

Использование алгоритмов обработки естественного языка и методов машинного обучения для прогнозирования популярности новостей

Владимир Зайцев

группа 14806, Кафедра КМИТ
Югорский Государственный Университет
zaytsev@usc.edu

научный руководитель: Владимир Бурлуцкий

27 Июня, 2012

Предметная область

- Новостные сайты – публикация контента и получение прибыли за счет просмотров рекламы.
- Владельцы сайтов стремятся максимизировать свою прибыль через увеличение числа просмотров новостей.

Цель: отбирать новости с наибольшим потенциалом (количество кликов, просмотров, комментариев, реакций в социальных сервисах).

Исходные данные: новостные сайты с информацией о просмотрах новостей, социальные сервисы (Twitter, Facebook, VK и другие)

Инструменты:

- Машинное обучение
- Коллаборативная фильтрация

Information for the World's | www.forbes.com


Log in | Sign up | Connect | Facebook | Twitter | LinkedIn | YouTube | Help

Search news, business leaders, and stock quotes

U.S. Europe Asia Follow Us

Subscribe >

We Need More Immigrants, But Let's Be Smart About It



Perhaps it would be better if our policies were less about politics and more about gaining specific skills and abilities from other countries. [Continue »](#)

Joel Kotkin

Is This The Week The IPO Window Creaks Open?

Steve Schaeffer

Why The RIM Acquisition Rumors Don't Ring True

Eric Savitz

IRS Shakes Up Its Own

Im Kelly

Why The CEO's Lieutenant May Be The Secret To Success

Alice G. Walton

How American Business Is Reaching Out To The Jobless

J. Maureen Henderson

The Best Gift You Can Give Your Employees

Ryan Scott

BUSINESS

The New Case for Women on Corporate Boards: New Perspectives, Increased Profits

Kate Taylor

How to Give Four Generations Feedback

Most Read on Forbes

1. **89 Business Cliches That Will Get Any MBA Promoted And Make Them Totally Useless**

Новость с наибольшим потенциалом просмотров

Имеется множество объектов X (новостей), множество допустимых ответов Y (значений "популярности"), существует также целевая функция $y^* : X \rightarrow Y$, значения которой $y_i = y^*(x_i)$ известны только на конечном подмножестве объектов $X^l = \{x_i, \dots, x_l\}$, $X^l \subset X$.

Задача восстановления регрессии по эмпирическим данным:

По выборке X^l восстановить зависимость y^* , то есть построить *решающую функцию* $a : X \rightarrow Y$ ($Y \subset \mathbb{R}$), которая приближала бы целевую функцию y^* причём не только на объектах обучающей выборки X^l , но и на всём множестве X .

Представление объектов: признаковая модель

Каждая новость это – набор признаков: заголовок, содержание, автор, дата публикации, контекст и т.д.

Для представления текстовых признаков как правило используется **векторная модель** из информационного поиска:

$$TF_t = \frac{n_t}{\sum_k n_k}, \text{ – частота термина в документе } d \quad (1)$$

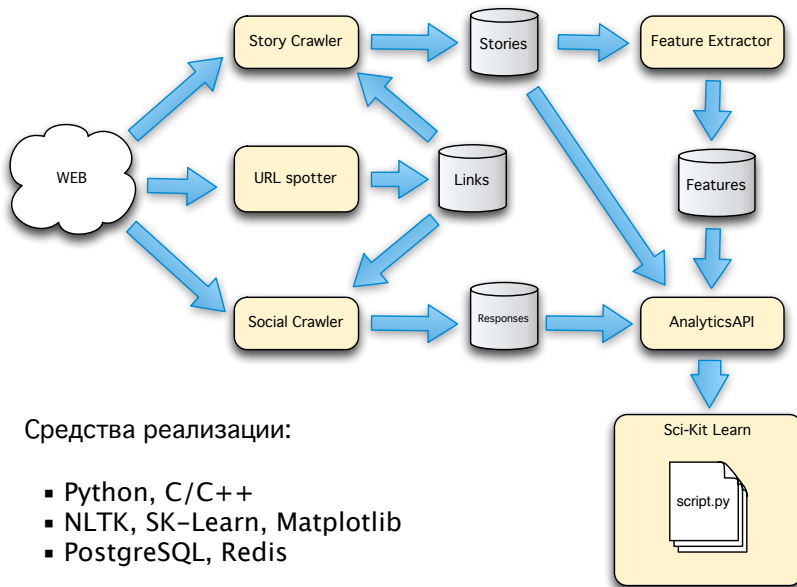
$$DF_t = \frac{|d_i \supset t|}{|D|}, \text{ – частота документов содержащих данный терм} \quad (2)$$

$$TFIDF_{t,d} = TF_{t,d} \times \log \frac{1}{DF_t} \text{ – мера "важности" термина} \quad (3)$$

Разработать систему, позволяющую автоматизировать:

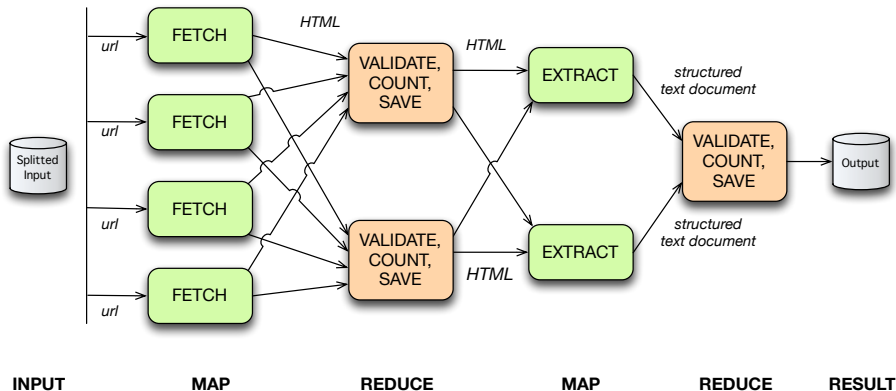
- сбор данных из новостных источников и социальных сервисов;
- преобразование исходных данных в вектора признаков;
- настройку и тестирование регрессионных моделей.

Обзор архитектуры



Структура компонентов: приложения Map-Reduce

Пример: сборщик новостей (Story Crawler)



service: River Fetcher

Common Parameters

ID	1001	Done
Status	SLEEP	0%
Total Tasks	0	
Complete Tasks	0	
Errors Occurred	0	
Tasks Dropped		0
Tasks Cached		0

Worker pool size: [Change](#)

[Run](#)[Stop](#)[Kill](#)

Интерфейс: просмотр наборов данных (2/3)

COLLECTOR SERVICES

[River Fetcher](#)

[Link Spotter](#)

[Page Fetcher](#)

[Page Parser](#)

[SM Probe Fetcher](#)

System Monitor

BUNDLES

[Look At Me](#)

[Forbes Blogs / All](#)

DATASETS

[EN.b.insider.10.11](#)

[EN.Forbes](#)

[RU.look.at.me](#)

dataset: EN.Forbes

Raw Rivers

Mime types	text/html	58799
Sources	Forbes Blogs / All	
Collecting Period	23:54 Mar 27 – 01:54 Mar 28	
See data	Get random example	

Extracted Urls

Sources		
Percentage of Downloaded	%	
See data	Get random examples	

Raw Documents

See data:	Get random example	53409
-----------	------------------------------------	-------

Structured Documents

Publish Dates	2005(3), 2007(10), 2008(9), 2009(355),
---------------	--

Интерфейс: просмотр исходных объектов (3/3)

Collector | Application Index

localhost:8000/apps/collector/dataset/15#

Reader

Farseer

analyti

Raw River

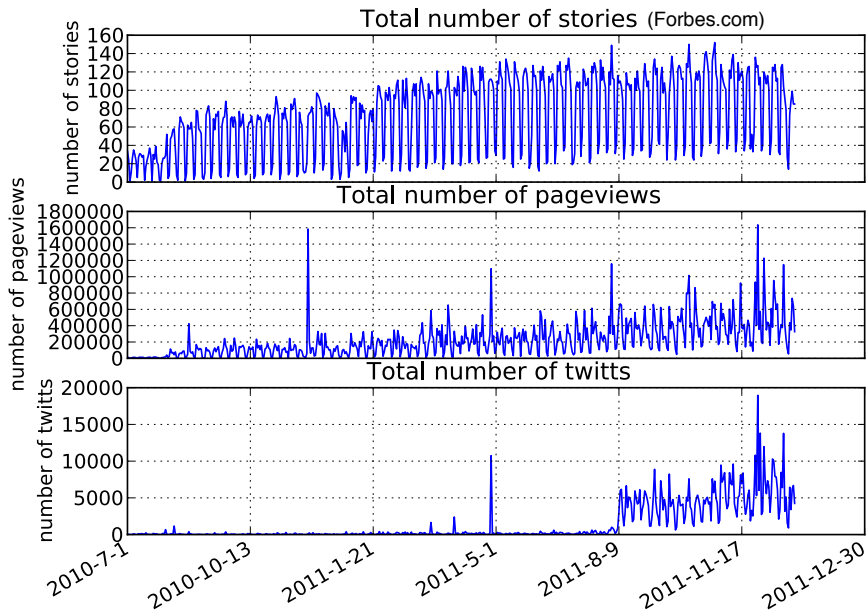
78923

```
pp/">Alex Knapp</a></cite></span><div class="desc">Forbes Staff</div>
<div class="date" rel="Sun Sep 18 19:17:03 EDT 2011">Sep 18, 2011</div>
</div>
<article>
<hgroup>
<h2 class="editable editable-hed"><a href='http://www.forbes.
com/sites/alexknapp/2011/09/18/a-flying-robot-plays-catch-and-holds-the-key-for-efficient-hvac-systems/'>A Flying Robot Plays
Catch and Holds the Key For Efficient HVAC Systems</a></h2>
</hgroup>
<a href='http://www.forbes.co
m/sites/alexknapp/2011/09/18/a-flying-robot-plays-catch-and-holds-the-key-for-efficient-hvac-systems/' class="editable editab
le-img"><img src='http://blogs-images.forbes.com/thumbnails/blog_1401/pt_1401_3409_o.jpg?t=1316388186' /></a>
<p>Researchers at the U.C. Berkeley EECS prog
ram have developed a flying robot that's capable of learning how to catch a ball. That's not completely new - I blogged about
a set of researchers at the Flying Machine Arena who developed a similar robot a few months ago. But what is interesting abo
ut this robot is that the means through which its learning, which Dr. Anil Aswani describes as "Learning-based model predicti
ve control" or LBMP for short. <a href='http://www.forbes.com/sites/alexknapp/2011/09/18/a-flying-robot-plays-catch-and-hold
s-the-key-for-efficient-hvac-systems/'>read &raquo;</a></p>
</article>
</li>
```

Close

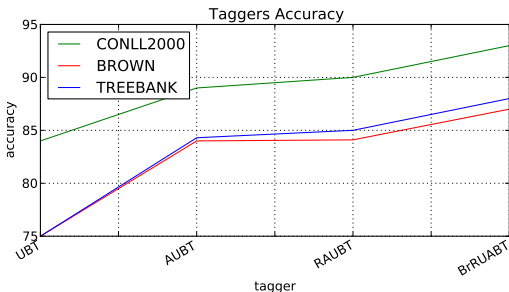
Forbes Blogs / All

Набор данных: 220,000 документов / 40 Gb HTML



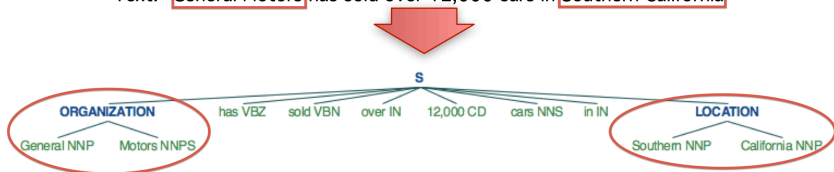
Извлечение признаков из текста: NLTK

- Сегментация
- Определение частей речи (Part-Of-Speech Tagging):



- Распознавание именованных сущностей (NER-Chuncking):

Text: "General Motors has sold over 12,000 cars in Southern California"



- Отбор термов на основе их частот ($600,000 \Rightarrow 22,000$).
- Удаление пунктуации ($22,000 \Rightarrow 20,000$)
- Метод LASSO ($20,000 \Rightarrow 177$):

$$\hat{Y} = \hat{\beta}_0 + \sum_{j=1}^p x_j \hat{\beta}_j; \quad \hat{Y} = X^T \hat{\beta} \quad (4)$$

$$\hat{\beta}^{lasso} = \arg \min_{\hat{\beta}} \left\{ \frac{1}{2} \sum_{i=1}^N (y_i - \beta_0 - \sum_{j=1}^p x_{ij} \beta_j)^2 + \lambda \sum_{j=1}^p |\beta_j| \right\} \quad (5)$$

Незначимые с точки зрения алгоритма признаки получают нулевые коэффициенты и удаляются из модели.

Алгоритмы:

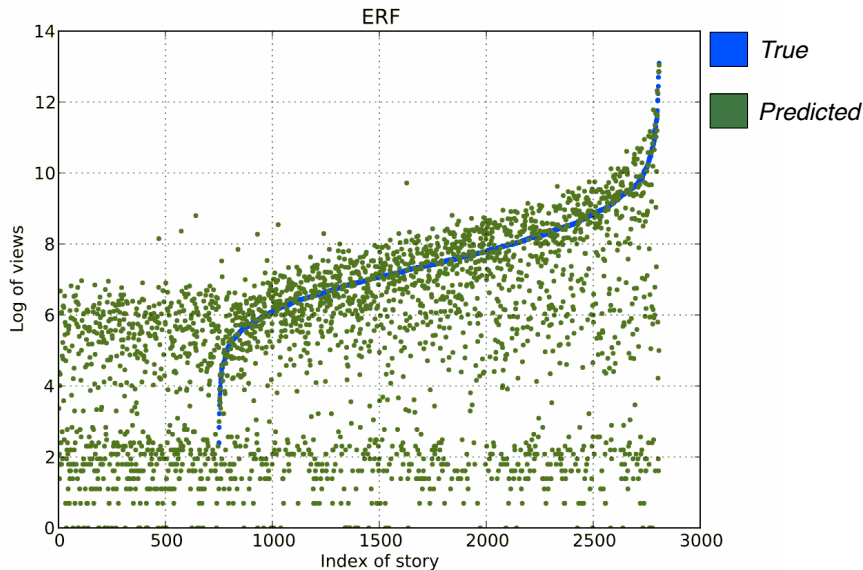
- Метод k -ближайших (KNN);
- Ридж-регрессия (Ridge-regression, L2)
- Случайный лес (Extremely Randomized Forest).

Настройка: минимизация эмпирического риска (ERM) методом скользящего контроля (K-Fold Cross Validation).

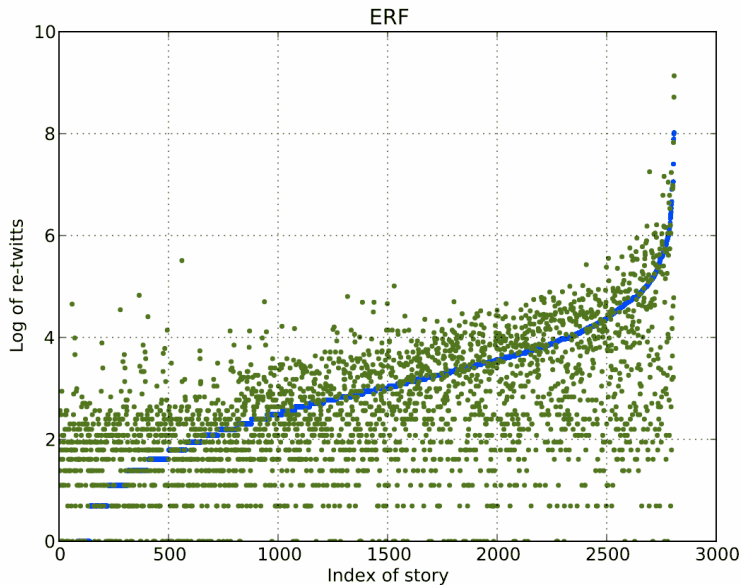
Функционал качества:

$$Q(a, X^l) = \frac{1}{l} \sum_{i=1}^l (a(x_i) - y^*(x_i))^2 \quad (6)$$

Результаты (просмотры): ERF



Результаты (реакции в Twitter): ERF



Реализована система, обеспечивающая автоматизацию процесса прогнозирования популярности интернет-новостей.

Система включает в себя инструменты, необходимые для сбора исходных данных, преобразования текстов в векторную модель с использованием алгоритмов обработки естественных языков.

Собираемые данные предоставляются потребителю в удобном структурированном виде, совместимом с пакетом *scikit-learn*, и могут использоваться для прогнозирования.

Вопросы?

Владимир Зайцев
zaytsev@usc.edu