Яндекс

Твиттер в поиске Яндекса

Иван Комаров

разработчик

Я.Субботник в Екатеринбурге, 6 июля 2013

Мотивирующий пример

26 октября 2012 года пользователям Рунета стал доступен сайт zapret-info.gov.ru. Мы хотим его как можно быстрее начать показывать по запросу **[реестр запрещённых сайтов]**.

Какие данные мы можем для этого использовать?

Ссылки на страницу

- Страница не очень быстро обрастает ссылками.
- Тексты ссылок малоинформативны.

Роскомнадзор напоминает о вступлении в силу с 01 ноября 2012 года изменений в Федеральный закон 149-ФЗ "Об информации, информационных технологиях и защите информации", сообщается на сайте "Единого Реестра доменных имен, указателей страниц сайтов в сети "Интернет" и сетевых адресов, позволяющих идентифицировать сайты в сети "Интернет", содержащие информацию, распространение которой в Российской Федерации запрещено" Zapret-info.gov.ru.

"Единый реестр доменных имен, указателей страниц сайтов в сети "Интернет" и сетевых адресов, позволяющих идентифицировать сайты в сети "Интернет", содержащие информацию, распространение которой в Российской Федерации запрещено", расположен на сайте http://zapret-info.gov.ru/.

Накануне ведомство запустило портал, на котором пользователи смогут жаловаться на интернет-ресурсы. Кроме того, был опубликован порядок ведения реестра сайтов с

будет проверить через специальный ресурс, открытый Роскомнадзором. Но сделать выгрузку всего реестра вправе

Текст и URL страницы

- Мало текста, слова запроса разбросаны по странице.

ЕДИНЫЙ РЕЕСТР
доменных имен, указателей страниц сайтов в сети «Интернет»
и сетевых адресов, позволяющих идентифицировать сайты в сети «Интернет»,
содержащие информацию, распространение которой
в Российской Федерации запрещено

- Мы практически ничего не знаем про URL и домен.

Твиттер

- Ссылки на интересные страницы появляются быстро.
- Каждая ссылка снабжена компактным описанием.

Yakimets Roma
@yakimetsr

Реестр запрещенных сайтов: zapret-info.gov.ru Тут можно узнать запрещен тот или иной сайт в Инете. Ну и настучать на сайты можно тоже там.

Evgeny Vilkov
@EvgenyVilkov

реестр запрещенных сайтов будет размещен по адресу zapret-info.gov.ru. Хорошее дело в домене gov не разместишь!

Pavel Plotnikov
@PavelPV

Реестр запрещённых сайтов в Российской Федерации (открыт частично) zapret-info.gov.ru

RussianNewsline
@RussianNewsline

Реестр запрещенных сайтов обзавелся адресом: На сайте zapret-info.gov.ru можно будет узнать, какие веб-страницы... bit.ly/XrqxoF

Что дальше?

- Я расскажу про **5** задач, которые нам пришлось решать.
- Попробую обойтись без закапывания в технические детали.
- Эти задачи — только верхушка айсберга.

I. Распознавание

языков

Twitter Firehose

- Шланг со всеми твитами в мире.
- Типичные объёмы за день:
  - **500 млн** записей в JSON-формате (включая служебные).
  - **1,3 ТБ** в несжатом виде, **200 ГБ** в сжатом.

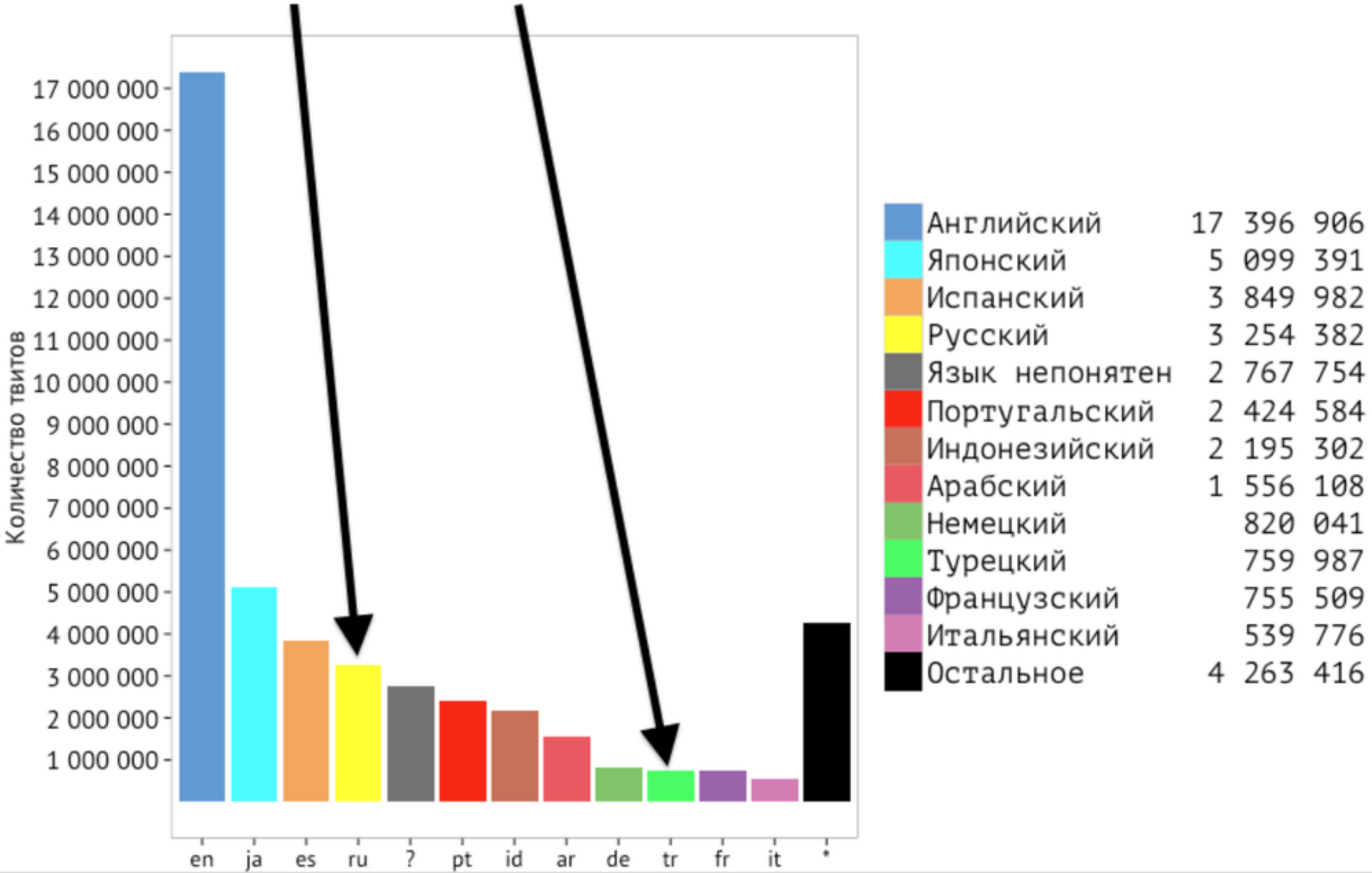
Как оставить только нужные твиты?

Машинно-обученные классификаторы!

	Яндекс	twitter
Обучающая выборка	Запросы пользователей	Твиты
Охват языков	Малая часть	Все возможные
Дата появления	Начало времён	Февраль 2013 года

Твитов со ссылками в день — около **45 млн**.

Нас интересуют только **русские** и **турецкие** твиты (в сумме около **4 млн**).



Неоднозначность

На каком языке написан этот твит?

Фрося
@fr0syfr0

Ура новенькие котяты!!
new.livestream.com/FosterKittenCa... born May 8th
Mama is Kari. The boys are Grant, Adam, and Tory. The girl is Jamie

Все врут: twitter

Caylie
@fopehybopuj

Почему многие геи против однополых браков? dlvr.it/3VnXZh

Anna Zeiman
@AnnaZeiman

OneTwoTrip и Anywayanyday: история конфликта siliconrus.com/2013/06/onetwo

QAZETA.COM
@Vuqar2011

Bu gün TQDK buraxılış imtahanları keçirəcək wp.me/p2lgIU-3JJ

	Я	Человек
uk	ru	ru
bg	ru	ru
tr	az	az

Все врут: Яндекс

Nikolay Kolev
@mazalo

Едно изключение от поп-фолк клишетата fb.me/2uWPYdosn

Mustafa Selcuk
@mustafaselcuk

Twitter'daki 17 yalan haber ensonhaber.com/twitterdaki-ge...

Azgii
@azaa9098

Ё.Отгонбаярын багын найз А.Чинбат хэргээ хүлээжээ medee.mn/main.php?eid=3...

Все врут: методы борьбы

- Переобучать классификатор (дорого!).
- Агрегировать твиты по пользователям.
- Фильтровать по алфавиту.
  - Украинский: **і, ї**
  - Сербский/македонский: **њ**
  - Азербайджанский: **ə**
- Болгарский победить сложнее всего (**ъ**?).

II. Короткие ссылки

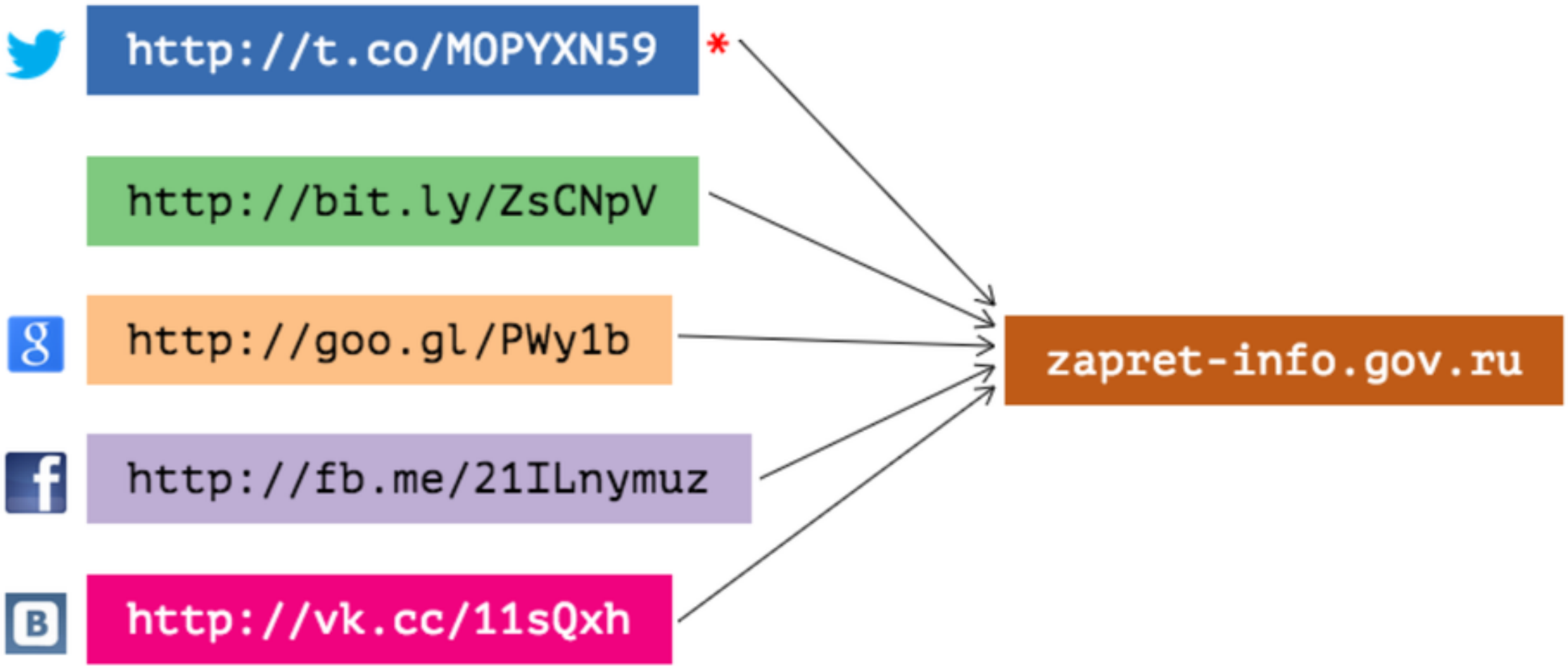
- Значительная часть ссылок в Твиттере испорчена укорачивателями.
- Нельзя просто так взять и понять, что скрывается за ссылкой.

Сергей Рябов
@ryabov_s

Вот и реестр запрещенных сайтов заработал. vk.cc/110pP8

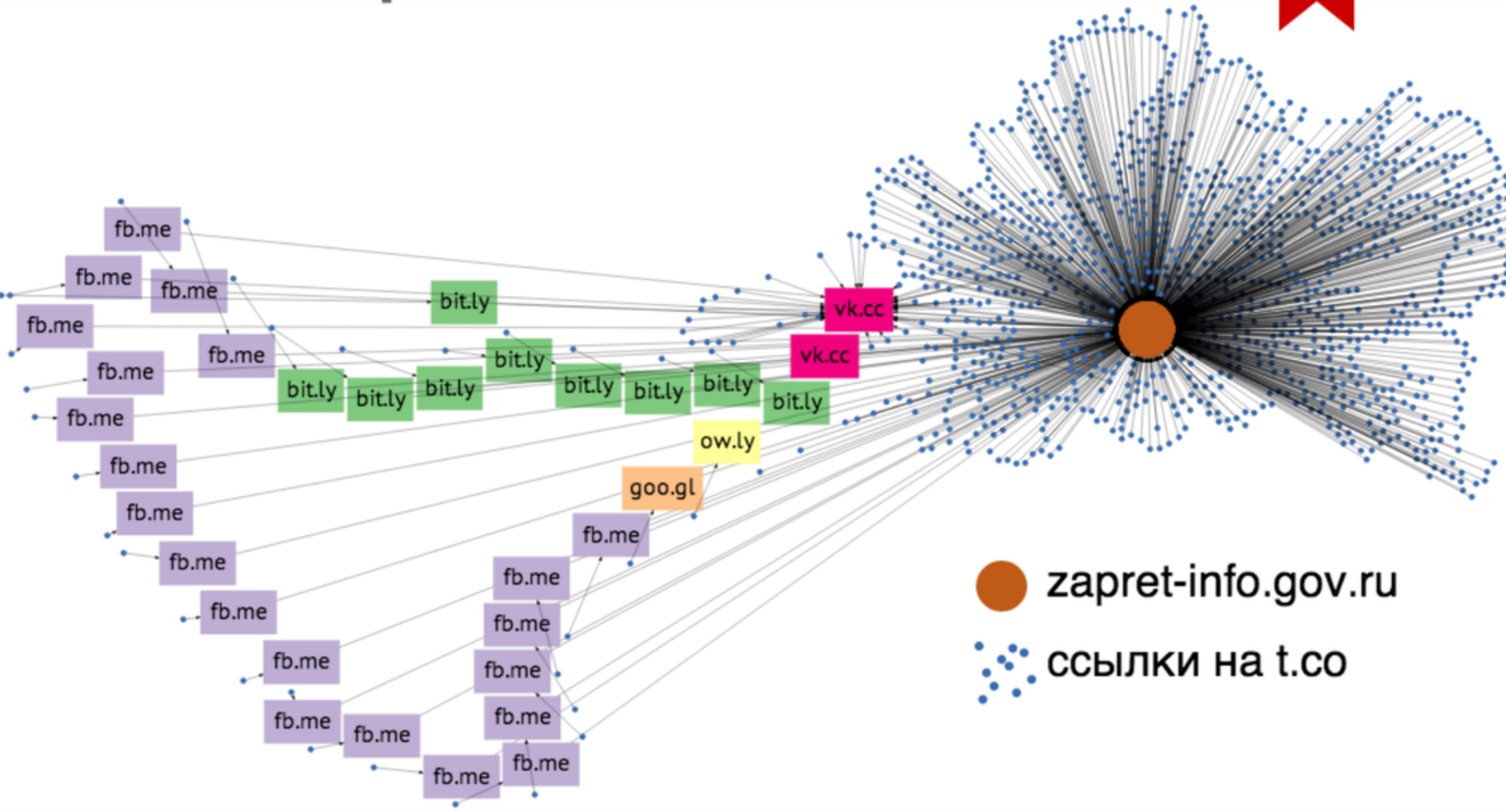
Эта ссылка ведёт на всё тот же `http://zapret-info.gov.ru`.

Популярные укорачиватели



*добровольно-принудительное укорачивание

Полная картина



Хранение редиректов

- Мы сохраняем весь граф редиректов для каждой ссылки.

  Зачем нужен граф?

  - Уменьшает число обращений к укорачивателям.
  - Единая база редиректов для всего Яндекса.

Некоторые любят подлиннее

pagan
@torque10

Will Facial-Recognition Software Finally Reveal Mona Lisa's True Identity? | bit.ly/IMPLnu RT @v_shakthi @ashwinsanghi @AndisKakeli

III. Текстовая релевантность

Релевантен ли твит запросу?

навальный

komagor
@komagor

Алексей Навальный о Координационном Совете - YouTube youtube.com/watch?v=ByyWGR...

ДА ИЛИ НЕТ?!

Однозначные запросы

`джигурда`

Vadim Samartsev
@vdmsmrcv

youtube.com/watch?v=mdThpb... Никита Джигурда избил #православного эксперта в программе "Прямой эфир".

✓

`pussy riot`

Tereza
@TerezaNotMother

Музыкальная посылка от Сявы для Pussy Riot. Шик) youtube.com/watch?v=dIQKWc...

✓

Многословные запросы

`северное братство хомяков`

Русская планета
@rusplt

Лидер сообщества «Северное братство» признал свою вину в экстремизме: Ученый Петр Хомяков, являющийся руководи... bit.ly/PZqkVi

✓

`карта россии`

Vladimir Kaltyrin
@kaltyrin

Карта пунктов приёма вторсырья | Гринпис России act.gp/xGXBHV via @gp_russia

✗

Неправильное выделение объекта

за рулем

Катёнатор 16+
@noomrem

Соц.кампания,цель которой — обратить внимание водителей на проблему употребления алкоголя за рулем.Они там все упрлс. youtube.com/watch?feature=...

- Скорее всего, имелся в виду журнал «За рулём».
- Важно, в каком контексте используются слова запроса в твите.

Текст «не про то»

народный фронт

Ivan Semyonov
@Sem7ser

"@interfax_news: Пятеро боевиков убиты в Дагестане: bit.ly/18A5DLW #novosti" - Они не хотеть вступать в Народный фронт

лужков , матвиенко

LENTA.RU

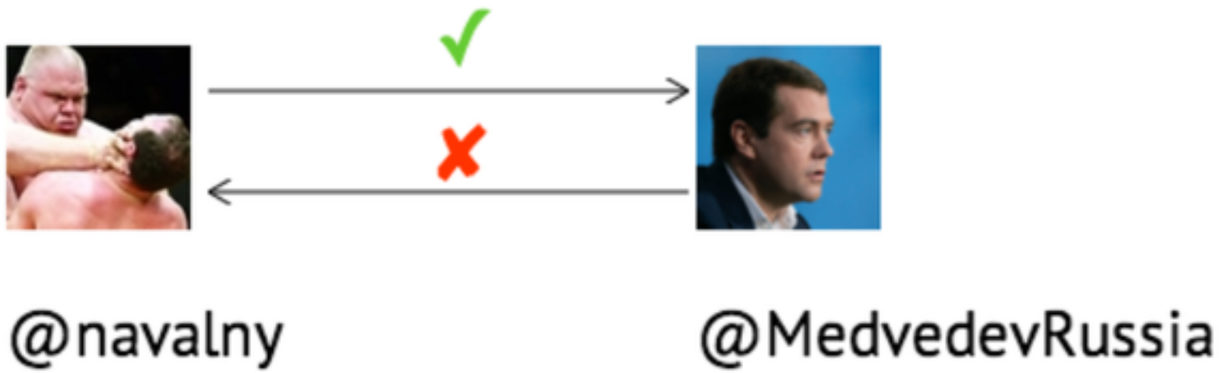
Дорогая редакция
@lentaruofficial

Питер интернетизирован на 59%, Мск - на 58% http://bit.ly/h3kHg7. При Лужкове такого не было, а при Матвиенко эта сторона наиболее опасна

Граф дружбы

Ориентированный граф (**V**, **E**), где:

- **V** — пользователи.
- **E** — пары читающих друг друга пользователей.



IV. Социальный граф

TTX

- Twitter API для сбора графа.
- Размеры (только русские и турки):
  - |V| = **21 млн**.
  - |E| = **1,6 млрд**.
- **MESH** для обработки (см. доклад **Паши Артёмкина**).

Применение: авторитетность

- Все пользователи Твиттера равны, но некоторые равнее других.
- Нужно число, оценивающее ценность пользователя.
- Количество читателей — очень плохо.
- Алгоритмы, использующие граф целиком, гораздо лучше. *

* На самом деле нет.

TunkRank

- Итеративный графовый алгоритм.
- Как PageRank, только для Твиттера.
- Грубая и тупая модель, но работает.

$$TunkRank(user) = \sum_{f \in Followers(user)} \frac{1 + RetweetProbability \cdot TunkRank(f)}{|Following(user))|}$$

V. Спам

Чем занимаются спамеры?

- Автоматический фолловинг.
- Покупка и взлом пользователей.
- Воровство контента.
- Создание псевдосообществ.
- Нет предела совершенству!

Как мы с этим боремся?

- [redacted].
- [redacted].
- [redacted].

  [redacted]: [redacted], [redacted] и [redacted].
- Нет предела совершенству!

Спасибо за внимание!

#TweetLangChallenge

#yasubbotnik

Это не спам!

Почта: dfyz@yandex-team.ru

Твиттер: @i_komarov

Реальная возможность **заработать** призы!