

文本挖掘的两种简单方法

管理学院电子商务研究所

钱洋

□ 开源数据集(Open Source Data)

□ 网络爬虫(Crawler Data)

Kaggle.com | Kaggle Open Datasets | Write Code + Run Data Analysis

www.kaggle.com/

Download Open Datasets on 1000s of Projects + Share Projects on One Platform. Learn to Write Machine Learning Kernels + Run Analysis + Share with Community. Over 14,000 Datasets. Free Dataset Downloads. Analyze + Visualize Data. Local + International.

Genomics

Ask Bigger Questions By Efficiently Processing Petabytes Of Data.

BigQuery

A Fast, Scalable & Fully Managed ML Cloud Data Warehouse For Analytics.

Free Trial

Google Cloud Platform Free Tier
Learn and build on GCP for free.

All Products

Secure Your Data, Gain Real-Time Insights, Boost Productivity & More

Kaggle: Your Home for Data Science

<https://www.kaggle.com/>

Kaggle is the place to do data science projects. See how it works Play. Computer. Register with just one click: We won't share anything without your permission.

来自kaggle.com的搜索结果

Competitions

Two Sigma: Using News - Titanic - Google Cloud & NCAA® ML

Datasets

U.S. Education Datasets: Unification Project. Roy Garrard ...

Learn

Python - Pandas - Deep Learning - Microchallenges - ...

Kernels

Kernels. Documentation. New Kernel. Public. Your Work ...

Datasets

Documentation

New Dataset

Search 17066 datasets

Feedback

Filter

Sort by: Hottest



Heart Disease UCI

ronit

2 years 3 KB 8.8 1 File (CSV)

1964

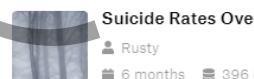


Mail Customer Segmentation Data

Vijay Choudhary

10 months 2 KB 8.8 1 File (CSV)

273

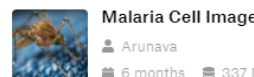


Suicide Rates Overview 1985 to 2016

Rusty

6 months 396 KB 8.2 1 File (CSV)

1037

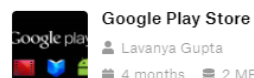


Malaria Cell Images Dataset

Arunava

6 months 337 MB 7.5 1 File (other)

431

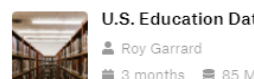


Google Play Store Apps

Lavanya Gupta

4 months 2 MB 7.1 3 Files (CSV, other)

1704



U.S. Education Datasets: Unification Project

Roy Garrard

3 months 85 MB 7.6 8 Files (other, CSV)

228

□ 开源数据集(Open Source Data)

□ 网络爬虫(Crawler Data)

阿里云天池: 天池大数据众智平台

tianchi.aliyun.com/ ▼

天池是阿里云旗下大数据平台, 围绕云生态挖掘输送优秀人才。旨在打造“数据众智、众创”平台, 欢迎来自世界各地的技术人员来天池参与百万奖金的天池大赛, 进行 ...

[天池大赛](#) · [数据集](#) · [关于我们](#) · [技术圈](#)

天池大数据竞赛-天池大赛-阿里云天池 - Aliyun

tianchi.aliyun.com/competition/gameList/activeList ▼

天池大数据竞赛,是由阿里巴巴集团主办,面向全球科研工作者的算法竞赛。通过开放海量数据和分布式计算资源,大赛让所有参与者有机会运用其设计的算法解决 ...

天池大赛 - Aliyun

<https://tianchi.aliyun.com/competition/> ▼

赛事简要: 新人赛以“地铁乘客流量预测”为课题, 提供杭州地铁票务数据, 要求参赛者预测未来流量变化, 实现地铁乘客流量预测, 用大数据和人工智能技术助力未来 ...

官方数据	公共数据	我的数据	输入搜索内容	Q	+ 新建
A Labeled Chinese Dataset for Diabetes / 中文糖尿病标注数据集					
数据集来源于中文糖尿病领域权威期刊, 数据包括基础研究、临床研究、药物使用、临床病例、治疗方法等多个方面, 时间跨度达到7年, 涵盖了近年来糖尿病领域最广泛的研究内容和热点。					
天池小T	2019-04-29	524	58	3547	4
Dense Arbitrary shaped text Dataset / 密集不规则文本行数据集					
Dense Arbitrary shaped text Dataset / 密集不规则文本行数据集					
天池小T	2019-03-21	145	23	1777	0
Product Description Dataset / 商品描述文案数据集					
用于商品描述生成的数据集: The dataset used for the product description generation.					
天池小T	2019-01-31	157	9	1418	0
Cloud Theme Click Dataset / 云主题点击数据集					
数据集为淘宝APP中云主题场景的用户点击日志, 用以对用户在不同场景、以及新场景下的推荐进行优化验证。This dataset is click data of Cloud Theme, which is an important recommendation procedu...					
天池小T	2019-01-31	61	1	776	0

PaperWeekly 4月3日 08:04 来自 微博 weibo.com

【国内数据竞赛优胜解决方案集锦】国内各个竞赛平台也在逐步发展，很多国内竞赛的获奖团队会热心地公开自己的算法甚至是源码。Github上有一个repo专注于搜集整理国内算法竞赛的各个优胜解决方案，目前为止已经收集了国内几家较大的数据竞赛平台的42场比赛的104个解决方案，其中72个方案附有开源代码。

... 展开全文

☆ 收藏 | 170 | 22 | 88

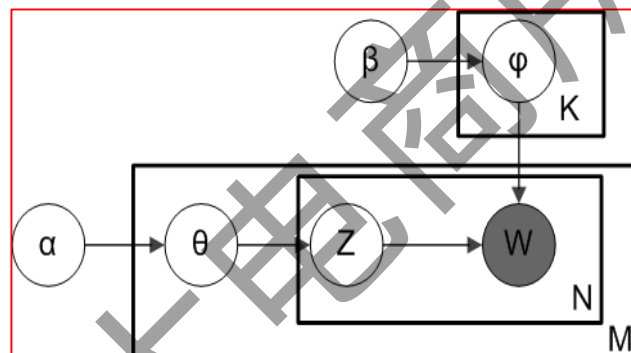
PaperWeekly 3月14日 08:01 来自 微博 weibo.com

【推荐系统综述】A Survey on Session-based Recommender Systems
#综述论文# #RecSys# 本文是第一篇全面深入总结session-based recommendations的综述文章，值得推荐。文章系统总结了目前一种新型推荐范式：session-based recommendations的特点、挑战和目前取得的进展，对整个推荐系统研究领域和相关 ... 展开全文

Input Data

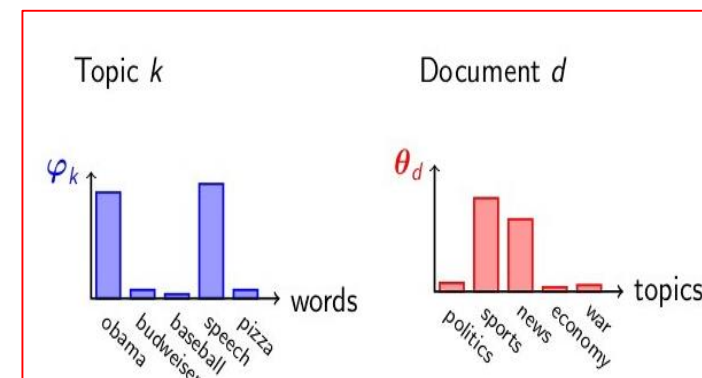
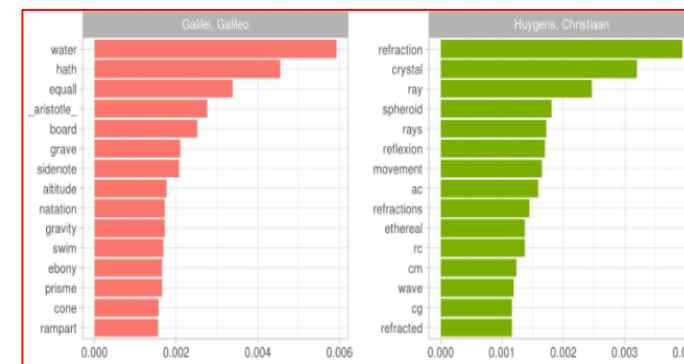
$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

$tf_{i,j}$ = number of occurrences of i in j
 df_i = number of documents containing i
 N = total number of documents



Model

(TF-IDF 和 LDA)



Output

现身说法:移动互联网让买房更简单 <http://t.cn/RLNbjje0> 就喜欢简单高效! 哈哈哈哈哈👍 我在用 @Weico微博客户端, 智能夜间模式、个性定制阅读环境、还有离线模式节省流量! 想体验更好? ★★★★★ 戳链接下载Weico Pro 4:<http://t.cn/RAPgR4n>

- ◆ 以句子的形式呈现
- ◆ 存在标点符号
- ◆ 存在URL
- ◆ 存在非法字符
- ◆ 存在停用词

◆ 数据处理，尤其文本处理，大部分时间都花在预处理上。如：

- 将单个用户的数据进行合并。
- 分词、去停用词。
- 去标点符号、去URL链接、去非法字符等。



程序讲解

◆ 英文数据处理与中文数据处理，存在哪些差异？



It's Not Just Coffee: Technology Now Fueling Massive Growth in Seattle

John Boitnott / AI, Entrepreneurs, Tech

Seattle has long been known for inciting the country's coffee culture, but now that we're all sufficiently caffeinated, these days it's the technology scene that's causing a buzz. Situated well north of the traditional tech enclave known as Silicon Valley, Seattle has come into its own as a magnet for innovative companies and a supportive [...]



How IoT Affects The Future Of Web Developments

Harshal Shah / AI, Code, IOT

ATMs, Apple Watch, Fitbit, Self-driving cars and more are part of a list of devices using IoT is endless. The concept of IoT began years ago; we just didn't use the term IoT. Kevin Ashton is credited with being the first to coin the phrase Internet of Things (IoT) in 1999. The Internet Of Things [...]



5 ways Artificial Intelligence is Revolutionizing eCommerce Marketing

Rameeza Yasin / AI, Marketing, Productivity

Marketing strategies are continuously changing due to increasing technological advancement. Technology is responsible for raising the expectations of people all around the globe. With an increase in innovations, it's pretty obvious that customers are expecting more and companies are doing everything to can to remain in their good list. Consequently, competition between companies seems to [...]

<https://readwrite.com/category/ai/>

WAIT,
WHAT?





谢谢