Яндекс

Роботы и люди в Twitter'e

Дмитрий Кузнецов
разработчик

Я.Субботник, Екатеринбург, 2 июля 2011 года

Зачем мы сегодня здесь?

Яндекс

блоги

☑ только в записях из микроблогов, в сообщениях автора «yandex»…

Показаны сообщения **1** — **10** из **769** найденных.

http://twitter.com/yandex/statuses/83811791846637568

yandex: Рамблер теперь использует поиск Яндекса http://t.co/XyrT1EX

1 ч. 23 мин. назад · yandex · twitter.com/yandex · это: **спам**

http://twitter.com/yandex/statuses/83504711579348992

yandex: RT @ddnv: Внутренний график разработки Яндекс.Почты для iPhone. http://t.co/0vODdao

вчера, 16:00 · yandex · twitter.com/yandex · это: **спам**

http://twitter.com/yandex/statuses/83450856137166848

yandex: Встречайте приложение Яндекс.Почты для iPhone: с push-уведомлениями, группировкой писем и отличным д
http://t.co/yt12TtV

вчера, 12:26 · yandex · twitter.com/yandex · это: **спам**

http://twitter.com/yandex/statuses/83175589179162624

yandex: Firefox 5 только что вышел. Ещё быстрее и лучше. Загружайте версию от Яндекса на http://t.co/aaROM3l

21 июня 2011, 18:13 · yandex · twitter.com/yandex · это: **спам**

http://twitter.com/yandex/statuses/81698101953441792

yandex: Погодная тема в Яндекс.Почте RT @jogur_t: Мало кто знает, что в ней 4 сезона и больше 160 экранов погоды

17 июня 2011, 16:22 · yandex · twitter.com/yandex · это: **спам**

Информационный мусор

Twitter

Twitter растёт, и мы любим эксперименты!

Русский Twitter

Русский Twitter

Сегодня

700 тыс + пользователей

400 тыс + твитов каждый день

2010 год

300 тыс +

200 тыс +

По данным Яндекс.Поиска по блогам

Русский Twitter

В России

700 тыс + пользователей

400 тыс + твитов каждый день

В Мире

200 млн +

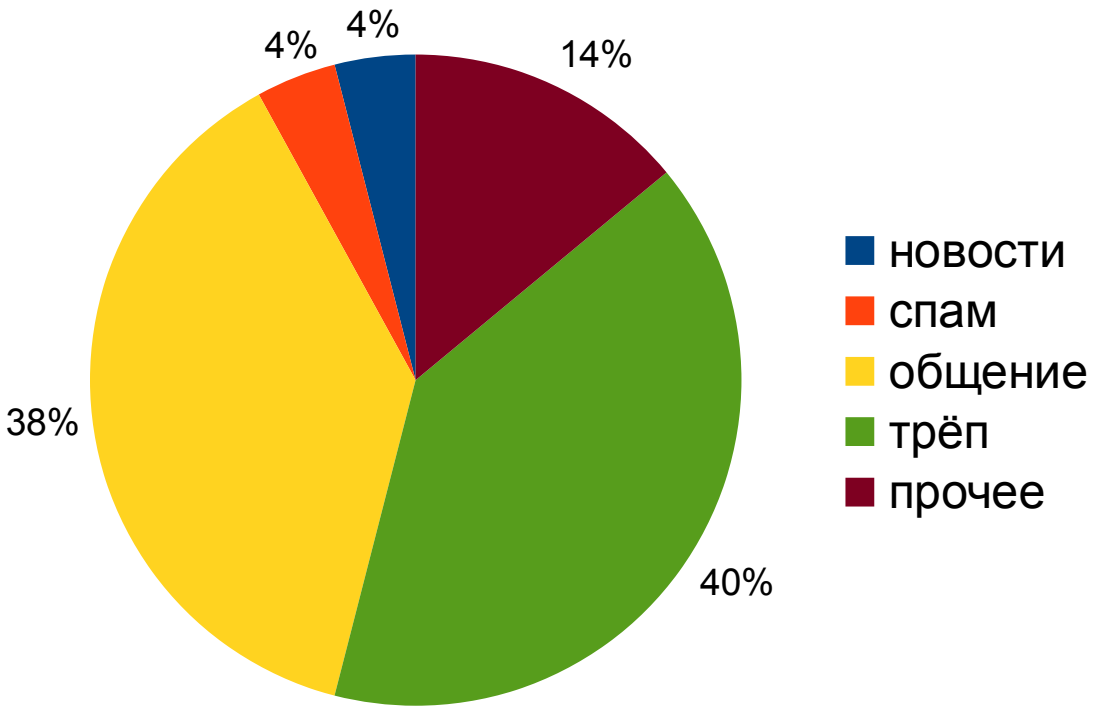
155 млн +

5 % пользователей пишут 75% всех твитов

По данным Яндекс.Поиска по блогам и http://business.twitter.com

Twitter : контент

Twitter : контент
Кто о чём пишет



*PearAnalytics

Русский Twitter

русскоговорящие пользователи Twitter'а — почти замкнутая система

Twitter : контент

каждый третий твит содержит ссылку

http://clck.ru/5dzj

1. рассказать другим о чём-то интересном

2. самопродвижение и реклама

Twitter : кто пишет?

Twitter : роботы

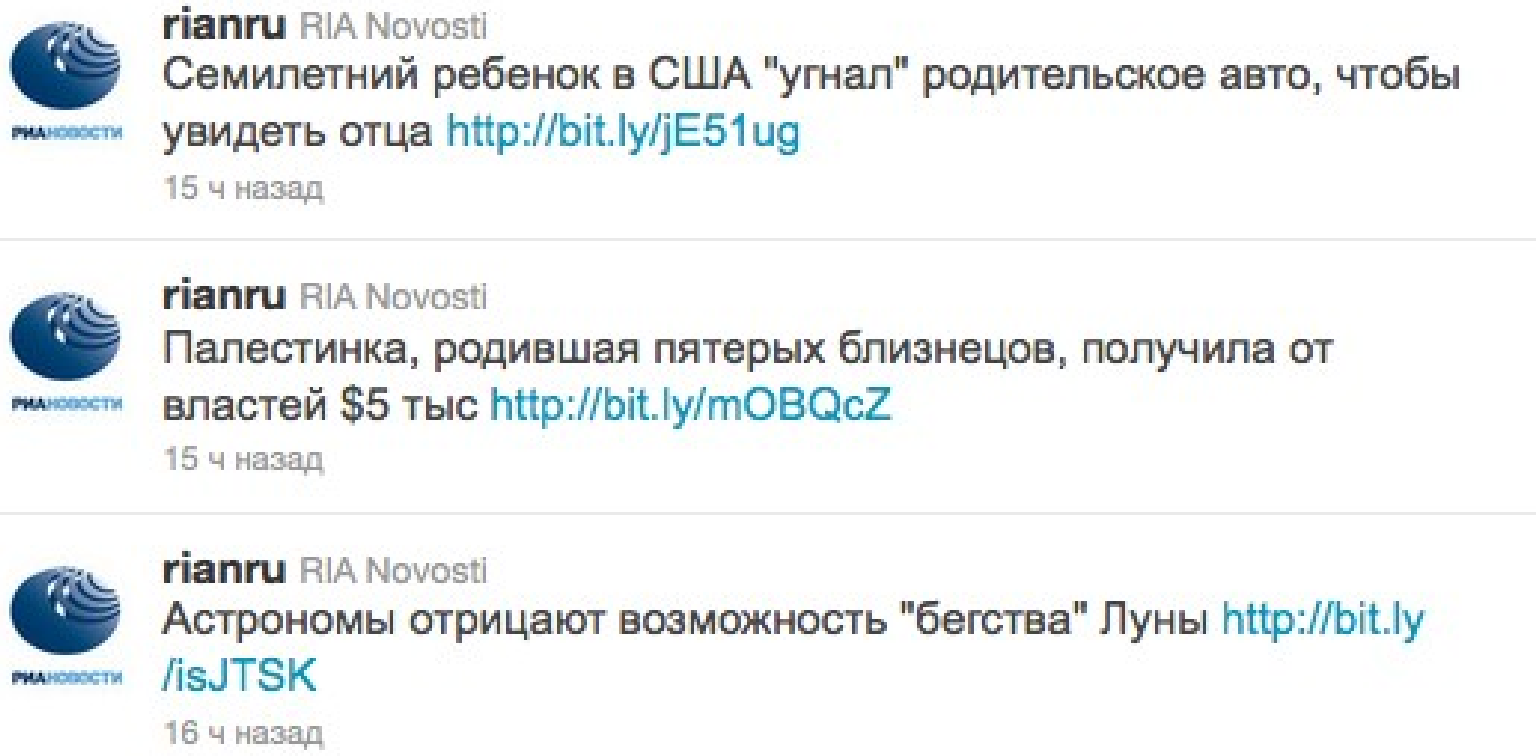
Я

Twitter : роботы

контент сгенерирован автоматически или «редакцией»

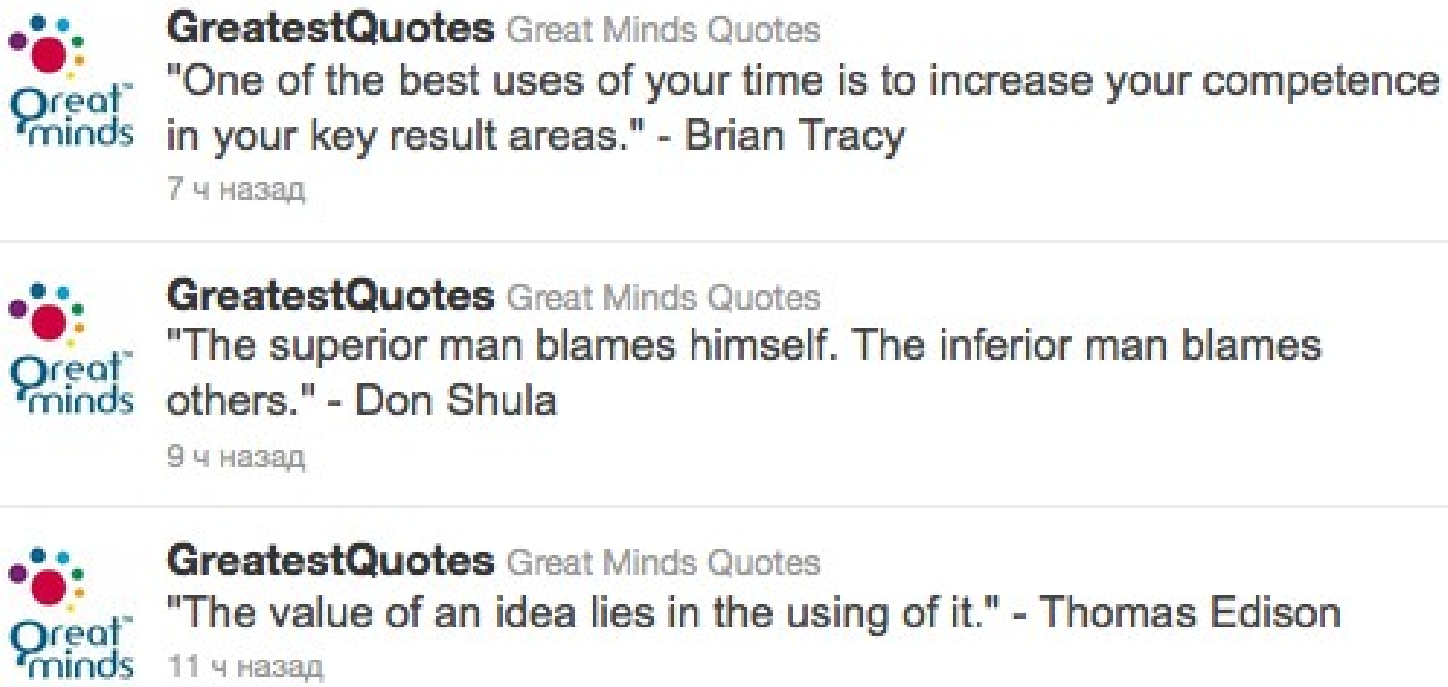
Twitter : роботы

трансляции с сайта или блога



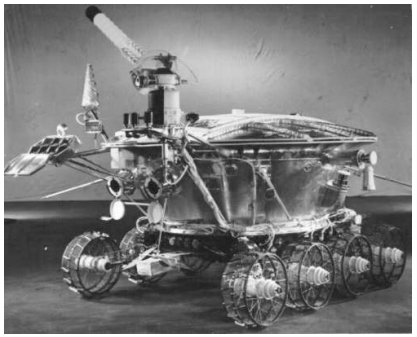
Twitter : роботы

ссылки — не всегда основное содержимое твитов

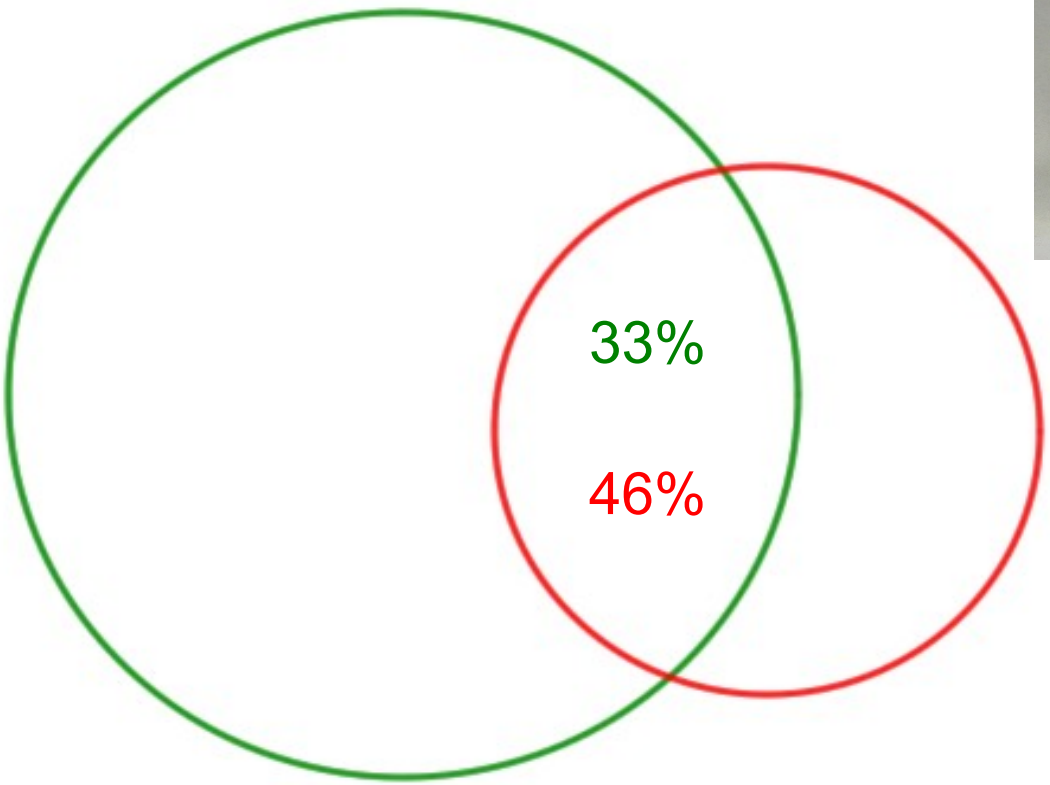


Twitter : роботы

спам



роботы



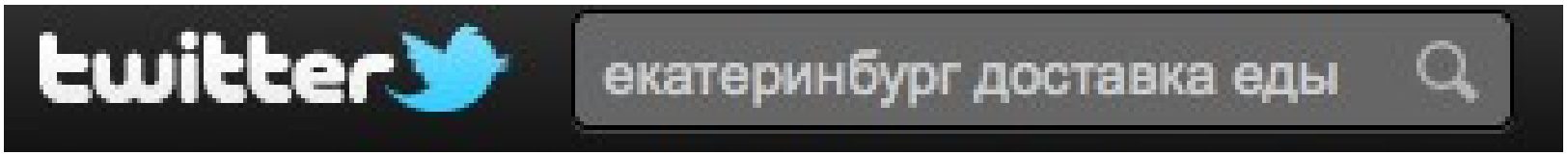
Twitter : роботы
Сколько?

- 10 % пользователей являются роботами

- 25 % всех твитов произведены на свет роботами

Twitter : роботы

Теперь роботы тоже умеют искать



Twitter : роботы

сегодня аккаунт есть не только у президентов

feral_pigeon feral_pigeon
coo coo coo ...
22 июня

feral_pigeon feral_pigeon
peck.
21 июня

feral_pigeon feral_pigeon
fast walk bob bob bob bob bob bob bob bob bob bob
21 июня

Twitter : роботы

сегодня аккаунт есть не только у президентов

tweetingseat TweetingSeat

Tweeting from @smallsocietylab 'Pop-up Science Museum' in the @dcadundee http://twitpic.com/5ctev3

17 июня

tweetingseat TweetingSeat

Tweeting from @smallsocietylab 'Pop-up Science Museum' in the @dcadundee http://twitpic.com/5ctett

17 июня

tweetingseat TweetingSeat

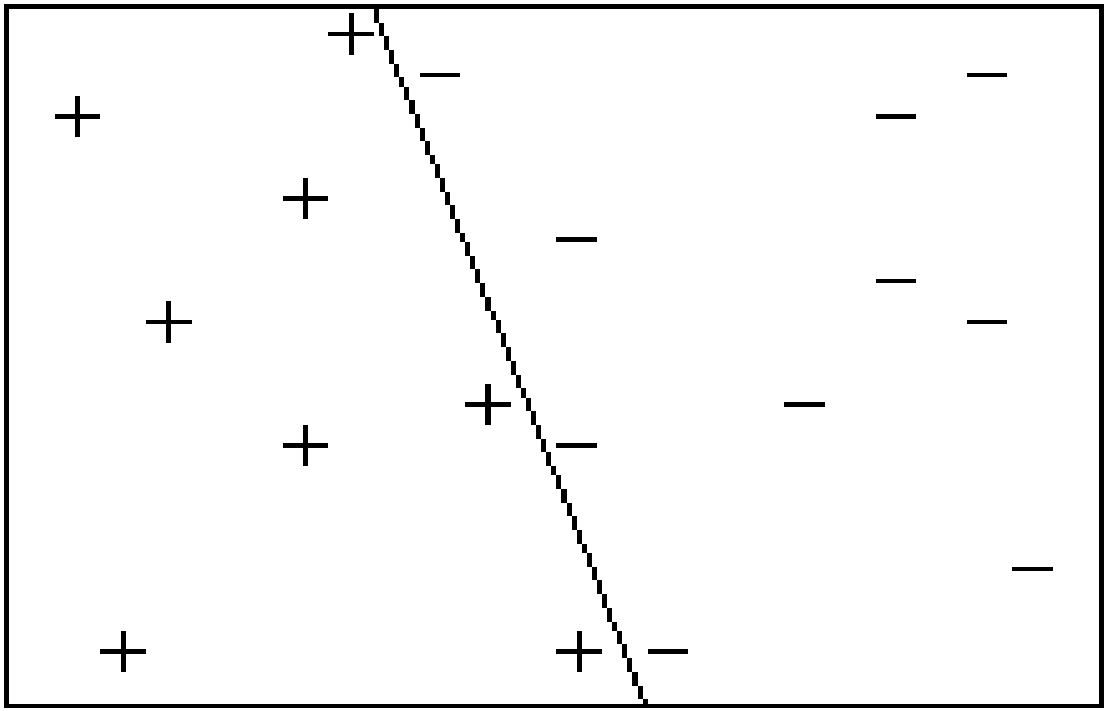
Tweeting from @smallsocietylab 'Pop-up Science Museum' in the @dcadundee http://twitpic.com/5cterd

17 июня

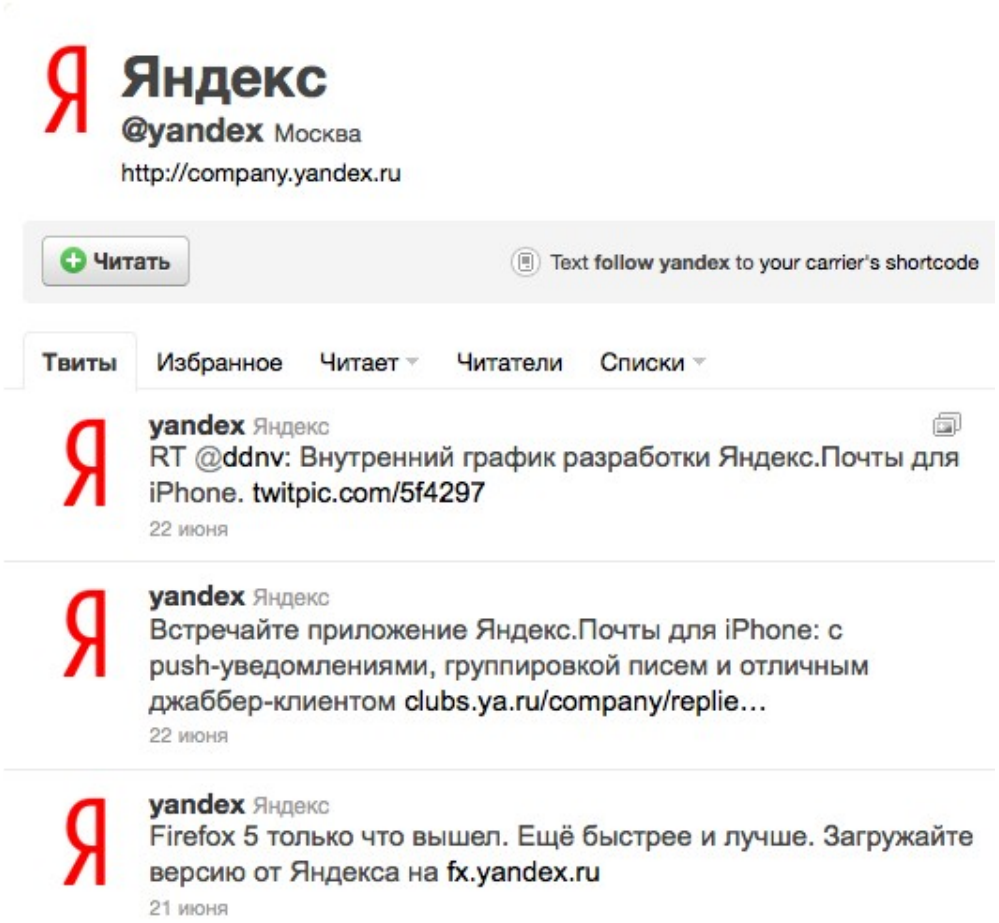
Мы научились их отличать!

Как?

Задача классификации — машинное обучение



Классификация



факторы из содержимого твитов

Пример

каждый твит содержит ссылку, и пользователь ни с кем не разговаривает

много смайликов и ответов другим пользователям

Пример

каждый твит содержит ссылку, и пользователь ни с кем не разговаривает

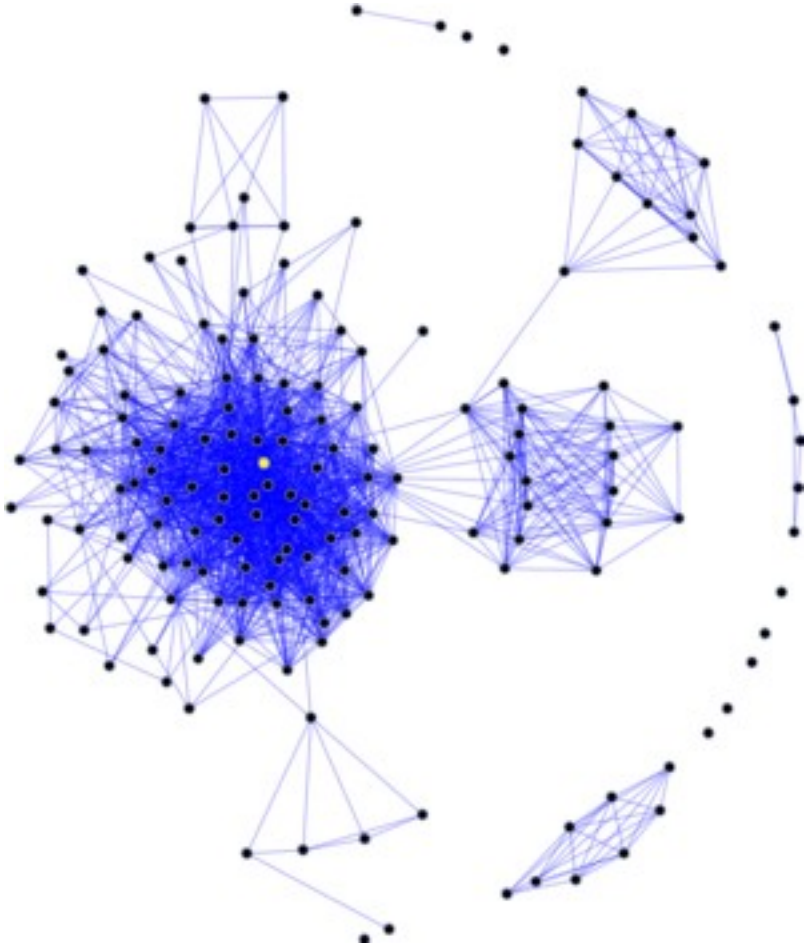
доля ссылок на пост

стилистические факторы

много смайликов и ответов другим пользователям

разговор с «людьми»

Классификация



факторы из социального графа

Итого

извлечение факторов

+

суровая математика

= классификация

Повторяйте это дома!

Повторяйте это дома!

Twitter API: http://dev.twitter.com/doc

RapidMiner: http://rapid-i.com

Weka: http://www.cs.waikato.ac.nz/ml/weka

SVM-Light: http://svmlight.joachims.org

Дмитрий Кузнецов
Разработчик

drsmith@yandex-team.ru