ИНСТРУМЕНТЫ ТЕКСТОВОЙ АНАЛИТИКИ SAS



ЦЕННАЯ ИНФОРМАЦИЯ НЕ ИСПОЛЬЗУЕТСЯ

- Для построения более точных аналитических моделей требуется не только продвинутые алгоритмы, но и больше информации
- При этом в реальном бизнесе информации о клиентах много, однако существенная её часть представлена в неструктурированном тестовом виде (соцсети, комментарии на форумах, ...)
- При этом аналитические модели работают
 только со структурированной информацией

Системы поддержки принятия решений с учетом дополнительной текстовой информации могли бы быть существенно точнее



ИСТОЧНИКИ ТЕКСТОВЫХ ДАННЫХ

Внутренние

- 1. Анкеты и опросы
- Коллекции внутренней документации
- 3. Транскрипты разговоров из колл-центров
- 4. Корпоративная переписка и др

Внешние

- 1. Профили из соцсетей
- 2. Контент посещаемых сайтов
- 3. Информация из блогов и с тематических форумов
- 4. Судебные базы данных и др.



ЗАДАЧИ ТЕКСТОВОЙ АНАЛИТИКИ

Обогащение информации по клиентам

- Новые характеристики клиентов: интересы, события, социальный статус
- Повышение точности аналитических моделей и бизнес-правил

Мониторинг источников информации

- Исследование тематик новостей и дискуссий, контекста упоминаний продуктов компании и мнений об их характеристиках
- Оценка отношения к бренду, ключевым сотрудникам компании и работе филиалов
- Оценка отношения к брендов компаний-конкурентов
- Что думают? Почему? О чём? Кому передать? Как исправить?



ОБОГАЩЕНИЕ ИНФОРМАЦИИ ПО КЛИЕНТАМ

Интернет Социальные сети Поисковые запросы

Анкеты Транзакции Геолокация Колл-центр Жалобы Обращения

Важные события из жизни

Интересы Социальный сегмент

Общие факты:

- ✓ отдыхает на дорогих курортах
- ✓ увлекается алкоголем
- ✓ проявляет негатив к работе

Недавние/планируемые события:

- ✓ переезд, потеря работы
- ✓ свадьба, рождение ребенка

SAS TEXT ANALYTICS

Новые интерпретируемые характеристики клиентов

Повышение точности моделей

CRM, РИСК-МЕНЕДЖМЕНТ AHTИФРОД, COLLECTION



SAS Text Analytics

Text Mining

Категоризация
Извлечение фактов

Оценка эмоциональной окраски Анализ мнений

Статистический анализ текста

SAS Text Miner

Контекстно-семантический анализ Компьютерная лингвистика

> SAS Content Categorization SAS Contextual Analysis SAS Sentiment Analysis

SAS TEXT MINER СТАТИСТИЧЕСКИЙ АНАЛИЗ ТЕКСТА

TEXT MINING METHODOLOGY

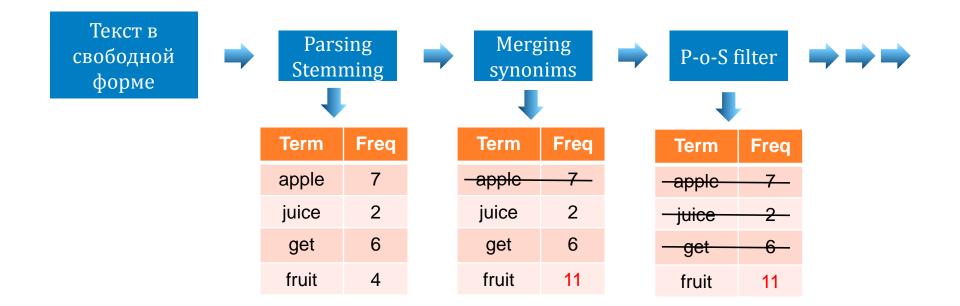
Импорт Лингвистический препроцессинг текста

Текстовое представление в числовое

Data Mining поверх числового представления



лингвистический препроцессинг



SAS TEXT MINER TOOLS

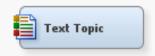
BASIC



Разбиение текста на термины, Р-о-S фильтр, синонимы, стоплисты



Взвешивание терминов, поиск синонимов, частотный фильтр



Выделение и интерпретация главных тем (SVD-разложение)

ADVANCED



Импорт больших коллекций длинных документов, определение языка текста

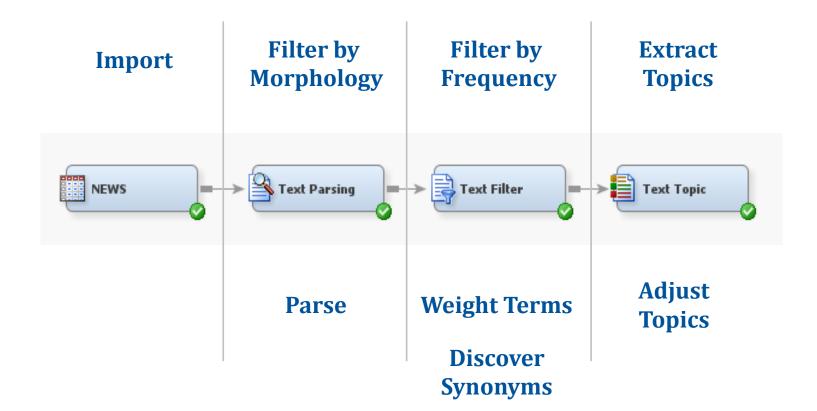


Построение Boolean правил категоризации



Поиск похожих по содержанию документов

SAS TEXT MINER TYPICAL WORKFLOW



SAS TEXT MINER TEXT PARSING NODE

TEXT PARSING NODE - P-o-S AND STEMMING

Part-of-Speech tagging – определение части речи терминов

- → исключение определенных частей речи из анализа
- → более качественный и быстрый **stemming**

Stemming – приведение терминов к начальным формам

- → качественное выделение главных тем коллекции документов
- → значительное ускорение последующего анализа текста



TEXT PARSING NODE - TERM FILTERING OPTIONS

Три способа фильтрации терминов в Text Parsing Node

1. Morphology-Based

Исключение определенных частей речи из анализа *Пример: исключаем все глаголы и предлоги*

2. Role-Based

Исключение терминов из анализа по их роли в предложениях Пример: исключаем все знаки пунктуации и аббревиатуры

3. Stop Lists

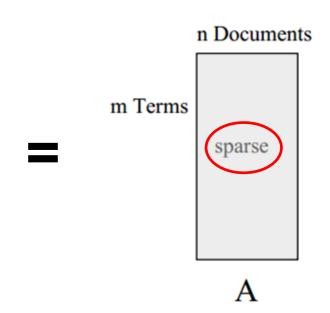
Явный список исключаемых терминов Пример: исключаем слова «apple» и «juice»



TERM-DOCUMENT MATRIX AFTER TEXT PARSING

Ha выходе Text Parsing Node – **RAW term-document matrix**

Terma	Documenta								
	cl	ď	c3	c4	сħ	шl	m2	ш3	m4
computer	1	1	0	0	0	0	0	0	0
EPS	0	O	1	1	0	0	0	0	0
human	1	0	0	1	0	0	0	0	0
interface	1	0	1	0	0	0	0	0	0
гевропае	0	1	0	0	1	0	0	0	0
ayatem	0	1	1	2	0	0	0	0	0
time	0	1	0	0	1	0	0	0	0
1186F	0	1	1	0	1	0	0	0	0
graph	0	0	0	0	0	0	1	1	1
minors	0	0	0	0	0	0	0	1	1
aurvey	0	1	0	0	0	0	0	0	1
trees	0	0	0	0	0	1	1	1	0



SAS TEXT MINER TEXT FILTER NODE

TEXT FILTER NODE – OVERVIEW

Функции Text Filter Node:

- 1. Spell Checking проверка орфографии
- 2. **Weighting** назначение терминам весов в зависимости от их встречаемости и силы связи с целевой переменной
- 3. Frequency-based term filtering
- 4. Interactive Synonyms Search инструменты для интерактивного создания словарей синонимов

TEXT FILTER NODE - SPELL CHECKING

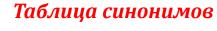
Редко встречаются **freq < 10**

Из словаря или часто встречаются freq > 6

~расстояние Левенштейна

Term	Parent Candidate	Distance
одноклассники	одноклассники	0
<mark>о</mark> дноклассники	0дноклассники	2
одн0клас_ники	одноклассники	6
одн0клас_ники	0дноклассники	10











Term	Parent		
одноклассники	одноклассники		
одноклассники	одноклассники		
одн <mark>0</mark> клас_ники	одноклассники		
одн <mark>0</mark> клас_ники	одноклассники		

TEXT FILTER NODE – TERM WEIGHTING

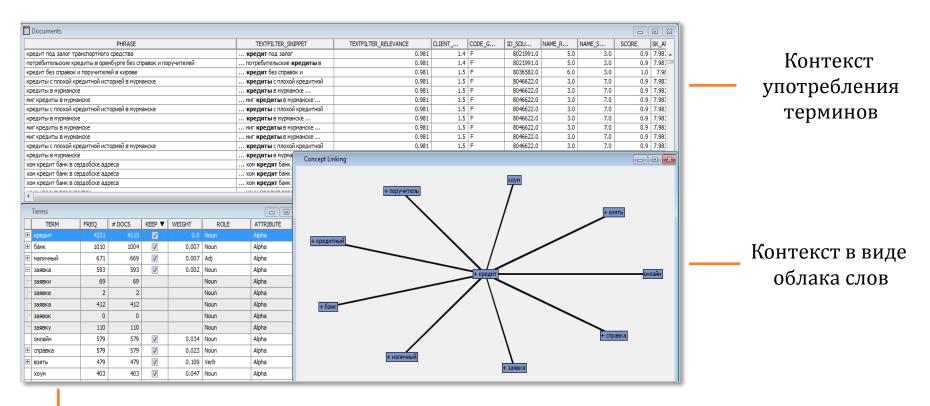
- 1. None = FREQ
- 2. Binary(FREQ)
- 3. Log(FREQ)

1. IDF
$$w_i = \log_2\left(\frac{1}{P(t_i)}\right) + 1$$

2. Enthropy
$$w_i = 1 + \sum_j \frac{(f_{ij}/g_i) \cdot \log_2(f_{ij}/g_i)}{\log_2(n)}$$

3. Mutual
$$w_i = \max_{C_k} \left[\log \left(\frac{P(t_i, C_k)}{P(t_i) P(C_k)} \right) \right]$$

TEXT FILTER NODE – INTERACTIVE FILTERING

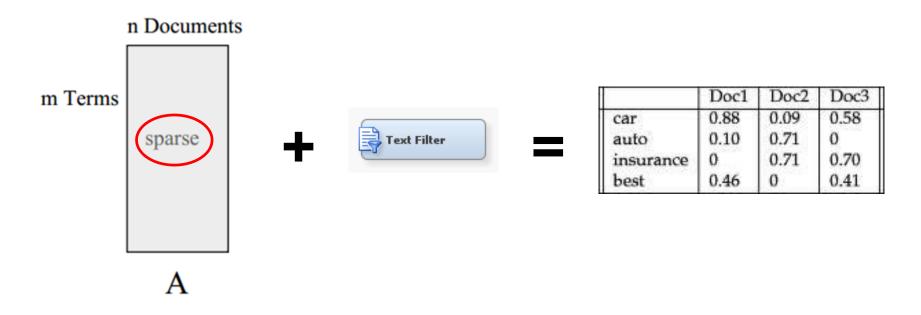


Интерактивное составление словарей синонимов / стоп-листов



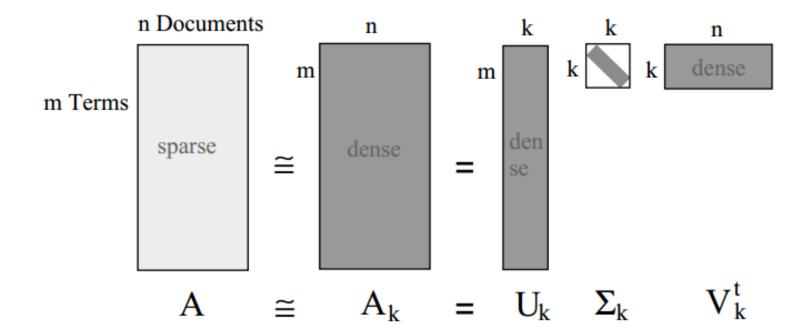
TERM-DOCUMENT MATRIX AFTER TEXT FILTER

Ha выходе Text Filter Node – WEIGHTED term-document matrix

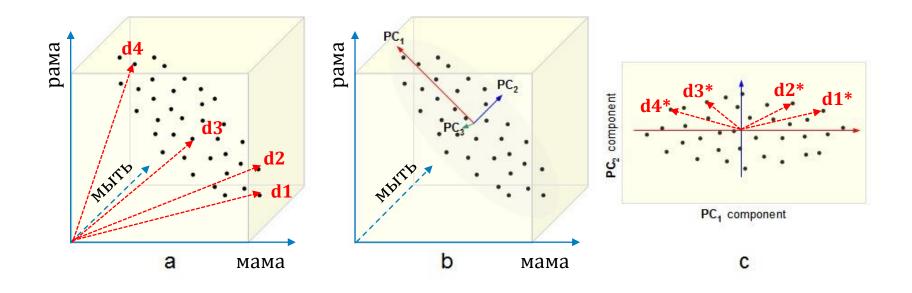


SAS TEXT MINER TEXT TOPIC NODE

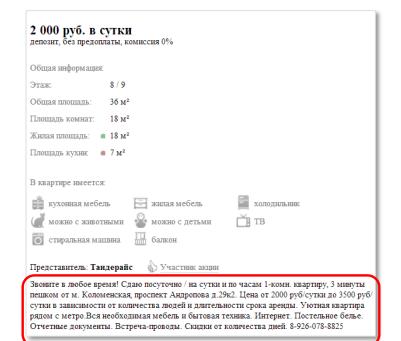
SVD DECOMPOSITION



INTUITION BEHIND SVD PROJECTION



TEXT TOPIC – SVD PROJECTION



Topic ID	Topic	Number of Terms	#Docs ▼
	1+школа,+сад,+детский,+детский сад,+магазин	143	4143
	8+минута,пешком,+станция,+метро,+станция метро	165	3432
	54+возможен,+торг,ипотека,срочно,ипотеке	141	2994
	25+хороший состояние,+хороший,+свободный продажа,+свобод□	112	2991
	14+комната,+больший,+просторный,+лоджия,+гардеробный	172	2989
	24+тихий,+зеленый,+зелёный двор,+двор,+тихий зелёный двор	78	2770
	29+документ,+гот,+сделка,+первичный,срочно	75	2605
	10 ул,д.,+1-ком/квартира,/у,м.	245	2577
	23+мебель,+качественный,+бытовой,+техник,техникой	223	2575
	30+чистый подъезд,+подъезд,+сосед,+хороший сосед,+приличн□	133	2567
	20+пол,+плитка,ламинат,+паркет,+потолок	163	2550
	47 +уютный, +светлой, +уютный квартира, +светлой квартира, оче	120	2515
	66+транспортный,шаговой,+доступность,+сообщение,+транспо	220	2514
	15+взрослый собственник,+взрослый,+собственник,+один,+при	26	2484
	65+балкон,застеклен,+раздельный,+застекленный балкон,+ст□	200	2404
	39+москва,+новый,+московский,км,+лес	299	2380
	32+встроенный,+встроенный кухня,+шкаф,+кухня,+встроенный □	91	2355
	45+застекленный,+лоджия,+застекленный лоджия,+больший,+бо	110	2350
	7+площадь,кв,+этажный,+общий,+однокомнатный	174	2336
	9+огороженный,+территория,+подземный,+паркинг,+подземны□	153	2325
	33+вид,+панорамный,+панорамный вид,+прекрасный,+окно	124	2118
	493-х,менее,+собственность,+договор,+недвижимость	205	2100
	6+машина,+стиральный,+мебель,+стиральный машин,+холодиль	226	2062
	36+показ,+время,+любой,+любой время,+звонить	229	2015
	27+альтернатива,+легкий,+подобран,+лёгкий альтернатива,д□	81	2005



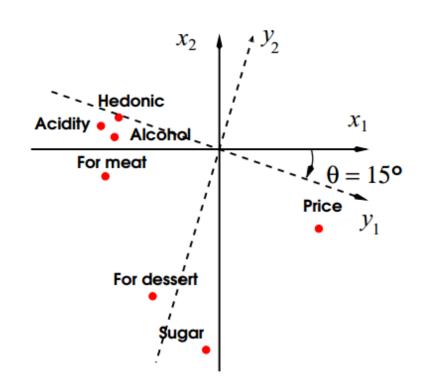
TEXT TOPIC - SVD ROTATION

SVD Rotation - алгоритм повышения интерпретируемости тем:

1. Преобразовать координаты

$$(x, y)_{NEW} = (x^3, y^3)_{OLD}$$

2. Рассчитать SVD на $(x, y)_{NEW}$

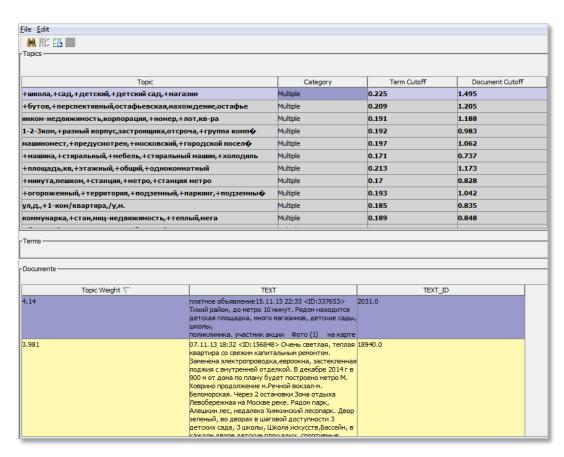


TEXT TOPIC – INTERACTIVE TOPIC VIEWER

✓ Экспертная доработка автоматически рассчитанных тем

✓ В каких темах «засветился» конкретный термин?

✓ Какие документы «засветились» в конкретной теме?

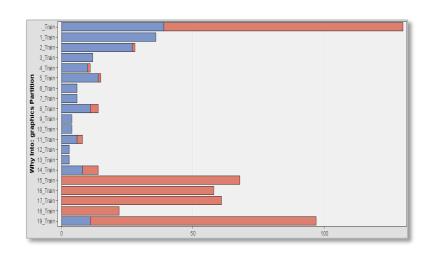




SAS TEXT MINER ADVANCED NODES

TEXT RULE BUILDER

Text Rule Builder строит Boolean правила категоризации документов



```
Content Categorization Code
     F graphics =1 ::
     (0R
     , "graphics"
     , (AND, (OR, "images", "image"))
     , (AND, (OR, "algorithm", "algorithms"))
     , "graphics"
     , (AND, (OR, "programs", "program"), (NOT, (OR, "year", "years")))
     , (AND, (OR, "defines", "defined", "define"))
     , "vesa"
     , "software"
     , (AND, (OR, "fractal", "fractals"))
     , "impulse"
     , (AND, (OR, "file", "files"))
     , "yeh"
     , (AND, (OR, "sphere", "spheres"))
     , "email" )
     F graphics =0 ::
     , (AND, (OR, "team", "teams"))
     , "gordon"
     , (AND, "ca" , (NOT, (OR, "chip" , "chips" )))
     , (AND, (OR, "sciences", "science"))
     , (AND, ">" , (NOT, (OR, "mode" , "modes" ))))
```

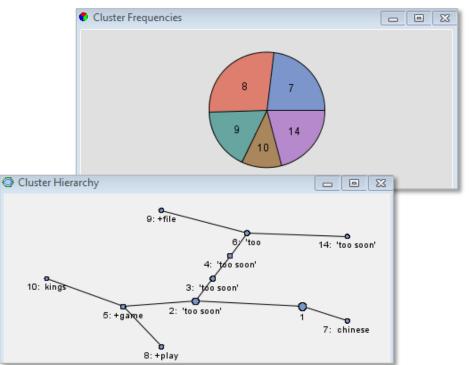
TEXT CLUSTER

Text Cluster разбивает документы на кластеры для:

- предварительной сегментации объектов перед построением моделей
- выявления похожих документов (например, для антиплагиата)

Алгоритмы:

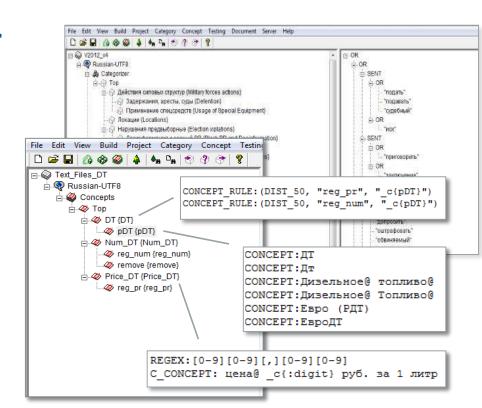
- 1. Expectation-Maximization
- 2. Hierarchical Clustering



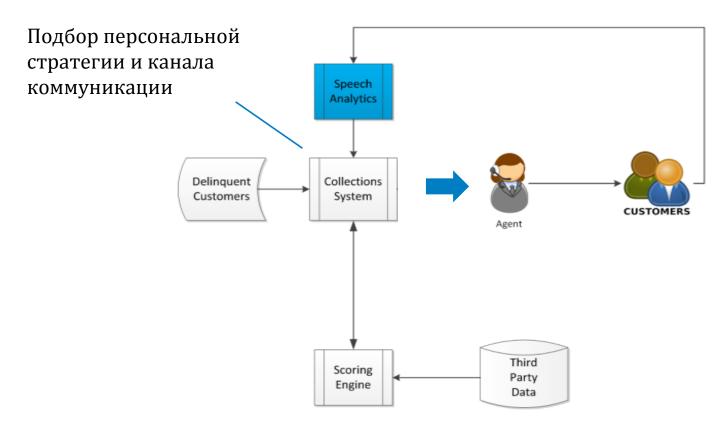
SAS ENTERPRISE CONTENT CATEGORIZATION КОНТЕКСТНО-СЕМАНТИЧЕСКИЙ АНАЛИЗ ТЕКСТА

SAS ENTERPRISE CONTENT CATEGORIZATION

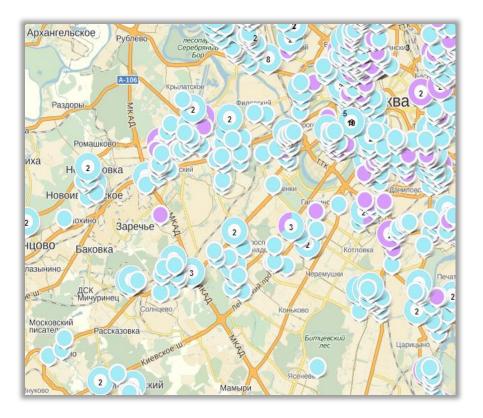
- Специализированный инструмент для контекстно-лингвистического разбора текста
- Коллекция линвистических, для выделения объектов, объектов и событий
- Иерархическая организация правил для ускоренной разработки
- ➤ Быстрый старт работ при наличии SAS Text Miner



ОПТИМИЗАЦИЯ РАБОТЫ КОЛЛЕКШЕН-ЦЕНТРА



ЦЕНОВЫЕ МОДЕЛИ НЕДВИЖИМОСТИ



Зачем строить?

- ГосорганыПланирование городской застройки
- Финансовый сектор
 Проверка стоимости залоговой недвижимости

ЦЕНОВЫЕ МОДЕЛИ НЕДВИЖИМОСТИ

Структурированная часть

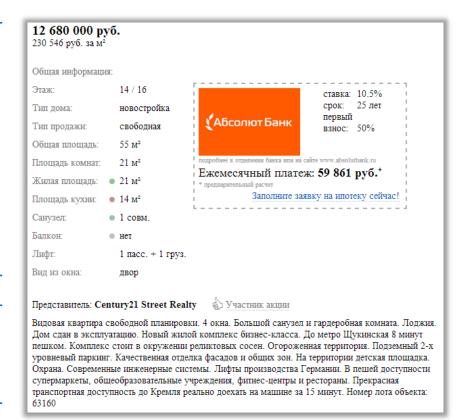
Только базовая информация об объекте

→ Модели начальной точности

Неструктурированное текстовое описание

Важные характеристики объектов!

→ Модели точнее



ЦЕНОВЫЕ МОДЕЛИ НЕДВИЖИМОСТИ

1. Text Mining

Какие свойства недвижимости риэлторы считают важными?

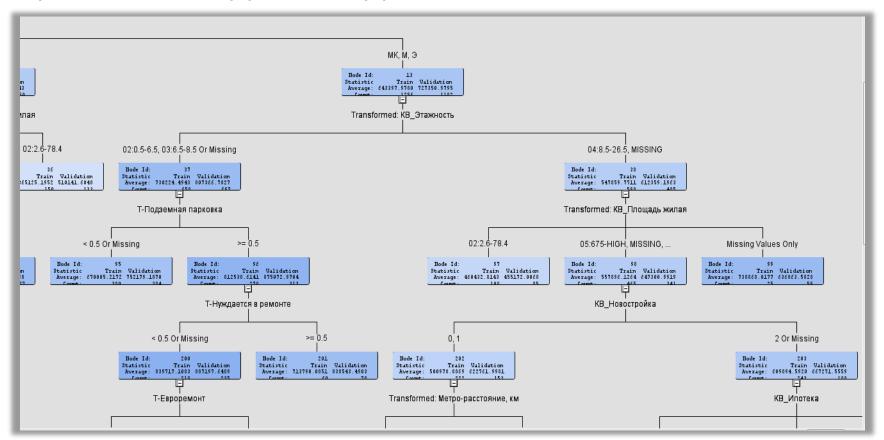
Торіс	# Docs ▼
+школа,+сад,+детский,+магазин,+детский сад	3846
+минута,пешком,+станция,+метро,м.	3398
+комната,+просторный,+больший,+гардеробный,+кухня	3237
+площадка,+детский,+детский площадка,+спортивный,+терри	3141
дкп,3-х,менее,+собственность,ипотека	3065
+новый,+муниципальный,+москва,+сантехник,+новый дом	3052
+установлен,+пластиковый,шаговой,+станция,+тереть	2942
+тереть,кв,+встроенный кухня,+кухня,+один	2908
+тихий,+зеленый,+двор,+зелёный двор,+место	2864
+торг,+возможен,срочно,ипотека,ипотеке	2863
+мебель,+бытовой,техникой,+евроремонт,+встроенный	278
+потолок,+евроремонт,+натяжной,+натяжной потолок,+высот	2764
м.,ул,кв,д.,/у	2760
+торговый,+торговый центр,+1-ком/квартира,+ходьба,+минут□	2747
+пол,+плитка,ламинат,+паркет,+стеклопакет	2746
шаговой,+доступность,3-х,+инфраструктура,+магазин	2738
+комплекс,+возможен,ипотека,+жила,+быть	2725

2. Deep Linguistic Analysis

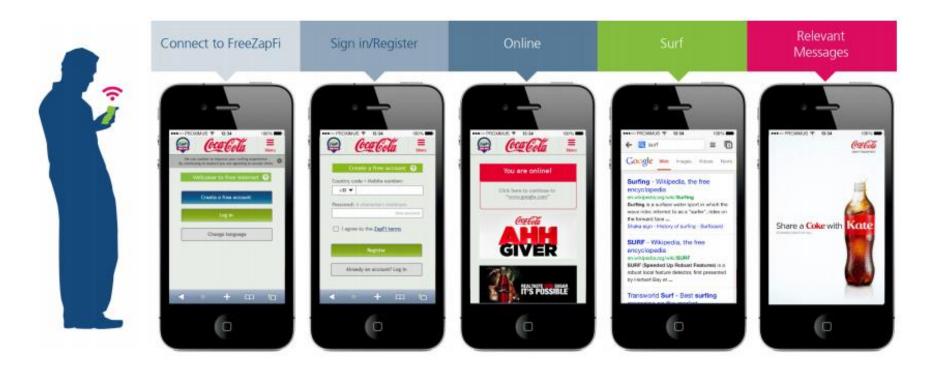
Какими свойствами обладает данный конкретный объект?

- ✓ смежные комнаты
- ✓ подземная парковка
- ✓ окна во двор
- ✓ детсад и школа

ЦЕНОВЫЕ МОДЕЛИ НЕДВИЖИМОСТИ







Платформа SAS: анализ клиентский траффика в режиме реального времени

Продвинутая и текстовая аналитика для построения детального профиля каждого клиента

Анализ текстового содержимого посещенных сайтов существенно дополняет понимание потребностей и профиля интересов клиента



For Consumers

 Sponsored Free Wi-Fi (fast & secure)

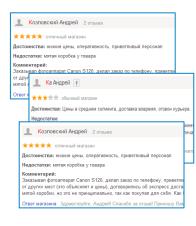
For Brands, Merchants & Operators

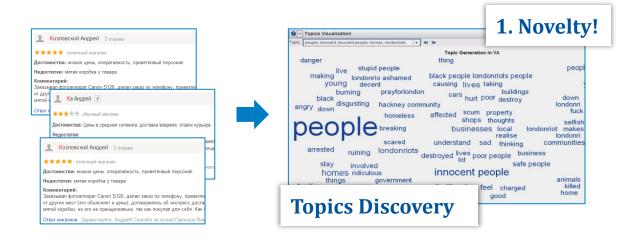
Mobile Marketing Platform (interaction, insights & mobile campaigns)













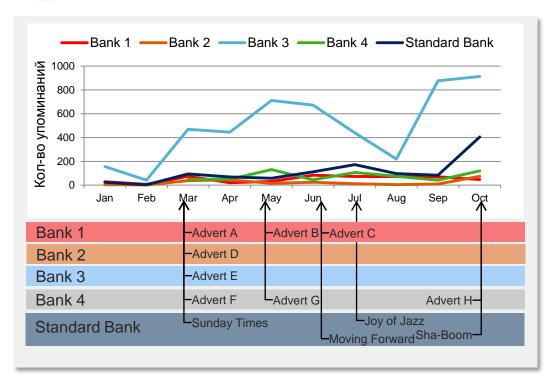


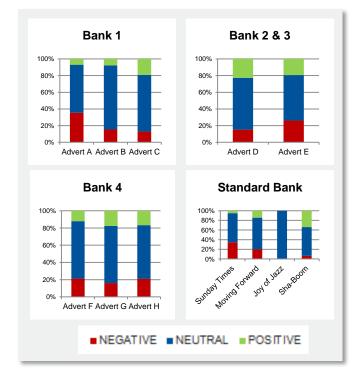
Эффект: устранение дефектов в продуктах ещё до того, как поступит первый гарантийный талон!



STANDARD BANK: МОНИТОРИНГ КОНКУРЕНТНОЙ СРЕДЫ

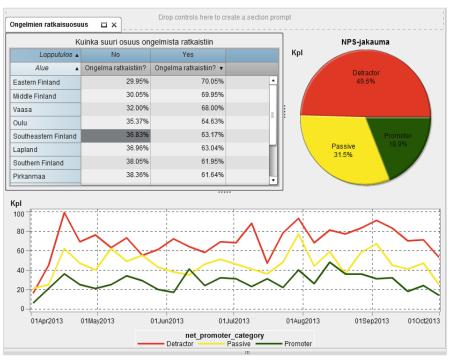


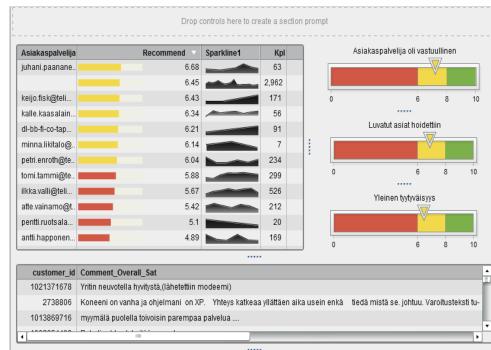




TELIA SONERA: CUSTOMER EXPERIENCE MONITORING

РАЗГОВОРЫ ИЗ КОЛЛ-ЦЕНТРОВ







СПАСИБО ЗА ВНИМАНИЕ!

