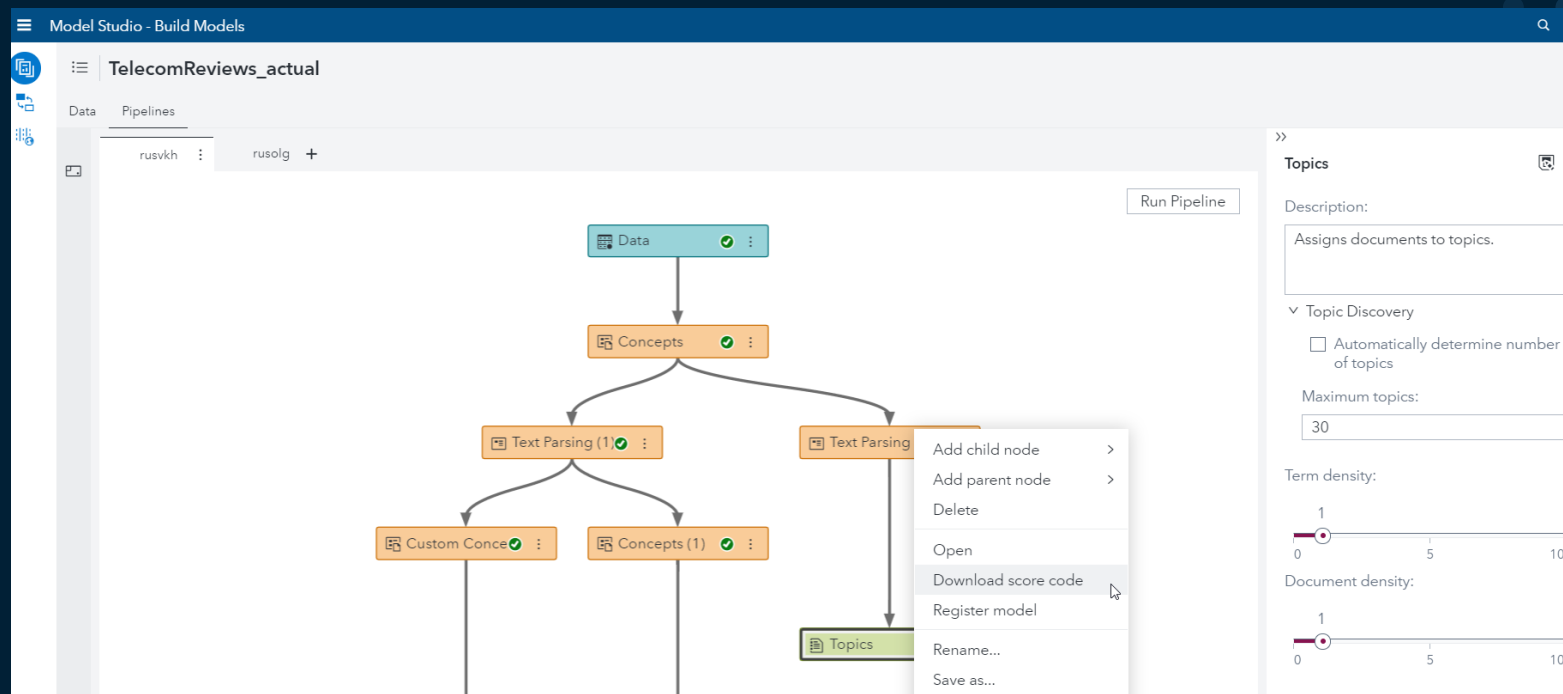
План занятий: прикладная часть

Темы 3-4. «Текстовая аналитика в деле»:

- банки и телеком
- образование и закупки

- 
- исследование пользовательского опыта, анализ событий
 - классификация ответов учеников и объектов медицинских закупок
 - анализ отзывов потребителей, юридических документов

Практика 1: SAS Viya



SAS Visual Text Analytics User Manual

Практика 2: немножко Python

```
63 # получаем эмбединги для обучающей и тестовой выборки
64 train_embeddings = []
65 for text in train_data:
66     embedding = embed(str(text))[0]
67     train_embeddings.append(embedding)
68
69 test_embeddings = []
70 for text in test[text_var]:
71     embedding = embed(str(text))[0]
72     test_embeddings.append(embedding)
73
74 ###
75 # обучаем классификаторы
76 nbg = GaussianNB()
77 nbb = BernoulliNB()
78 svm = SVC(kernel='poly', gamma='scale', probability=True)
79 lg = LogisticRegressionCV(cv=5, multi_class='ovr', max_iter=1200)
80
81 classes = train[class_var].tolist()
82
83 nbg.fit(train_embeddings, classes)
84 nbb.fit(train_embeddings, classes)
85 svm.fit(train_embeddings, classes)
```

Будем работать онлайн в Google Colab

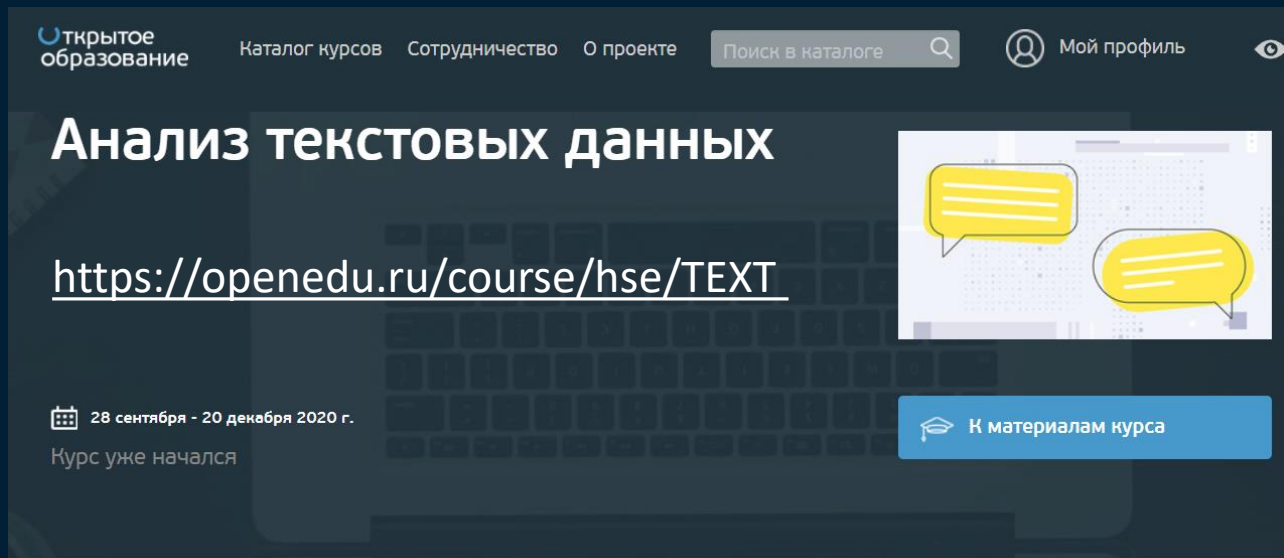


Тема №1: Текстовая аналитика в бизнесе

Кому, зачем, как?



Курс «Анализ текстовых данных» на OpenEdu



The screenshot shows the OpenEdu website interface. At the top, there is a navigation bar with links: 'Открытое образование', 'Каталог курсов', 'Сотрудничество', 'О проекте', a search bar 'Поиск в каталоге', a user profile icon with 'Мой профиль', and an eye icon. The main header features the course title 'Анализ текстовых данных' in large white letters. Below it is the URL <https://openedu.ru/course/hse/TEXT>. To the right of the URL is a graphic with two yellow speech bubbles. Below the URL, on the left, is a calendar icon and the text '28 сентября - 20 декабря 2020 г.' and 'Курс уже начался'. On the right is a blue button with a graduation cap icon and the text 'К материалам курса'.



Бесплатный
онлайн-курс
(короткие
видео)

Онлайн-курс **«Анализ текстовых данных»** посвящён обработке текстов методами машинного обучения. В ходе обучения слушатели курса узнают о различных задачах, связанных с анализом текстов, освоят методы предобработки текстовых данных, изучат основные подходы к решению задач на основе классического машинного обучения и глубоких нейронных сетей.

[О курсе](#)

[Формат](#)

[Программа курса](#)

Поделиться



12 недель

длительность курса



около 5 часов в



Книга Speech and Language Processing (Jurafsky, Martin)

Speech and Language Processing (3rd ed. draft)

[Dan Jurafsky](#) and [James H. Martin](#)

Draft chapters in progress, October 16, 2019

web.stanford.edu/~jurafsky/slp3



Стэнфордский курс по Nature Language Processing



CS224n: Natural Language Processing with Deep Learning

Stanford / Winter 2020



web.stanford.edu/class/cs224n



A series of horizontal bars of varying lengths and colors (teal, blue, and dark blue) are positioned on the left side of the slide, creating a modern, abstract background element.

Тема №1: Текстовая аналитика в бизнесе

Кому, зачем, как?



Прирост данных за 1 минуту



470,000
ТВИТОВ



16,000,000
сообщений



2,400,000
поисковых запросов



156,000,000
электронных писем

Текст – крупнейший источник данных,
созданный человеком

Текстовая аналитика в бизнесе прежде всего решает задачу извлечения релевантной информации

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 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

Текстовая аналитика в бизнесе прежде всего решает задачу извлечения релевантной информации

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 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!

Интернет: негатив

Связь: негатив

Общая негативная оценка

Контакт-центр: проблема

Брак продукта

Сотрудники: негатив

Обращение к детализации

Коммуникация: КЦ

Смена оператора

Давний клиент

Задачи текстовой аналитики



Задачи анализа текстовых данных для клиентской аналитики

На индивидуальном уровне



Обогащение знаний о клиенте

Составление более комплексного
профиля абонентов, выявление
ключевых событий

На агрегированном уровне



Оценка восприятия компании (тепловая карта)

Агрегированный сбор мнений о
продуктах, рекламе, конкурентах,
службе поддержки



Система быстрого реагирования

Обнаружение отзывов,
требующих срочной реакции
сотрудников службы поддержки

A series of horizontal bars of varying lengths and colors (teal, blue, green) are positioned on the left side of the slide, creating a modern, abstract background element.

Voice of Customer

Анализ мнений потребителей

B2C | B2B



Voice of Customer

1

Анализируем обратную связь (отзывы и обращения), автоматически выявляем ключевые проблемы

2

Автоматизируем работу колл-центра и чатов

3

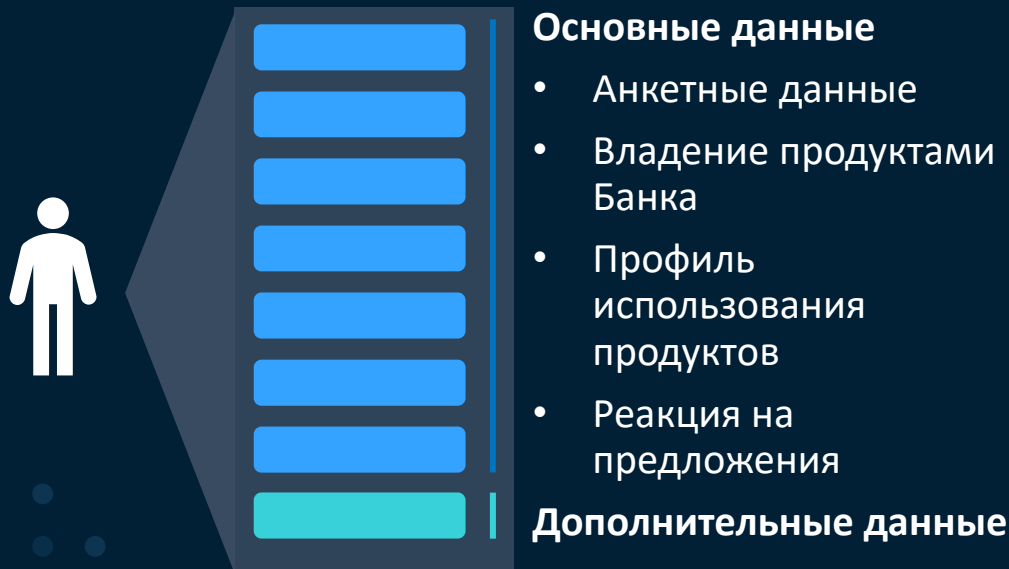
Извлекаем ценную информацию из любых текстовых документов большого объема

ЦЕЛИ:

- Лучший клиентский сервис (быстрые и полные ответы, нет ожидания даже при пиковой нагрузке)
- Анализ качества услуг и получение инсайтов
- Аудит работы сотрудников

Дополнительные атрибуты для анализа

Виды информации



- Удовлетворенность текущими аспектами продуктов
- Заинтересованность в дополнительных услугах
- Планы покупки недвижимости, авто

Кейс: Royal Bank of Scotland: Voice of Customer

Крупнейший коммерческий
банк в Шотландии

Цель:

- Повышение эффективности удержания клиентов

Результат:

- Повышение производительности анализа диалогов более чем в 1000 раз
- **Повышение эффективности выявления слабых сторон оператора**
- Выявление причин обращения в поддержку на всём массиве данных
- **Учет расхождений между эмоциональной окраской диалога и данными в CRM**



Амбиция: стать банком
с лучшим клиентским сервисом

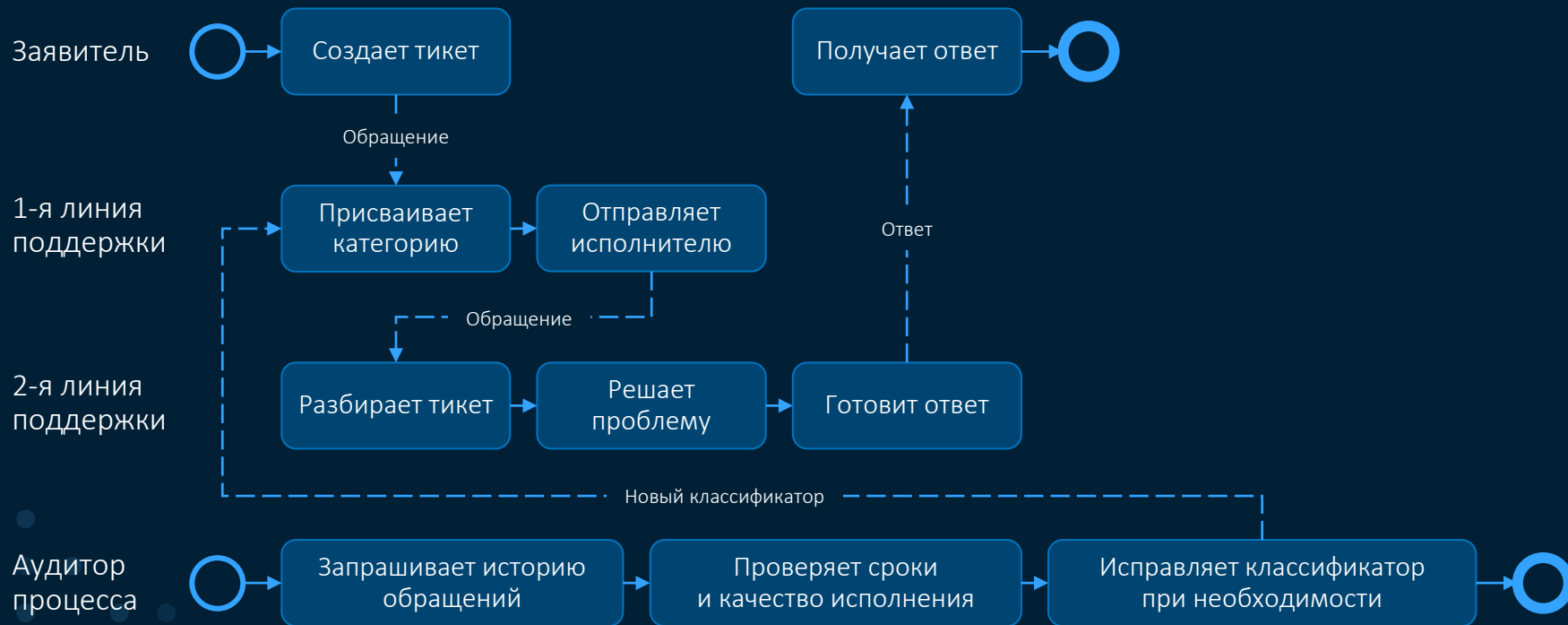
Проблема:

- Высокие трудозатраты на анализ 250 тыс. диалогов в месяц

Решение:

- Внедрение автоматической классификации диалогов
- Индустриализация выявления эмоциональной окраски диалога

Типовой процесс обработки обращений и проблематика



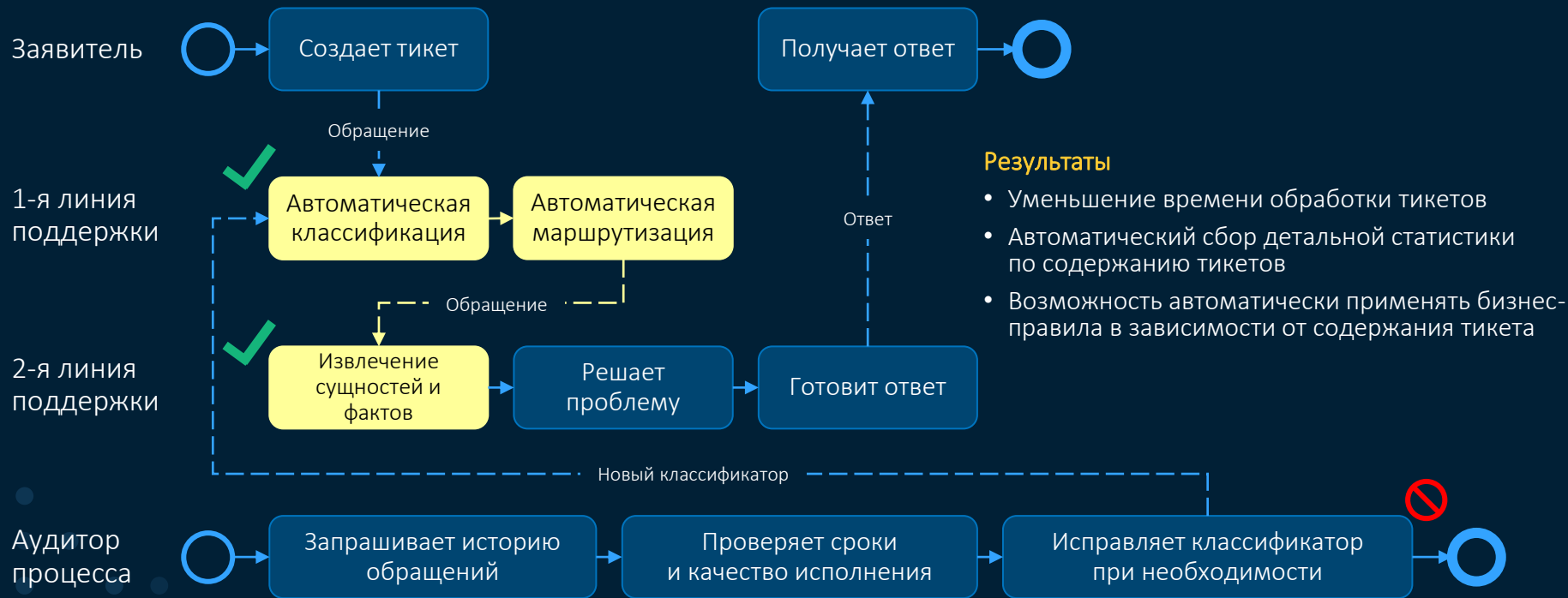
Типовой процесс обработки обращений и проблематика



Первый этап оптимизации – автоматическая классификация и маршрутизация обращений



Второй этап оптимизации – автоматический разбор обращений, извлечение сущностей и фактов



Опционально – внедрение инструмента для поиска новых категорий и редактирования справочника



Идеальный процесс автоматической обработки обращений



Решение текущих проблем

Предотвращение проблем в будущем

Пример тепловых карт на основе мнений пользователей

Тепловая карта удовлетворенности качеством обслуживания на основе обращений в колл-центр, отзывов в соц.сетях



Тепловая карта удовлетворенности обслуживанием в колл-центре

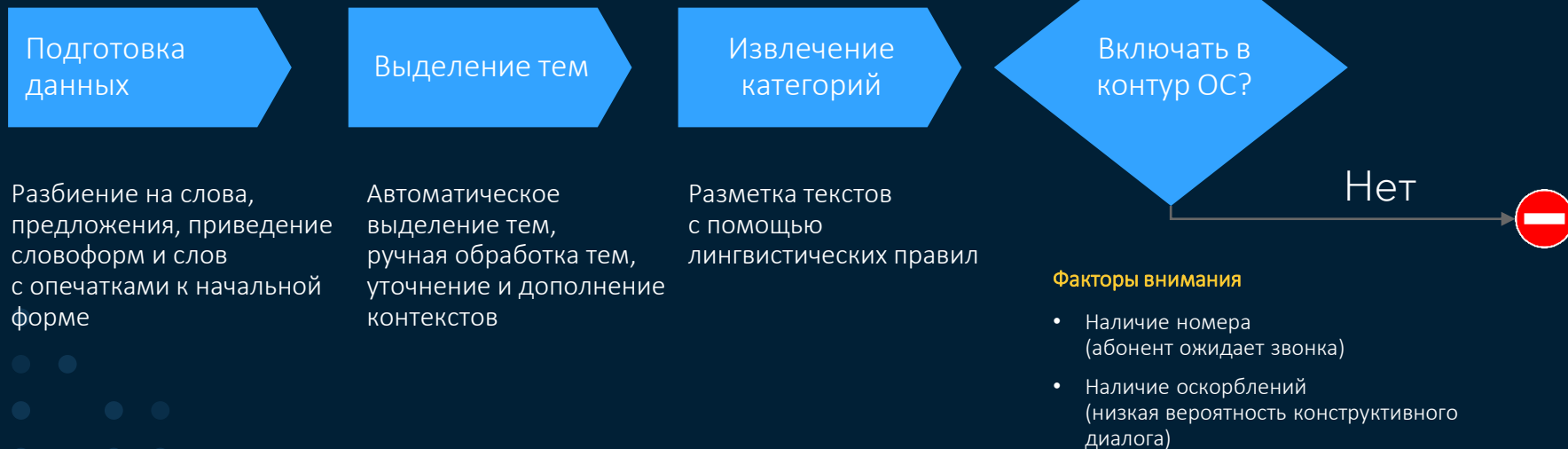
	Центральный	Северо-Западный	Южный	Северо-Кавказский	Приволжский	Уральский	Сибирский	Дальневосточный
0-1	1,4	0,7	2,7	0,5	3,7	1,3	3,4	3,4
1-2	4,2	1,8	1,6	3,8	0,7	2,9	4,8	2,7
2-3	4,5	1,2	2,7	3,1	0,8	3,8	0,7	2,3
3-4	2,8	3,6	1,0	1,7	2,8	1,5	4,8	3,0
4-5	4,5	3,0	4,0	0,3	3,4	3,8	1,1	2,8
5-6	1,5	0,5	1,2	3,2	4,8	0,5	3,7	2,9
6-7	3,7	1,3	1,5	0,0	0,6	4,3	3,3	3,1
7-8	3,7	0,7	0,2	0,2	4,6	2,3	3,7	2,8
8-9	2,6	0,8	4,9	3,3	1,5	3,4	1,0	2,7
9-10	2,5	1,1	2,9	4,5	1,7	1,3	2,4	0,7
10-11	2,8	0,2	0,4	2,6	2,4	2,0	1,6	0,6
11-12	1,5	4,2	4,6	3,7	2,9	2,6	2,8	1,4
12-13	3,2	4,8	2,2	2,6	1,2	1,9	3,9	4,6
13-14	3,9	3,5	0,8	2,0	4,8	4,7	2,0	1,4
14-15	4,8	4,0	1,8	3,7	0,9	0,9	3,6	0,5
15-16	0,4	3,2	2,1	1,4	2,8	1,7	0,3	3,4
16-17	2,1	1,1	1,7	3,9	2,5	4,9	1,7	3,1
17-18	3,4	2,3	1,7	5,0	4,4	3,0	2,2	1,4
18-19	3,1	1,1	0,9	3,9	4,1	2,5	4,7	0,4
19-20	0,9	0,7	3,2	0,7	2,0	4,3	1,5	2,3
20-21	0,9	5,0	1,7	1,0	4,6	2,3	0,1	4,4
21-22	5,0	4,4	4,1	4,3	2,2	4,4	4,1	0,1
22-23	4,2	3,7	1,4	0,5	3,1	2,7	1,8	0,8
23-24	0,8	2,4	4,8	3,4	0,2	2,3	2,7	2,5

Оценки 0...5

Нужно ли перезвонить абоненту?

Новый процесс в телекоме

Оператор связи раз в квартал проводит опрос среди абонентов с целью определить их удовлетворенность услугами. Оператор собирал результаты опроса, но никак не обрабатывал. SAS построил модель, которая определяет необходимость дополнительного контакта с клиентом по результатам опроса.



Телеком: Взвешенная модель: минус

« СНИМАЕТЕ ДЕНЬГИ НИ ЗА *** С*****Й!!!!!! ПОМЕНЯЛИ
ТАРИФ! ВООБЩЕ ОТКАЖУСЬ ОТ ***** Оператора.
П***** ВЫ НА***!!!! СПЛОШНОЕ НАЕ*****!!!! »

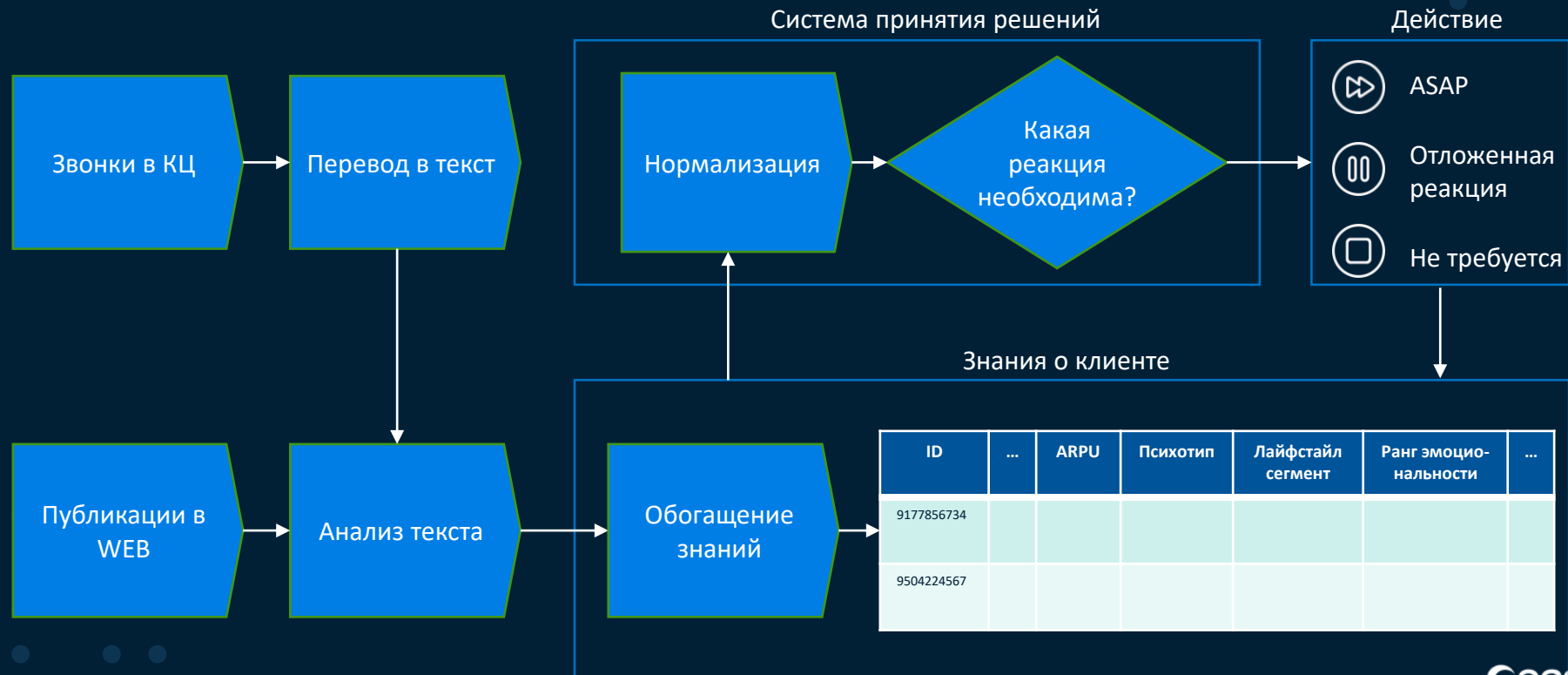
-0.5 балла

Телеком: Взвешенная модель: плюс

« ВЫ ОСТАВИЛИ МЕНЯ БЕЗ СВЯЗИ ЗА ГРАНИЦЕЙ!!!!
+7 9** *** ** 33.. СРОЧНО ПОДКЛЮЧИТЬ!!!! »

+1 балл

Пример системы быстрого реагирования на жалобы клиентов



Кейс: анализ мнений покупателей в ритейле

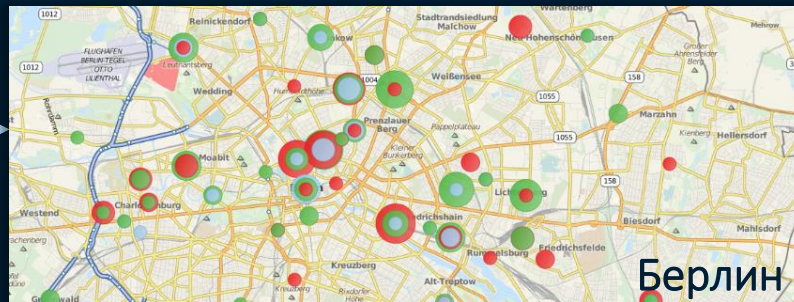
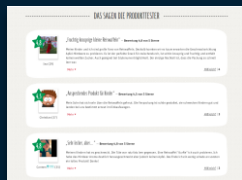
Крупный европейский ритейлер захотел узнать:

- мнения покупателей о продуктах под его собственной торговой маркой
- мнения покупателей о магазинах компании в Берлине

Отзывы на YELP



Отзывы на сайте
изготовителя

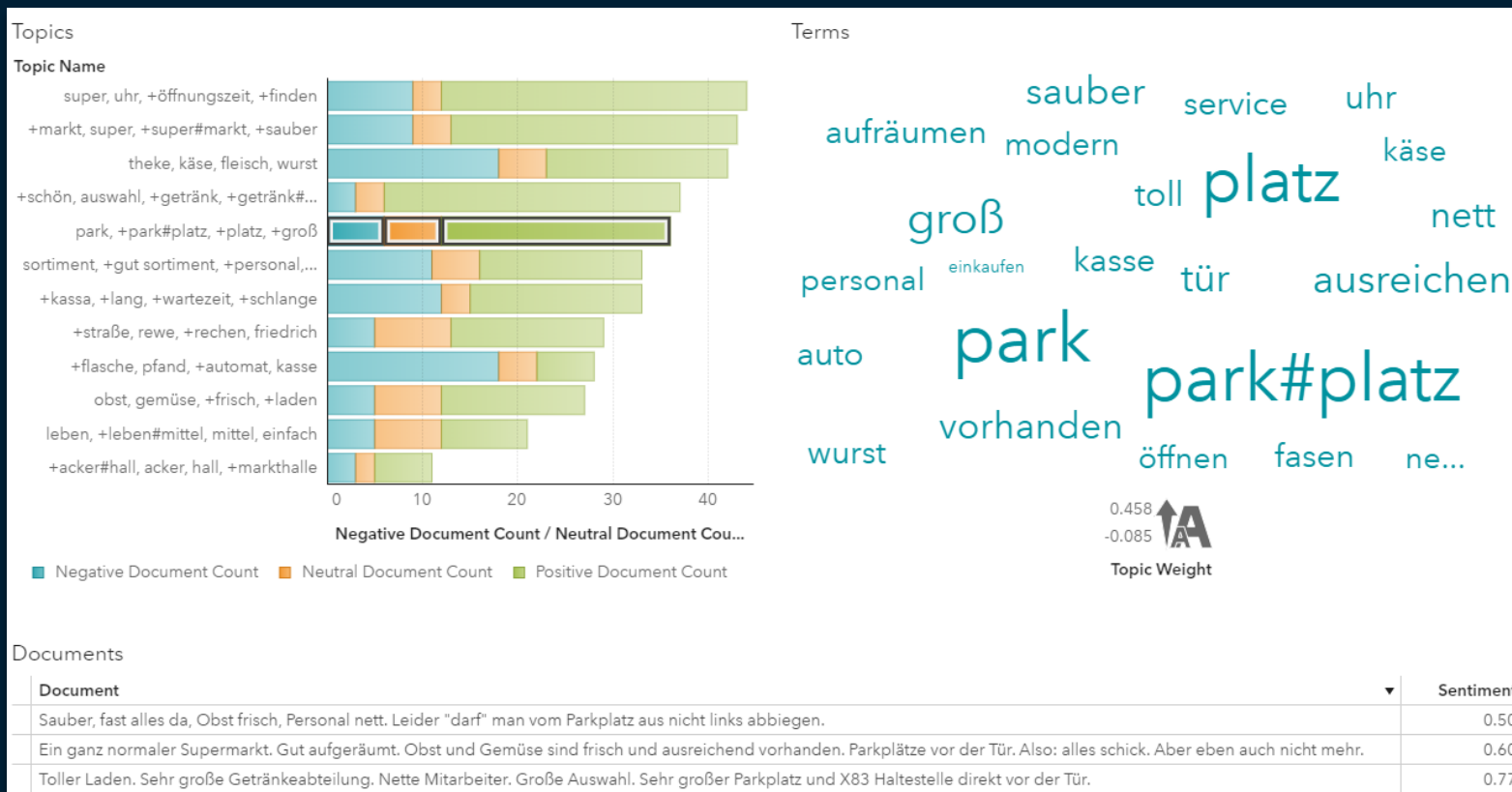


“

В вашем продукте слишком много сахара для того чтобы называть его органическим!

”

YELP: Автоматическое выделение тем



YELP: Автоматический анализ тональности



Обращения: жалобы, благодарности и т.д.

Можно использовать для:

- Урегулирования инцидентов и анализа рисков
- Сбора разрозненной обратной связи по продуктам для ВП
- Повышения First Call Resolution
- Обогащения данных по клиентам
- Любой задачи, для которой необходим анализ обратной связи от клиентов или сотрудников

Next Best Offer для клиента КЦ

Голос -> текст

Текст ->
причина
обращения

Подбор
предложения

1

УВЕЛИЧЕНИЕ КОНВЕРСИИ

- Приоритезация / пересчет персональных предложений для клиентов в режиме реального времени в зависимости от контекста обращения.
- Приоритезация информационных / сервисных сообщений в зависимости от контекста обращения.
- Управление скриптами продаж.
- Оптимизация расходов на коммуникации.

3,5% - 11%

2

СОКРАЩЕНИЕ ОТТОКА

- Приоритезация закрытых спец. предложений на удержания с учетом ценности клиента (CLTV), анализа операции, контекста обращения / жалобы, вероятности оттока и клиентского опыта
- Управление маркетинговым бюджетом – выбор «подарков» / стимулов в зависимости от контекста обращения/претензии

7% - 19%

**Текстовая информация может
анализировать в режиме
реального времени**

Задачи: анализ разговоров с коллекторами

Транскрипт разговора:
«Да-да, **заплатчу** обязательно.
Только не сейчас, **потом**.
Летом, **в июне** где-то. **Числа**
двадцатого»

Накопление
данных

Предобработка
текста

счета -> счёт
интернэт - интернет
учет синонимов
частеречная разметка

Переписка

Разработка
лингвистических правил

```
(UNLESS, bad_context,  
(SENT, (DIST_8, "_payment{заплатчу}",  
_later{потом})) )
```

Экспертные знания

Извлечение
фактов

Обещание заплатить: да
Дата ожидаемого платежа:
20.06.20

Векторное
представление

Текст -> N-размерный вектор



Тематический анализ

+вернуть, +месяц, +течение
+нет, +возможность, +проблема



Аналитические
модели

Машинное обучение



Насколько точно
операторы
следуют скрипту?



Что именно
операторы
обсуждают
с клиентами?



Какие новые
темы
разговоров
появляются?

Анализ платежей

B2B



Анализ назначений платежей

Актуальный профиль деятельности
45% входящего оборота за прошлый квартал – за молочные продукты. С учетом сумм, вероятно, фирма занимается оптовой торговлей данными товарами

ID клиента	Направление	Сумма	Назначение платежа
222777	BX	60 000	Оплата по договору №777-415 за молочную продукцию
222777	ИСХ	56 896	Оплата по договору №212-02567
222777	BX	11 560	Перевод ср-в по дог №12-14 от 24/05/2014 за поставку сыров
4159084	BX	4 567	Оплата за косметич услуги на выставке
4159084	BX	34 677	Для зачисления на счет Павлова А.П. - вознаграждение – вкл. НДС
4159084	BX	13 700	За мастер класс по стрижке
4159084	ИСХ	45 766	Оплата в банк по договору эквай-га

Особые виды связей ЮЛ/ФЛ

Последние полгода проводит обучение
Переводит зарплату сотрудникам

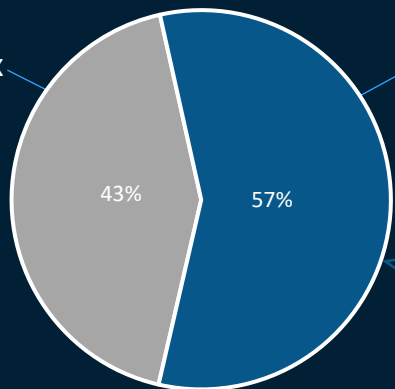
Владение продуктами конкурентов

Исходящие транзакции
за эквайринг в другой банк

Пример: отрасль можно определить почти по половине активных компаний

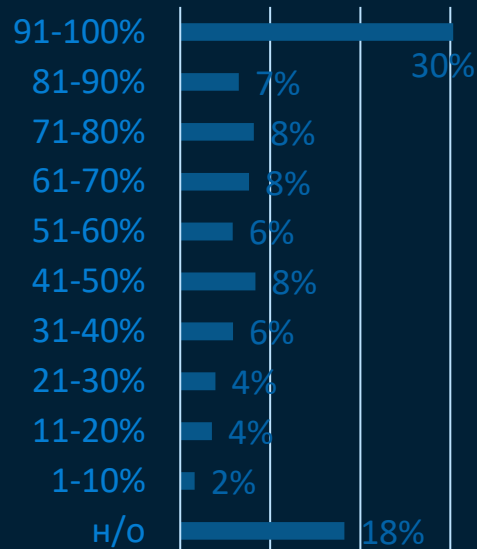
Распределение клиентов по входящим платежам (1 месяц)

Не было входящих платежей



Были входящие платежи

Распределение компаний по доле лидирующей категории во входящих платежах

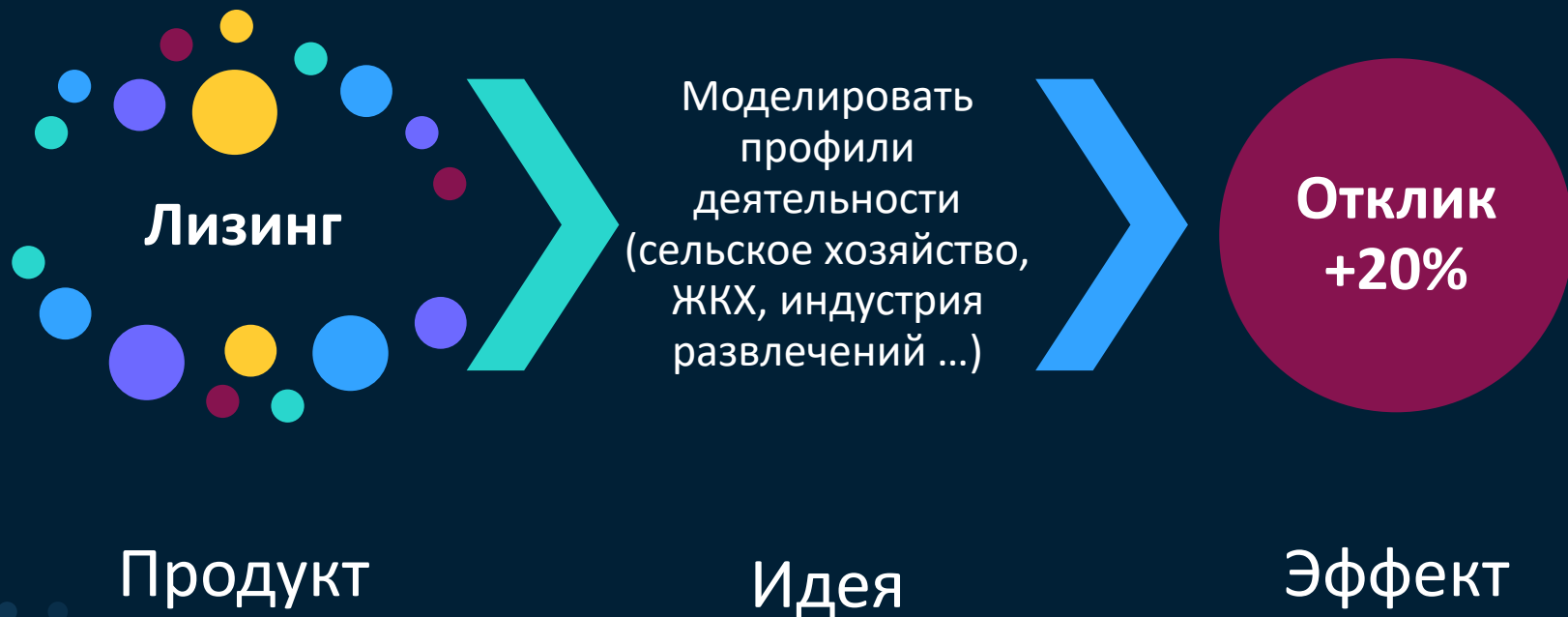


Предположение: основная отрасль компании = лидирующей категории, если лидирующая категория > 70%.

Таким образом, отрасль будет определена у 45% компаний, у которых были входящие платежи

Только у 18% платежей не смогли определить категорию

Назначение платежей: кампания по бизнес-правилам



Текстовая аналитика: зачем еще?

Можно использовать соцсети, сайты отзывов, СМИ для:

Оценки восприятия:

- рекламных кампаний (где и как реагируют на конкретные макеты / персонажей)
- мест размещения рекламы (где и как лучше реагируют на рекламу, в каких местах лучше размещаться)
- продуктов (как клиенты относятся к новым тарифам и опциям)
- розничных магазинов компании (какое мнение складывается у посетителей, вплоть до отдельных сотрудников в случае конфликтов)
- качества связи / обслуживания (в каких регионах/городах/районах наблюдаются проблемы со связью)
- конкурентов



Мнения потребителей и факты часто скрыты в глубинах информационного поля

Потребитель столкнулся проблемой и у него **нет выбора**



Написал адресную жалобу или отзыв

Поставщик продукта



Айсберг мнений

Потребитель столкнулся с проблемой и у него **есть выбор**



Взял другой продукт

Написал гневный пост

Рассказал друзьям

репутация



Форумы

Блоги

СМИ



фактов

Потребителю понравился продукт



Написал хвалебный пост

Что-то произошло на рынке продукта



Вышла статья в СМИ



Текстовая аналитика: зачем еще?

Можно использовать для:

- Повышения качества рекомендационных моделей для стриминговых сервисов (видео по запросу) за счет анализа рецензий на фильмы.
- Анализа медицинских записей (аудит процедур)

Текстовая аналитика: кому?

- Банки
 - Телеком-операторы
 - Страховые компании
 - Промышленность
 - Медицинские организации
 - Застройщики
 - Ритейл
-
- Департаменты качества, маркетинга, рисков

A series of horizontal bars of varying lengths and colors (teal, blue, and dark blue) are positioned on the left side of the slide, creating a modern, abstract background element.

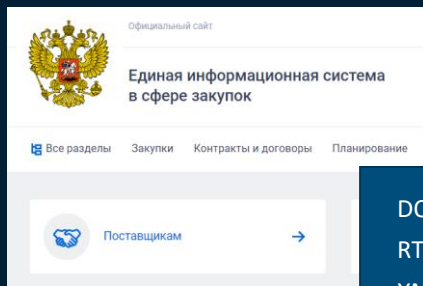
Demo

Протоколы собраний
EWS

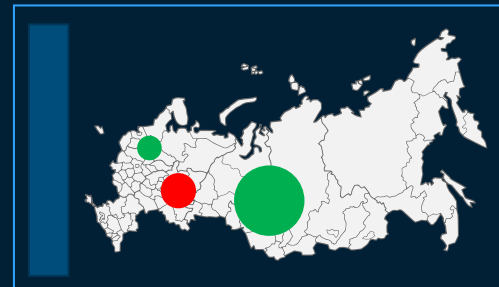
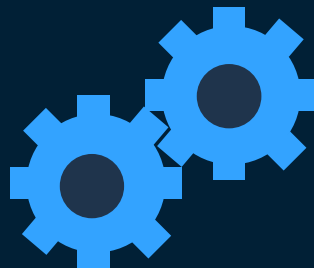


Кейсы: Тендерная аналитика

Изменение справочников



DOCX
RTF
XML



Сбор формализованных данных и закупочной документации

- Поставщики готовых данных
- Модуль сбора и очистки данных

Парсинг данных и расчет аналитических витрин

- SAS Data Quality
- Алгоритмы расчета показателей

Интерактивная отчетность, управление стратегией, сценарное моделирование

- SAS Visual Analytics

Кейсы: Анализ ответов учеников



Текстовая аналитика: а надо ли?

Нормальные бизнес-процессы –
в первую очередь

Технологии – потом

Вот потом она очень пригодится.

Немного помоделируем процессы

<https://precursorapp.com>*

sas.com

* не имеет отношения к SAS