SentiScan

система автоматического анализа тональности в social media

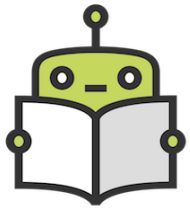
AI '14

whois

Леонид Литвиненко
CTO

&

Дмитрий Кан
CEO



SemanticAnalyzer
Language intelligence

YouScan: система мониторинга соцмедиа

Мониторинг ВК, Facebook, Twitter, Instagram, форумов, сайтов отзывов

Встроенная аналитика и инструменты командной работы

Используют отделы маркетинга, PR, поддержки клиентов крупных компаний

Лингвистические алгоритмы и API

www.semanticanalyzer.info

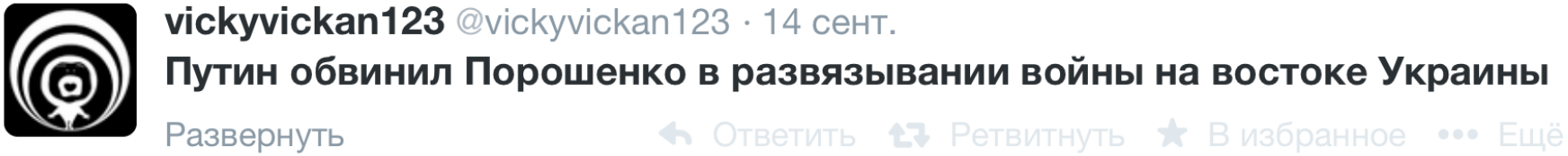
dk@semanticanalyzer.info

SemanticAnalyzer
Language intelligence

outline

- Постановка задачи
- Теоретические границы
- Why not machine learning
- Алгоритм
- Типичные проблемы
- Метрики качества
- Continuous integration
- Система оценки
- Планы на будущее

пример



Object: Путин

Sentiment tag?

неоднозначность постановки задачи

- Настроение автора?
- Настроение читающего?
- Личное отношение к теме?
- Что-то ещё, что предписал заказчик?

тональность для компьютера

- Общий тональный фон
- Извлечение мнений
- Извлечение мнения по конретному объекту
- Определение силы тональности или эмоции

от задачи к решению: уровень качества

MTurk, ручная аннотация: 80% согласия [1]

Реальное качество -- то, которое система показывает на неаннотированных данных

why not machine learning?

Некоторые методы дают 90% на тестовых данных

why not machine learning?

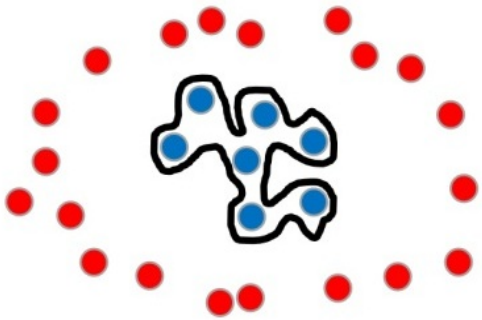
Сложно (невозможно?) сделать таргет-ориентированный метод

why not machine learning?

CEO of the company turned 50.

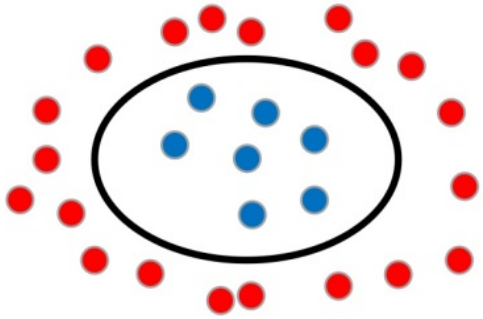
переобучение моделей

Переобученная модель



[2]

Оптимальная модель



потенцильно работающие методы

Naive Bayes -- для которотких сообщений (твитов) [3]

SVM -- для длинных (блог пост или статья)

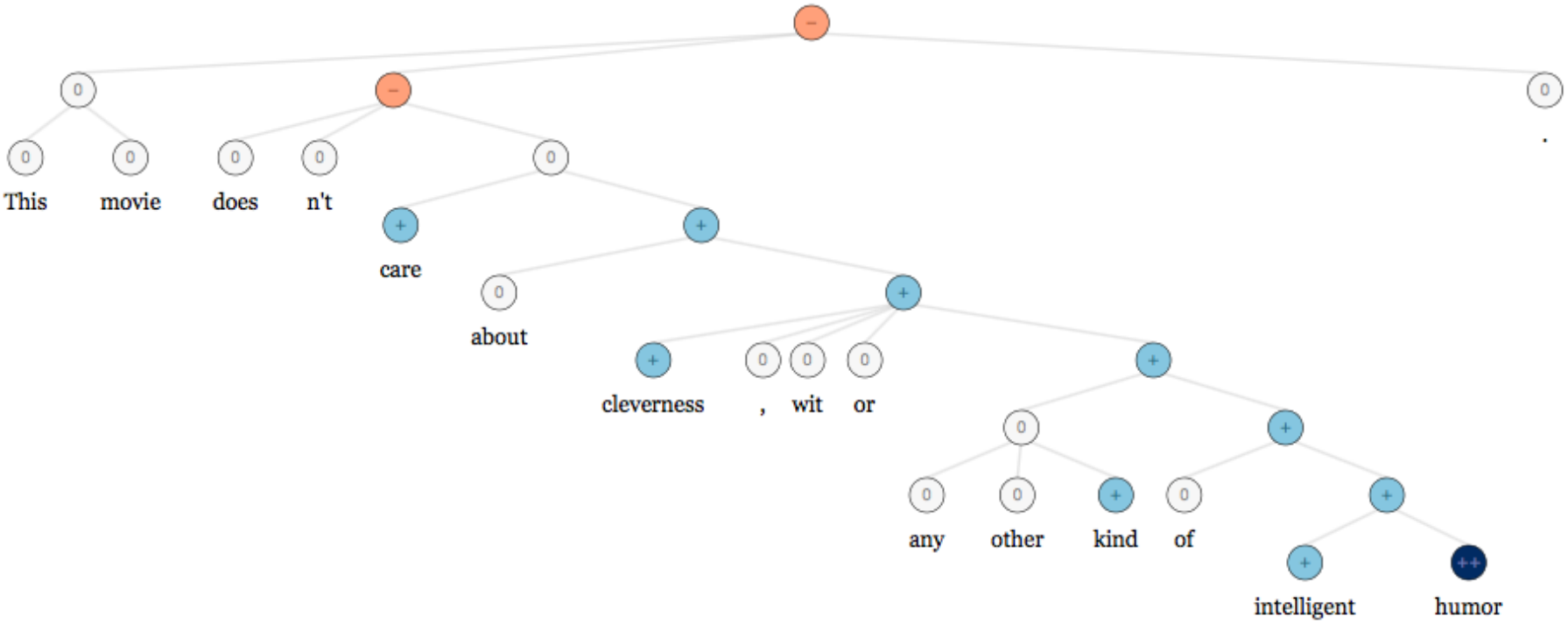
свежие веяния: deep learning

Back to the future: нейронные сети из 80-х [4]

Многонейронная логистическая регрессия

Требует больших вычислений

Не требует экспертных знаний


This
movie
does
n't
care
about
cleverness
,
wit
or
any
other
kind
of
intelligent
humor
.

[5]

тональность в реальной жизни

Контроль над качеством

Быстрая аннотация

Высокая точность (precision) при возможных потерях в выдаче

алгоритм (sketch)

- Лемматизация
- Полярные лексиконы
- Правила для русского языка
- Shallow parsing + анафора
- Учёт опечаток (Levenstein)

полярные лексиконы

- Автоматический mining из twitter: :) & :(
- Ручная вычитка частотных таблиц n-грамм
- Отбор полярно-однозначных кандидатов

полуавтоматическое расширение

"классный и"

Web Images Videos News Maps More Search tools

Page 2 of about 687,000 results (0.21 seconds)

Очень классный и удобный зимний комплект для ... - Io.ua
yuliyarom.io.ua/album693682 Translate this page
фото Очень **классный и** удобный зимний комплект для мальчика- аналог Benetton (2014-2015 гг) :Детские товары.

ОН КЛАССНЫЙ И ПРАВИЛЬНЫЙ В НЕМ ЕСТЬ ... - Дом-2
dom2.ru/community/113164710/121011652 Translate this page
Sep 28, 2014 - Анна Кручинина / АНЯ НАЧНИ ОТНОШЕНИЯ С НОВЕНЬКИМ ВАДИМОМ!ОН **КЛАССНЫЙ И** ПРАВИЛЬНЫЙ В НЕМ ЕСТЬ ПОТЕНЦИАЛ!

Ты классный и добрый,а на красоту пофиг,(хотя ... - Ask.fm
ask.fm/yt_laggerfeed/answer/111535494389 Translate this page
Ты **классный и** добрый,а на красоту пофиг,(хотя ты красивый) я помню всю нашу переписку. а вообще,снимай больше видео,радуй нас).

Vatzivassiloglou & McKeown, 1997

правила

не / без / нет / … + {LEXICON_ENTRY}
=> инвертированная полярность

правила

а / но / хотя / зато / ... - противопоставление полярностей левой и правой части

правила: мультипасс оконный алгоритм

без, не, еле, ни,
нет (проблем)

shallow parsing

Противительный союз

Отрицание

Большинство любит это, а мне это не нравится.

Score = totalPosScore - totalNegScore - { 0 | ½ * sentimentCount }

shallow parsing

а, но, зато,
только, хотя

Противительный союз

Мне понравился новый iPhone, а GalaxyS неудобен в использовании.

Score = totalPosScore - totalNegScore - { 0 | ½ * sentimentCount }

schematic flow on sentence level

Большинство любит это, а мне не нравится.

Phase 1 (negations): posScore = 0 - NEGATION_WEIGHT= -2

Phase 2 (individual words):
“любит” posScore = -2 + 1 = -1
“не” negScore = 0 + 1 = 1
“нравится” posScore = -1 + 1 = 0

Phase 3 (opposite conjunctions): sentimentCount = 3

totalScore = posScore - negScore - ½ * sentimentCount =
= 0 - 1 - 3/2 = -5/2

Sentiment: NEGATIVE

анафора

Джон встретил сегодня Ивана. Он передал Ивану текст договора.

анафора

Джон встретил сегодня Ивана. Он передал Ивану текст договора.

структурный подход

$\exists$ местоимение (м.р., ед.ч., им.п.) $\subset P_2$ □ (Джон $\not\subset P_2$) => Он ⬌Джон

Struct 1		Struct 2
{		{
Джон	→	Он
встретить		передавать
Иван	→	Иван
}		}

анафора: симметрия

Джон встретил Ивана. Тот передал Джону ответ директора.

анафора: симметрия

Джон встретил Ивана.

Тот передал Джону ответ директора.

поиск ссылки

$\exists$ местоимение (м.р., ед.ч., им.п.) $\subset P_2$ □ (Иван $\not\subset P_2$) => Тот $\Leftrightarrow$ Иван

Struct 1

{

Джон

встретить

Иван

}

Struct 2

{

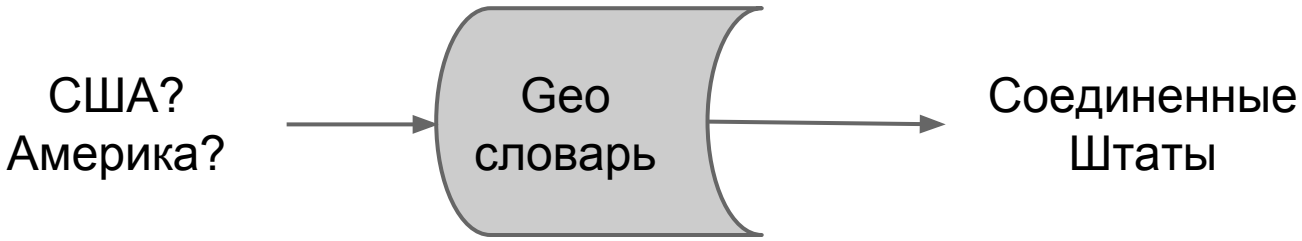
Тот

передавать

Джон

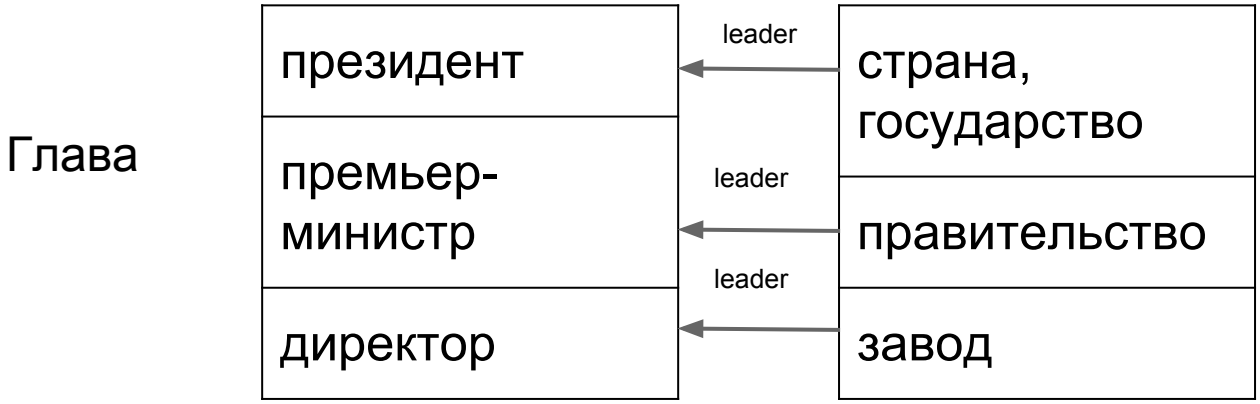
}

гео-словарный подход (Interfax)



Глава Америки… Сегодня президент США Б. Обама

[7]



типичные проблемы

Мне от похмелья помогают ужасные *бичевые* кружки магги с сухариками, открыла чудо

Алгоритм: NEUTRAL

Ручная: NEGATIVE

Исправление: пополнить словарь лемматизатора словами бичевый (-ая, -ое, ые)

типичные проблемы #2

своих близких. Мы знаем несколько мест, где можно это сделать в Апатитах: - Магазин электротоваров “АКЦЕНТ+” на ул. Дзержинского, 35 (брелоки). - Гипермаркет “Твой” (дорожные жилеты, брелоки). - Магазин “Зооландия” (майки, шлейки, ошейники для собак). - Магазин “Спортмастер” (браслеты, полоски на велосипедные спицы). - Магазины “Евророс” (жилеты, брелоки). - Книжный магазин “Студент” на ул. Бредова, 13 (брелоки). - Обувной отдел в магазине “Силуэт” (браслеты, брелоки). - Магазин игрушек “Фантик” (брелоки). - Заправка “Статойл” (дорожные жилеты ...

Алгоритм: NEGATIVE

Ручная: NEUTRAL

Исправление: В негативный словарь закрались слова: “близкий” и “близко”. Их удаление исправило ситуацию.

сложный случай

Ужасно замерзли(сейчас в Мак Кафе пьем горячий шоколад, ммм ... наслаждение)

Проблема: missing sentence boundary => sentiment overlap

пополнение базы лексем

пичаль	печаль
инет	интернет
ваще	вообще

качество =

какие задачи решает клиент

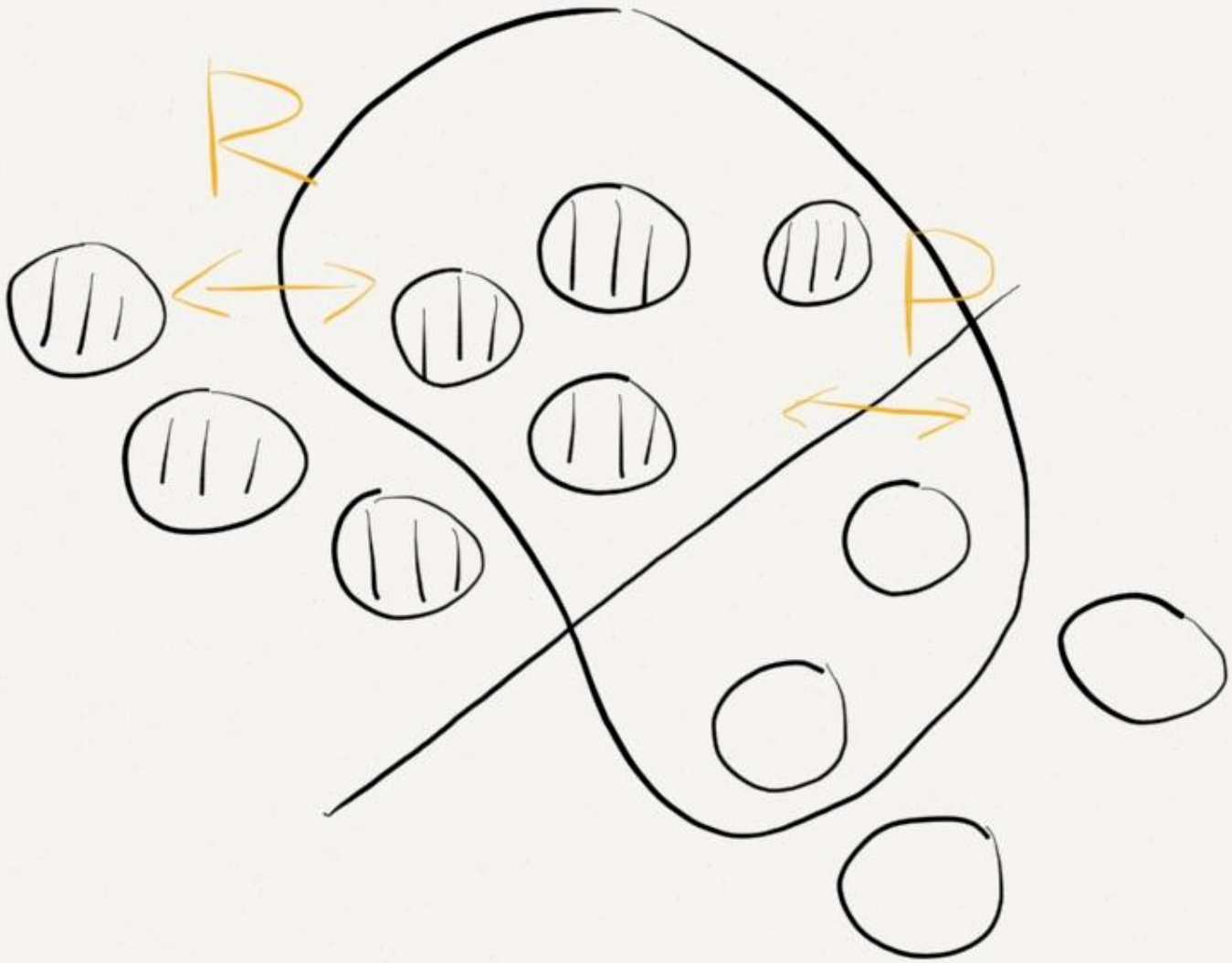
анализ отношения к продуктам

реагирование

какие задачи решает клиент

общая картина (recall)

точный негатив (precision)


R
P

$$\text{качество} = \frac{\text{max precision}}{\text{max recall}}$$

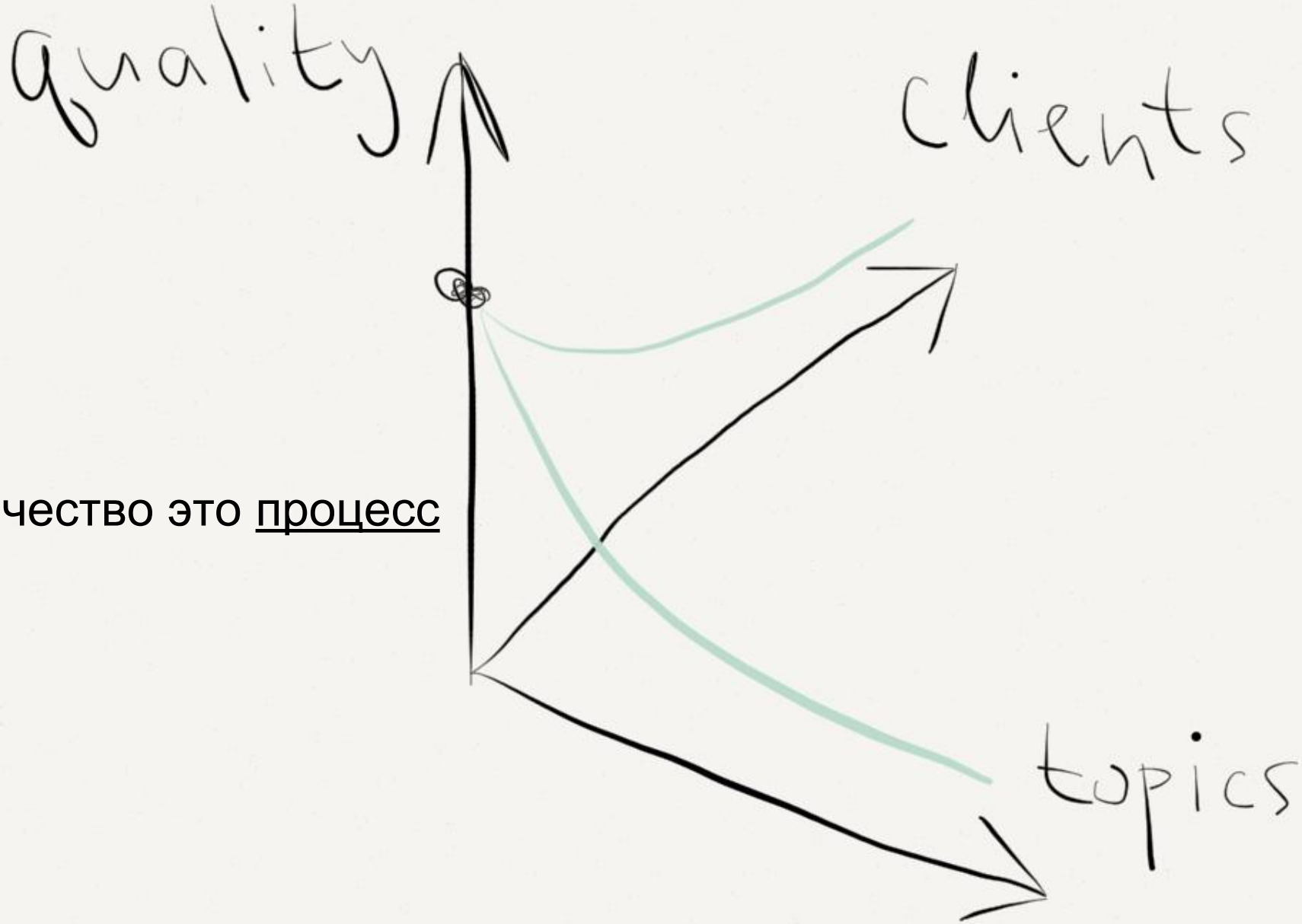
качество = max F (P, R)

$$F(P, R) = 2P * R / (P + R)$$

качество это не только

max F (P, R)

качество это <u>процесс</u>

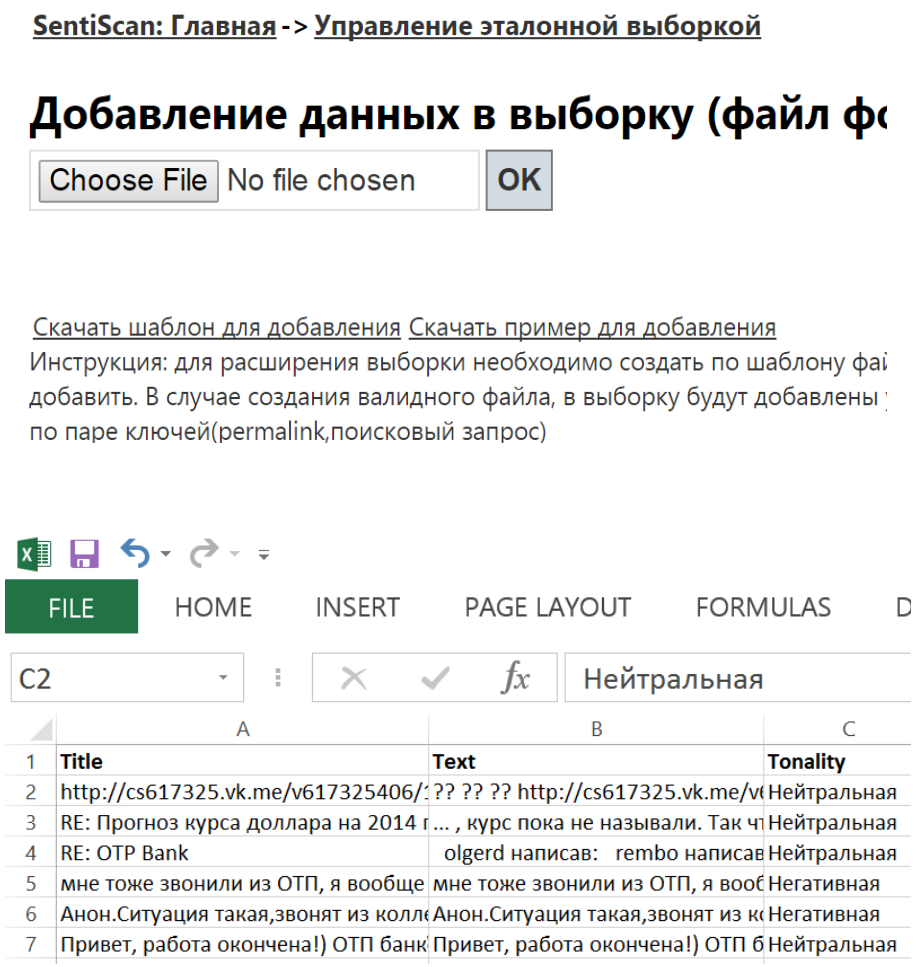


инструментарий

- эталонная выборка
- QA для новых версий
- A/B тесты конфигураций
- ошибки определения контекста
- поток исправлений от клиентов

эталонная выборка

- согласие в оценке 3-х лингвистов
- легко пополнять



QA новых версий

- по тональности
- конкретные тексты

Версия/Процесс	Accuracy				Precision	
	all	Pos	Neut	Neg	Pos	Neut
1.280 / 3	0.53	0.55	0.55	0.36	0.31	0.79
1.270 / 3	0.52	0.54	0.54	0.36	0.31	0.78
1.260 / 3	0.52	0.54	0.54	0.36	0.31	0.78
1.250 / 3	0.52	0.54	0.54	0.36	0.31	0.78
1.240 / 3	0.52	0.54	0.54	0.36	0.31	0.78
1.230 / 3	0.51	0.51	0.54	0.38	0.30	0.78

Версия: 1.26 Процесс: 3

Поисковый запрос		Accuracy				Precision	
ID	Текст	all	Pos	Neut	Neg	Pos	Neut
271	(McCafe или maccafe или "маккафе" или "мак кафе" или маккафэ или "мак кафэ")	0.63	0.6	0.7	0.32	0.76	0.69
272	(укрсиббанк или ukrsibbank или укрсіббанк или укрсиб или укрсіб или ukrsib или ukrsyb)	0.49	0.67	0.5	0.38	0.16	0.83
273	(VICHY или vishy или виши или вичи или vichi или Normaderm или нормадерм	0.42	0.51	0.42	0.24	0.21	0.83

todo

- те же метрики, разные тексты

A/B тесты конфигов

- title & text
- emoticons
- words boosters

Версия: 1.26 Процесс: 3

Поисковый запрос		Accuracy				Precision	
ID	Текст	all	Pos	Neut	Neg	Pos	Neut
271	(McCafe или maccafe или "маккафе" или "мак кафе" или маккафэ или "мак кафэ")	0.63	0.6	0.7	0.32	0.76	0.69
272	(укрсиббанк или ukrsibbank или укрсіббанк или укрсиб или укрсіб или ukrsib или ukrsyb)	0.49	0.67	0.5	0.38	0.16	0.83
273	(VICHY или vishy или виши или вичи или vichi или Normaderm или нормадерм	0.42	0.51	0.42	0.24	0.21	0.83

todo

- levenstein typos
- версии словарей

определение контекста

- ошибка лемматайзера

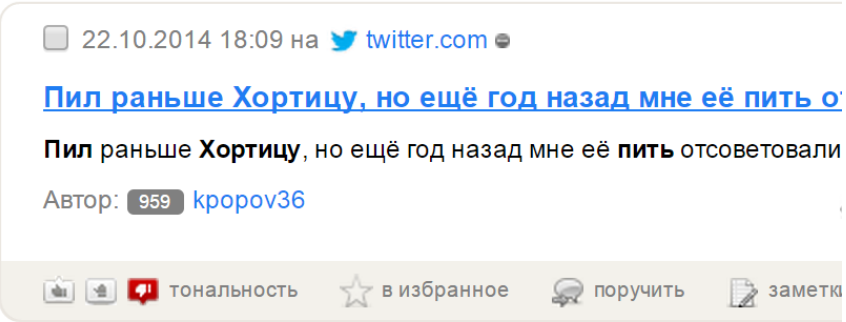
Accuracy			Precision			Синонимы
Pos	Neut	Neg	Pos	Neut	Neg	
0.6	0.67	0.3	0.74	0.66	0.11	0.78
0.6	0.47	0.4	0.19	0.78	0.35	0.75

todo

- тест с другими лемматайзерами

поток исправлений

- инсайты

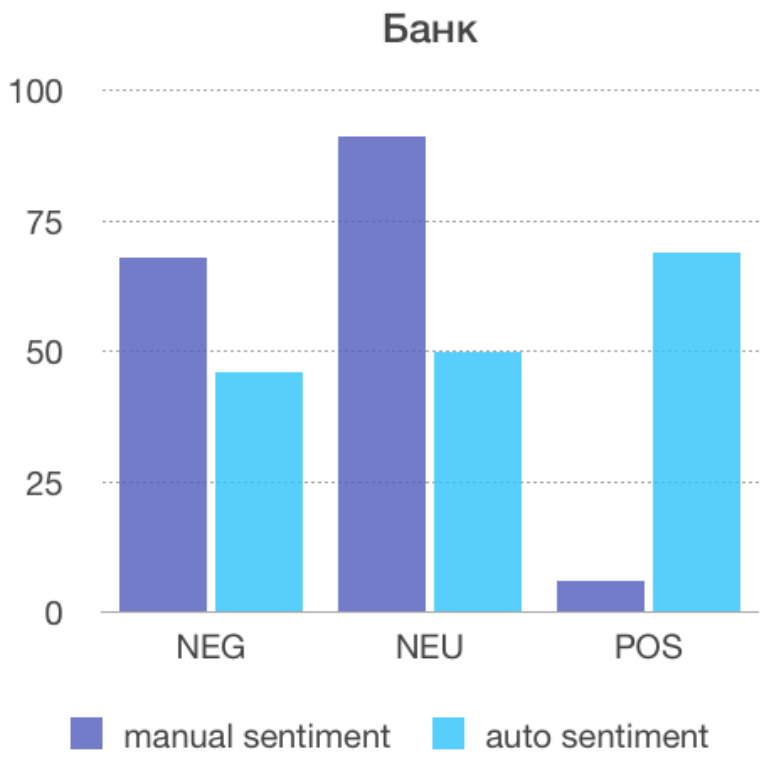


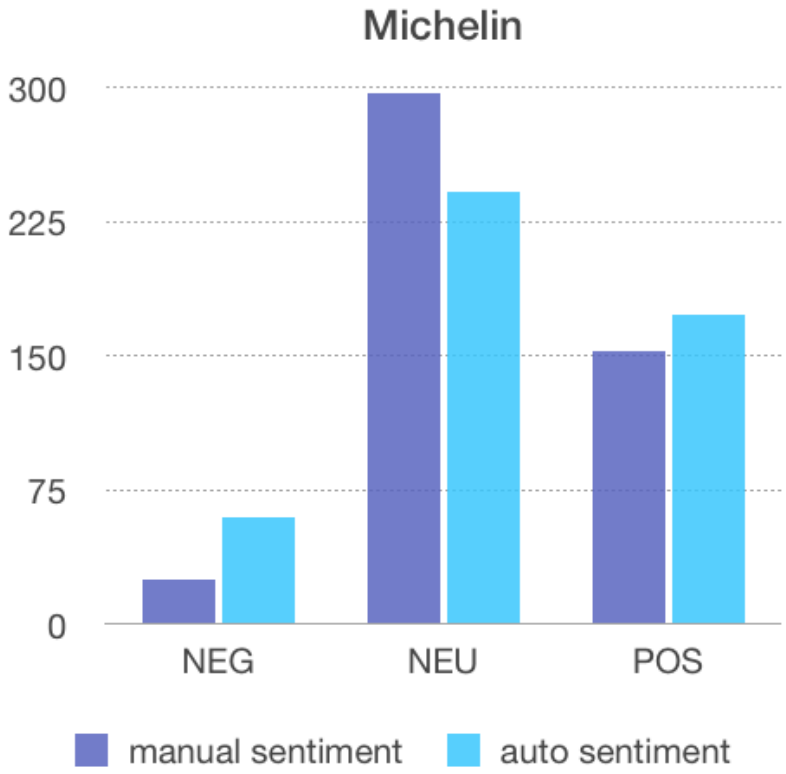

```
"mention": {
  "id": 537034125,
  "title": "<b>Пил</b> раньше <b>Хортицу</b>, но ещё год назад мне её
  "text": "<b>Пил</b> раньше <b>Хортицу</b>, но ещё год назад мне её
  "url": "http://twitter.com/KPopov36/statuses/524925590806462464",
  "date": "2014-10-22T18:09:53.4122217",
  "manualSentiment": -1,
  "autoSentiment": -1
},
```


todo

- больше инсайтов
- авто улучшение

human vs algorithm





Банк

и постригут

какого лоха

дополнительный доход

эти ребусы

обнулить абонку

Michelin

бешено дрифтит

единственная компания

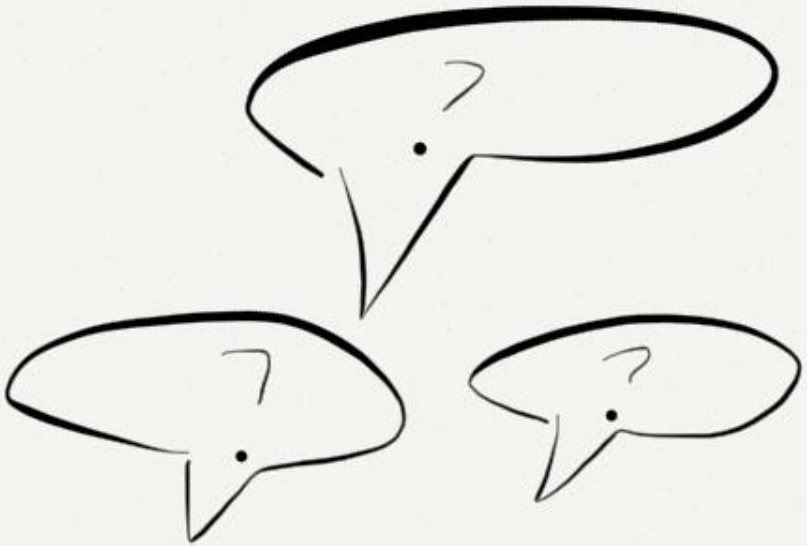
колёса tweel

пневма рулит

nokian hakkapeliitta

больше инсайтов

- частотные биграммы -> кандидаты в полярный словарь
- какое направление исправлений превалирует
- moving average precision/recall -> сигнал для движка



@dmitrykan
@llytvynenko

we're hiring :)

Литература

[1] Bermingham, A. and Smeaton, A.F. (2009). A study of interannotator agreement for opinion retrieval. In SIGIR, 784-785.

[2] http://habrahabr.ru/company/yandex/blog/206058/

[3] http://habrahabr.ru/post/229779/

[4] http://vimeo.com/77050653

[5] http://nlp.stanford.edu/sentiment/

[6] Hatzivassiloglou & McKeown: **Predicting the semantic orientation of adjectives. 1997**

[7] Zharikov A et al. Information Retrieval System for News Articles in Russian // The joint RuSSIR/EDBT 2011 Summer School, pp. 5-14. August 15-19, 2011