

# Определение наилучшего ответа на StackOverflow

Никита Подгузов

Научный руководитель: Рауф Курбанов

Санкт-Петербургский Академический университет

18 июня 2018 года

Возможности сервисов вопросов и ответов:

- Задавать вопрос (и отмечать правильный ответ)
- Отвечать на вопросы, заданные другими пользователями
- Голосовать за понравившиеся ответы


YAHOO! ANSWERS

Quora

 stackoverflow

### Особенности системы:

- Узкоспециализированная
- Большая база вопросов

 serverfault

Questions Tags Users Badges Unanswered Ask Question

### How do you search for backdoors from the previous IT person?

▲ We all know it happens. A bitter old IT guy leaves a [backdoor](#) into the system and network in order to have fun with the new guys and show the company how bad things are without him.

▼ I've never personally experienced this. The most I've experienced is somebody who broke and stole stuff right before leaving. I'm sure this happens, though.

★ So, when taking over a network that can't quite be trusted, what steps should be taken to ensure everything is safe and secure?

174

security

edited May 26 '11 at 16:02

asked Aug 18 '10 at 15:04

Peter Mortensen

Jason Berg

2,087 • 4 • 20 • 24

16k • 6 • 31 • 54

share improve this question

21 Answers

active oldest votes

▲ It's really, really, really hard. It requires a very complete audit. If you're very sure the old person left something behind that'll go boom, or require their re-hire because they're the only one who can put a fire out, then it's time to assume you've been rooted by a hostile party. Treat it like a group of hackers came in and stole stuff, and you have to clean up after their mess. Because that's what it is.

▼ **325**

✓

- Audit every account on every system to ensure it is associated with a specific entity.
- Accounts that seem associated to systems but no one can account for are to be mistrusted.
- Accounts that aren't associated with anything need to be purged (this needs to be done anyway, but it is especially important in this case)

edited Apr 13 '17 at 12:14

answered Aug 18 '10 at 15:40

Community

sysadmin1138

1

111k • 16 • 136 • 273

▲ I would say it is a balance of how much concern you have vs the money you are willing to pay.

▼ **97**

**Very concerned:**

If you are very concerned then you may want to hire an outside security consultant to do a complete scan of everything from both an outside and internal perspective. If this person was particularly smart you could be in trouble, they might have something that will be dormant for a while. The other option is to simply rebuild everything. This may sound very excessive but you will learn the environment well and you make a disaster recovery project as well.

edited Aug 18 '10 at 15:33

answered Aug 18 '10 at 15:18

Kyle Brandt

63.3k • 59 • 248 • 403

# Введение

## Постановка задачи

Проблемы:

- Большая доля "неразрешенных" вопросов
- Нет возможности помочь оценить правильность ответов пользователю, задавшему новый вопрос

Хотим научиться определять правильные ответы, используя базу вопросов ServerFault

# Актуальность

Google & StackOverflow

Google ruby capitalize

All Shopping News Images Videos More Settings Tools

About 369.000 results (0,34 seconds)

**Capitalize first letter in ruby - Stack Overflow**  
<https://stackoverflow.com/questions/3724913/capitalize-first-letter-in-ruby> ▼  
Sep 16, 2010 - Unfortunately, it is impossible for a machine to upcase/downcase/capitalize properly. ... That's why Ruby's String class only supports capitalization for ASCII characters, because there it's at least somewhat well-defined.

How to convert a string to lower or upper case in **Ruby** 21 Dec 2014  
**ruby** - Using 'capitalize' or 'capitalize!' 29 Sep 2014  
**Capitalize the first letter of each word - Ruby** 26 Jan 2014  
**Capitalize only first character of string and leave others alone ...** 12 Nov 2011  
[More results from stackoverflow.com](#)

**Class: String (Ruby 2.2.0) - Ruby-Doc.org**  
<https://ruby-doc.org/core-2.2.0/String.html> ▼  
Jump to **capitalize** - capitalize -> new\_str click to toggle source. Returns a copy of str with the first character converted to uppercase and the remainder to lowercase. Note: case conversion is effective only in ASCII region. "hello".capitalize #=> "Hello" "HELLO".capitalize #=> "Hello" "123ABC".capitalize #=> "123abc".  
::try\_convert - capitalize! - chop - gsub

Старая версия

Google ruby capitalize

All News Shopping Afbeeldingen Video's Meer Instellingen Tools

Ongeveer 370.000 resultaten (0,34 seconden)

**Capitalize first letter in ruby - Stack Overflow**  
<https://stackoverflow.com/questions/3724913/capitalize-first-letter-in-ruby>  
The upcase method capitalizes the entire string. I need to capitalize only the first letter. Also, I need to support several popular languages, like German and Russian. How do I do it?  
7 antwoorden

|   |  |   |
|---|--|---|
| <b>Best Answer</b><br>197 votes   | Antwoord 2 van 7<br>37 votes   | Antwoord 3 van 7<br>16 votes  |
| It depends on Ruby version you use. Ruby 2.4 and higher it just works, as since this version ruby supports Unicode case mapping. "мария".capitalize | capitalize first letter of first word of string "kirk douglas".capitalize #=> "Kirk douglas" capitalize first letter of each word in rails: "kirk" | Unfortunately, it is impossible for a machine to upcase/downcase/capitalize properly. It needs way contextual information |

**Class: String (Ruby 2.2.0) - Ruby-Doc.org**  
<https://ruby-doc.org/core-2.2.0/String.html> ▼ [Vertaal deze pagina](#)  
Source new **capitalize** - capitalize -> new\_str click to toggle source. Returns a copy of str with the first

Новая версия

# Введение

## Обзор имеющихся решений

- Burel et al. (2012)  
"Automatic Identification of Best Answers in Online Enquiry Communities"
- Tian et al. (2013)  
"Towards Predicting the Best Answers in Community-Based Question-Answering Services"
- Gkotsis et al. (2014)  
"It's all in the Content: State of the art Best Answer Prediction based on Discretisation of Shallow Linguistic Features"

# Введение

## Обзор имеющихся решений

- Тестирование проводилось на данных ServerFault
- Виды фичей:  $A \leftrightarrow A$ ,  $A \leftrightarrow Q$ ,  $A$ , *user/answer – rating, thread*
- Лингвистических фичи (длина текста, количество предложений, читаемость и др.)
- Vector Space Model + TF-IDF для определения похожести
- Вероятностная униграммная модель для оценки вероятности ответа
- Группировка ответов и дискретизация фичей
- Не учитывается наличие сниппетов кода
- Alternating Decision Tree / Random Forest Classifier

# Введение

## Минусы имеющихся решений

- Не используется текст вопроса
- Не учитывается контекст слова в предложении
- Игнорируется семантика



# Цель и задачи

Цель: Научиться определять правильность ответа на StackOverflow, используя тексты вопроса и ответов

Задачи:

- Произвести сбор данных для обучения и тестирования
- Реализовать несколько моделей классификаторов на основе нейронных сетей, использующих тексты ответов
- Улучшить точность классификации за счет текста вопроса
- Сравнить результаты с имеющимися работами

# Данные

## Общие факты

Данные:

- База вопросов ServerFault
- XML-файл размером  $\sim 0.9GB$
- Новые данные могут быть получены с помощью API
- Формат файла: *type\_id, id, score, date, body*

### Анализ:

- 684 тысячи постов, из них 257 тысяч вопросов и 427 тысяч ответов
- 130 тысяч вопросов (51%) без отмеченного правильного ответа
- 28 тысяч вопросов (11%), у которых нет ни одного ответа
- Большое количество технических терминов и слов с опечатками

После обработки получили 80 тысяч вопросов и 180 тысяч ответов

# Данные

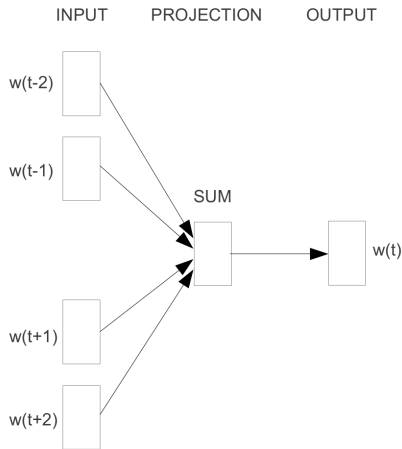
## Извлечение признаков

### Признаки:

- Векторные представления вопроса и ответа
- Лингвистические: количество ссылок, параграфов, сниппетов кода, длина текста, средняя длина предложения, различные индексы удобочитаемости
- *thread*: позиция ответа, относительная позиция ответа

# Подходы к представлению текста

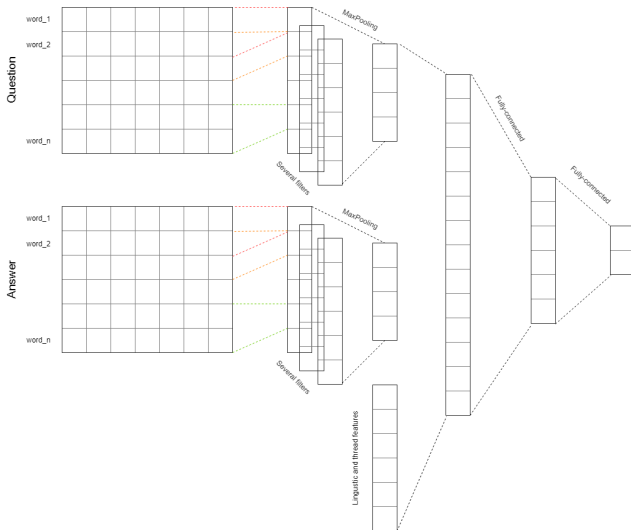
- Bag of words  
Минусы: не учитывается порядок слов и семантика, большая размерность
- Word2Vec  
Обучение на неразмеченном корпусе текстов, для каждого слова получаем вектор, отражающий его семантику.  
Минусы: не дает векторное представления для слов не из словаря
- Fasttext  
Модификация Word2Vec, основанная на работе с  $n$ -граммами букв



Архитектура модели CBOW

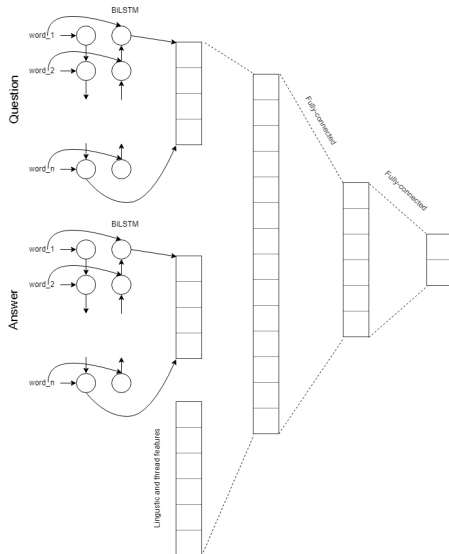
# Решение задачи

## Сверточная нейронная сеть



# Решение задачи

## Рекуррентная нейронная сеть



# Результаты

| Модель                  | Виды признаков  | Acc         | P    | R    | F <sub>1</sub> | AUC         |
|-------------------------|---|-------------|------|------|----------------|-------------|
| First-answer            | метаинформация  | 0.66        | 0.69 | 0.60 | 0.64           | 0.66        |
| Naive-Bayes with TF-IDF | словарные   | 0.6         | 0.58 | 0.54 | 0.56           | 0.59        |
| Burel et al. (2012)     | лингвистические,<br>словарные, поль-<br>зовательские,<br>метаинформация | -           | 0.77 | 0.77 | 0.76           | 0.83        |
| Tian et al. (2013)      | лингвистические,<br>метаинформация                                      | 0.72        | -    | -    | -              | -           |
| Gkotsis et al. (2014)   | лингвистические,<br>словарные, мета-<br>информация                      | -           | 0.83 | 0.66 | 0.74           | 0.85        |
| CNN                     | лингвистические,<br>текстовые, мета-<br>информация                      | <b>0.77</b> | 0.81 | 0.62 | 0.71           | <b>0.86</b> |
| RNN                     | лингвистические,<br>текстовые, мета-<br>информация                      | <b>0.78</b> | 0.82 | 0.64 | 0.72           | <b>0.87</b> |



- Подготовлен корпус данных, представляющий из себя ответы на вопросы с сайта Server Fault
- Реализовано несколько различных архитектур нейронных сетей для решения задачи определения наилучшего ответа на StackOverflow
- Показано, что текст вопроса является важным признаком для классификации
- Полученные результаты свидетельствуют о том, что нейронные сети справляются с задачей лучше методов, использованных в других статьях