

# 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/aaROM3I>

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 +

# Русский Twitter

## В России

700 тыс +  
пользователей

400 тыс + ТВИТОВ  
каждый день

## В Мире

200 млн +

155 млн +

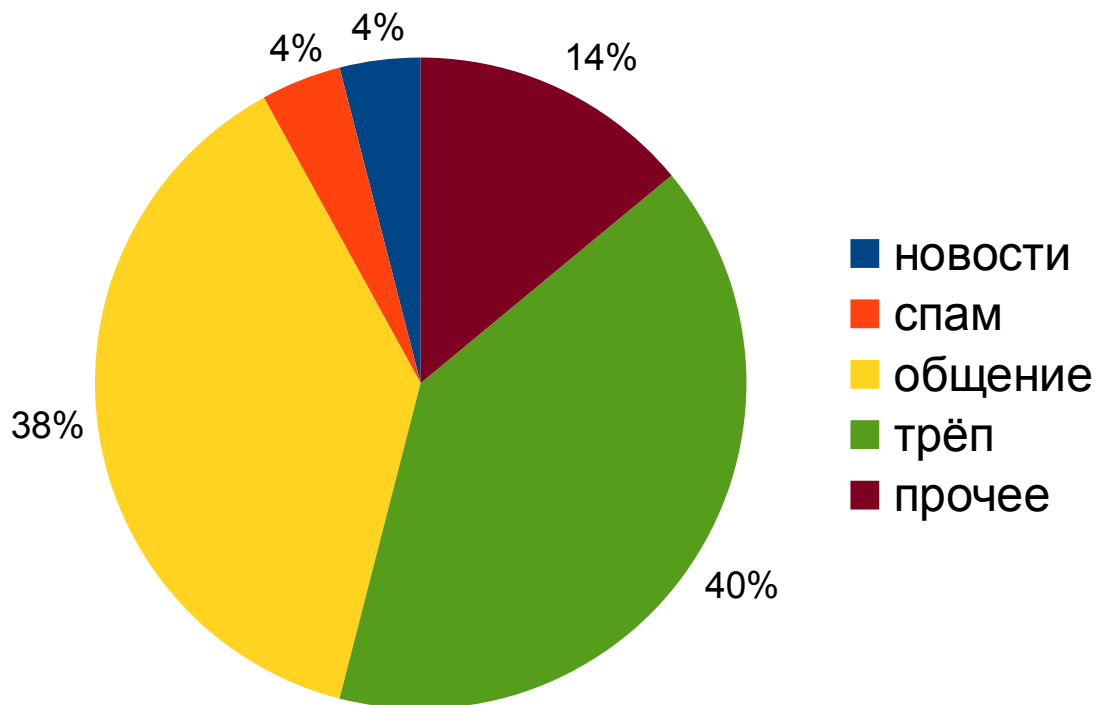
5 % пользователей пишут 75% всех ТВИТОВ



# Twitter : КОНТЕНТ

# Twitter : КОНТЕНТ

## Кто о чём пишет



# Русский Twitter

русскоговорящие пользователи Twitter'a — 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



# Twitter : роботы

контент сгенерирован автоматически или  
«редакцией»



# Twitter : роботы

трансляции с сайта или блога



**rianru** RIA Novosti

Семилетний ребенок в США "угнал" родительское авто, чтобы увидеть отца <http://bit.ly/jE51ug>

15 ч назад

---



**rianru** RIA Novosti

Палестинка, родившая пятерых близнецов, получила от властей \$5 тыс <http://bit.ly/mOBQcZ>

15 ч назад

---



**rianru** RIA Novosti

Астрономы отрицают возможность "бегства" Луны <http://bit.ly/isJTSK>

16 ч назад



# Twitter : роботы

ССЫЛКИ — НЕ ВСЕГДА ОСНОВНОЕ СОДЕРЖИМОЕ ТВИТОВ



**GreatestQuotes** Great Minds Quotes

"One of the best uses of your time is to increase your competence in your key result areas." - Brian Tracy

7 ч назад

---



**GreatestQuotes** Great Minds Quotes

"The superior man blames himself. The inferior man blames others." - Don Shula

9 ч назад

---

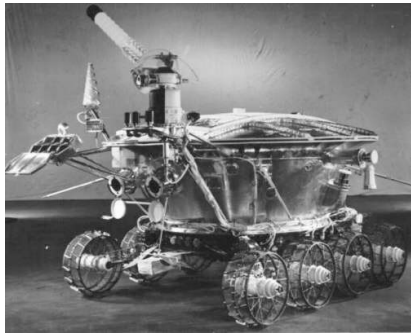


**GreatestQuotes** Great Minds Quotes

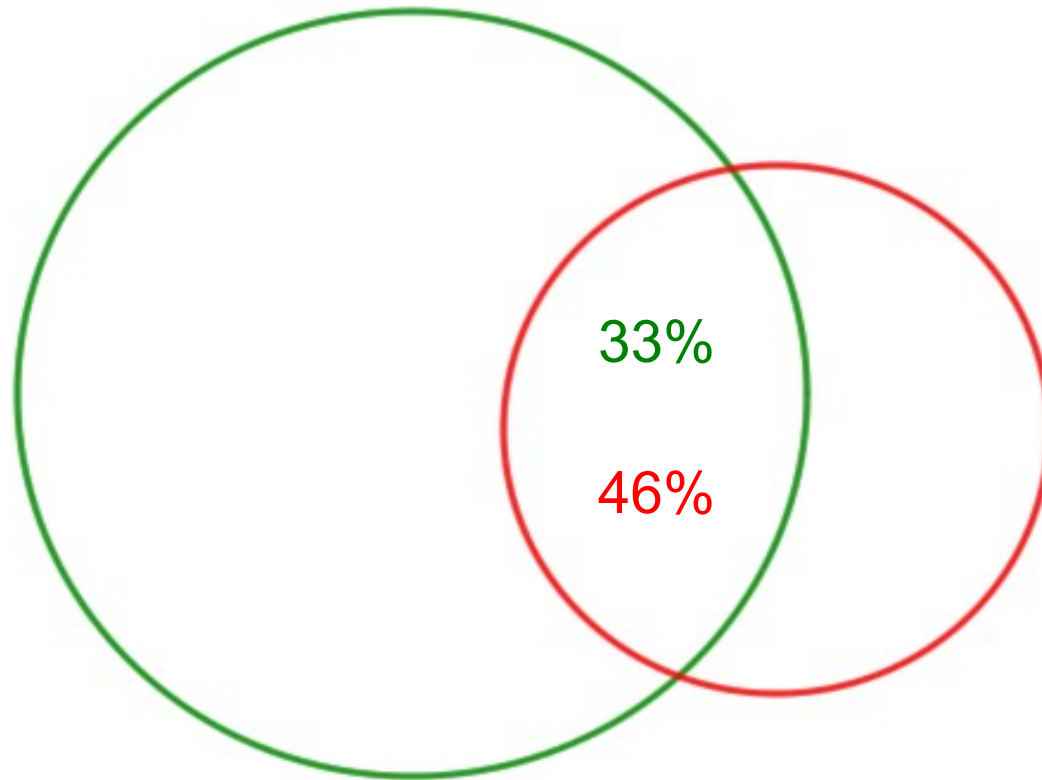
"The value of an idea lies in the using of it." - Thomas Edison

11 ч назад

# Twitter : роботы



роботы



спам



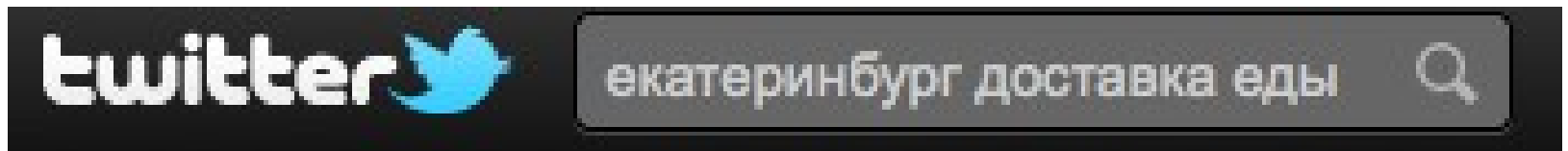
# Twitter : роботы

## Сколько?

- 10 % пользователей являются роботами
- 25 % всех твитов произведены на свет роботами

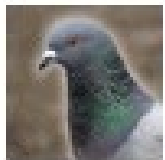
# Twitter : роботы

Теперь роботы тоже **умеют искать**



# Twitter : роботы

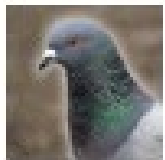
сегодня аккаунт есть не только у президентов



**feral\_pigeon** feral\_pigeon  
соо соо соо ...

22 ИЮНЯ

---



**feral\_pigeon** feral\_pigeon  
реск.

21 ИЮНЯ

---



**feral\_pigeon** feral\_pigeon  
fast walk 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



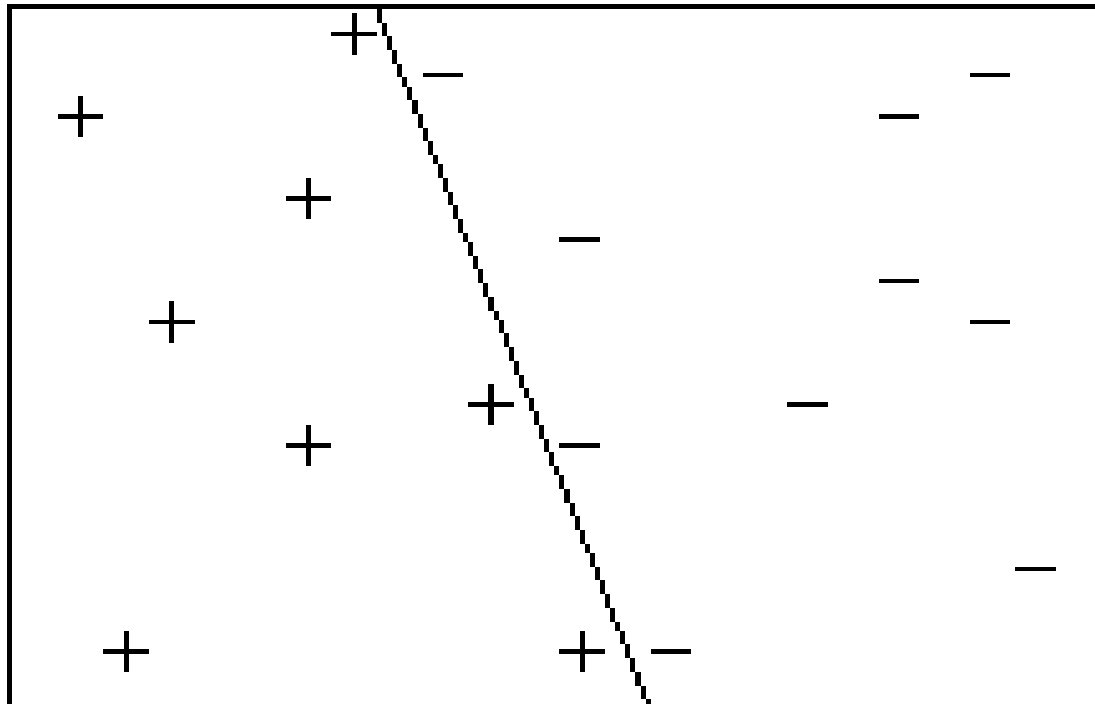
Tweeting from [@smallsocietylab](#) 'Pop-up Science Museum' in the  
[@dcadundee](#) <http://twitpic.com/5cterd>

17 июня

**Мы научились их  
отличать!**

# Как?

Задача классификации — машинное обучение






# Классификация

факторы из  
содержимого ТВИТОВ

**Я Яндекс**  
@yandex Москва  
<http://company.yandex.ru>

[+ Читать](#) [Text follow yandex to your carrier's shortcode](#)

**Твиты** | Избранное | Читает ▾ | Читатели | Списки ▾

**Я yandex** Яндекс   
RT @ddnv: Внутренний график разработки Яндекс.Почты для iPhone. [twitpic.com/5f4297](http://twitpic.com/5f4297)  
22 июня

**Я yandex** Яндекс  
Встречайте приложение Яндекс.Почты для iPhone: с push-уведомлениями, группировкой писем и отличным джаббер-клиентом [clubs.ya.ru/company/replie...](http://clubs.ya.ru/company/replie...)  
22 июня

**Я yandex** Яндекс  
Firefox 5 только что вышел. Ещё быстрее и лучше. Загружайте версию от Яндекса на [fx.yandex.ru](http://fx.yandex.ru)  
21 июня

# 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

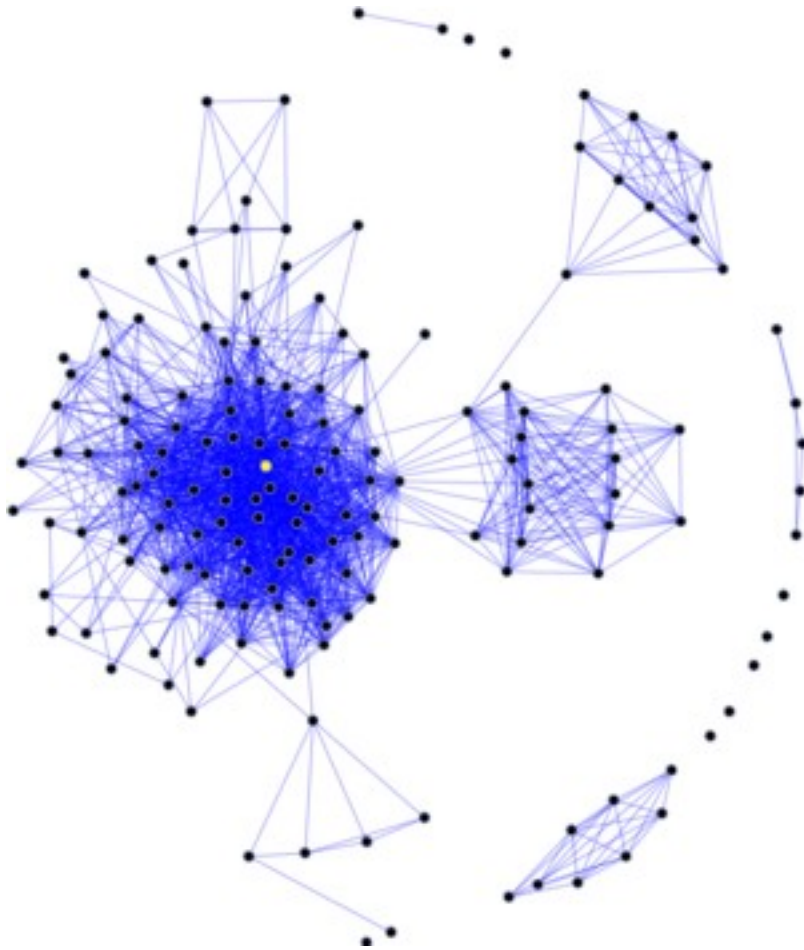
*доля ссылок на пост*

*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](mailto:drsmith@yandex-team.ru)