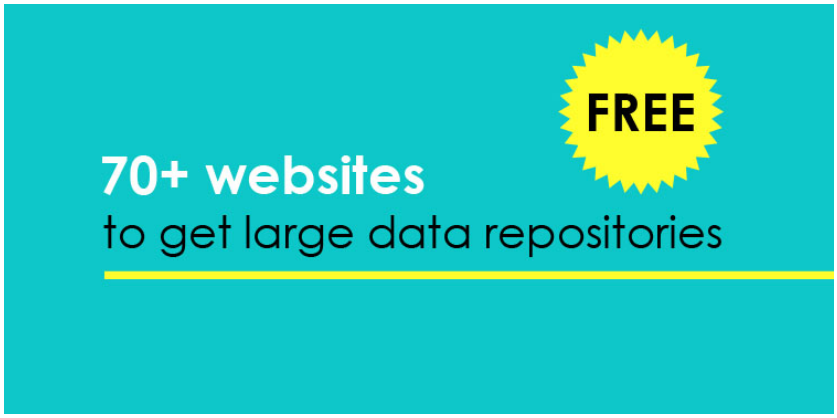
大数据：70多个网站让你免费获取大数据存储库

时间：2014-06-25 14:09 来源：199it 作者：Jane in



你是否需要大量的数据来检验你的APP性能？最简单的方法是从网上免费数据存储库下载数据样本。但这种方法最大的缺点是数据很少有独特的内容并且不一定能达到预期的结果。以下是70多家可以获得免费大数据存储库的网站。

Wikipedia:Database：向感兴趣的用户提供所有可用的内容的免费副本。可以得到多种语言的数据。内容连同图片可以下载。

Common crawl 建立并维护一个所有人都可以访问的开放的网络。这个数据保存在亚马逊s3bucket中，请求者可能花费一些钱来访问它。

Common crawl：建立并维护一个开放的网络，向所有人开放。

EDRM File Formats Data Set：由381个文件夹200种文件格式组成。

Apache Mahout TLP项目创建一个可扩展的机器学习算法。Mahout有许多免费的和付费的语料库语料。

EDRM Enron Email Data Set v2由安然公司邮件信息和附件组成，存在两组可下载的压缩文件中：XML和PST。

ClueWeb09用来支持信息检索和相关人类语言技术研究的资料库。它包含了从**2009**年**1**月到**2**月间收集的大约**10**亿个网页，包含**10**种语言。资料库被若干**TREC**会议的追踪检测使用。

DMOZ –最大的、最全面的人工编辑的开放式网站目录。它收集了不同类型的网站链接。**Dmoz**是互联网搜索引擎的一个主要来源。

theinfo.org –这是一个大数据集网站，在这里学者、设计师、艺术家等可以交流技巧和窍门，一起开发和共享工具，并开始整合他们独有的项目。

Project Gutenberg 提供超过**36000**免费电子书的下载，可以下载到个人电脑、**Kindle, Android, iOS or** 或其他便携式设备。

Million song data set：与**tracks** 和艺术家有关的数据

AWS (Amazon Web Services) Public Data Sets：提供了可以无缝融入**AWS**（亚马逊网络服务）云应用的公共数据集的集中存储库。

BigML big list of public data sources.

Bioassay data：研究文章**"**生物测定数据的虚拟筛选**"**，由**Amanda Schierz**编写，有**21**个生物测定数据集（活性**/**非生理活性成分），可以下载。

Bitly 1.usa.gov data：匿名点击政府链接

Canada Open Data：有许多政府和地理空间的数据集的试点项目

Canada Open Data：许多政府和地理空间数据集的试点项目。

Causality Workbench：数据存储库

Corral Big Data repository：在德克萨斯高级计算中心，提供以数据为中心的技术。

Data Source Handbook:公开数据指南

Datacatalogs.org：来自美国、欧盟、加拿大、**CKAN**以及其他的公开政府数据

Data.gov.uk：英国的公共可用数据（**London datastore**也是）

Data.gov/Education：对于教育数据资源的主要指南，包括高价值的数据集、数据可视化、课堂资源、创建自公开数据的应用程序以及其他。

DataMarket：可视化的世界经济、社会、自然和工业，拥有来自联合国，世界银行，欧盟统计局和其他重要数据提供者的一亿时间序列。

Datamob：可以很好利用的公开数据

DataSF.org：可向**City & County of San Francisco, CA.**购买的数据集信息交流中心

DataFerrett：一个用来访问和使用**The Data Web**的数据挖掘工具，许多网上美国政务数据集的集合。

EconData：大量经济学的时间序列，由许多美国政府机构编制。

Enron Email Dataset：来自大约**150**个用户的数据，这些用户大多数是安然公司高级管理人员

Europeana Data：包含**2000**万文字，图片，视频开放的元数据，以及由欧洲数位图书馆收集的声音，对于欧洲文化遗产内容值得信赖的、全面的资源。

Europeana Data：

FEDSTATS：一个美国统计资料的综合资源以及更多

FIMI repository for frequent itemset mining：工具和数据集

Financial Data Finder at OSU：大型财务数据集目录

GDELT:关于事件、位置和音调的全球数据，被英国卫报形容为**"**生命、宇宙和一切的大数据历史**"**

GEO (GEO Gene Expression Omnibus)：一个支持**MIAME**兼容数据提交的基因表达**/**分子丰度信息库，一个精心策划的网上资源，用于基因表达数据的浏览，查询和检索。

GeoDa Center：地理和空间数据

Google ngrams datasets：来自数Google扫描的百万书籍文本

Grain Market Research：财务数据，包括股票、期货等

Hilary Mason research-quality Big Data sets收集许多文本和图片数据集

HitCompanies Datasets：**HitCompanies**随机取样的**1**万个英国公司全面的数据，采用人工智能/机器学习进行自动更新。

ICWSM-2009 dataset：包含**2008**年**8**月**1**日到**10**月**1**日之间的**4400**万个博文

Infochimps：一个数据开放的目录和集合，允许分享、出售和下载关于任何内容的数据。

Investor Links：包含财物数据

KDD Cup center：数据、工作表和结果

Kevin Chai list of datasets：文本、**SNA**和其他领域

KONECT：科布伦茨网络收集，拥有大量各种类型的网络数据集，以便在网络挖掘领域进行研究。

Linking Open Data 工程，免费向所有人提供数据

MIT Cancer Genomics gene expression datasets and publications：来自麻省理工**Whitehead Center**用于基因组研究

ML Data：欧盟**Pascal2**网络数据储存库

NASDAQ Data Store：提供市场数据

National Government Statistical Web Sites：来自大约**70**个网站的数据、报告、统计年鉴、新闻和其他，包括非洲、欧洲、亚洲和拉丁美洲的国家。

National Space Science Data Center (NSSDC)：美国国家航空航天局的数据集，包含行星探索、空间和太阳物理学、生命科学、天体物理学以及其他方面。

Open Data Census：评估世界各地的开放数据的状态。

OpenData from Socrata：允许访问超过**10000**个数据集，包括商业、教育、政府和娱乐

Open Source Sports：大量运动数据库，包括棒球、足球、篮球和曲棍球

Peter Skomoroch dataset Bookmarks PubGene(TM) Gene Database and Tools：基因组有关的出版物数据库

Quandl, a collaboratively curated portal to millions of financial and economic time-series datasets.

qunb：一个用来发现和可视化的数据资料的平台

Robert Schiller data：住房建筑、股票市场和更多的来自于他的书 **Irrational Exuberance**的数据

SMD: Stanford Microarray Database,存储来自微阵列实验的原始的和标准的数据

Jerry Smith dataset collection：财经、政府、机器学习、科学和其他数据

SourceForge.net Research Data：包含大约**10**万个项目和超过**100**万注册用户的活动的历史和现状的统计数据的项目管理网站。

StatLib,卡内基梅隆大学数据档案

STATOO Datasets part 1和 **STATOO Datasets part 2**

Time Series Data Library

Visual Analytics Benchmark Repository.

UCI KDD Database Repository ： 适用于机器学习和知识发现研究的大数据集

UCI Machine Learning Repository.

UCR Time Series Data Archive：提供数据集、论文、链接和代码

United States Census Bureau.

Wikiposit：一个（虚拟的）融合了来自许多不同网站的数据（大多数是金融的），允许用户合并来自不同来源的数据

Wolfram Alpha disease and patient level dat.

Yahoo Sandbox datasets：语言、图表、评级、广告与营销、竞赛

Yelp Academic Dataset：30家大学的250个最接近商业的所有数据和评论，为学生和学者来探讨和研究

199IT编译自http://www.bigdata-madesimple.com/70-websites-to-get-large-data-repositories-for-free/

(责任编辑：mengyishan)

发表评论

请自觉遵守互联网相关的政策法规，严禁发布色情、暴力、反动的言论。

评价: 中立 好评 差评

匿名?

发表评论

最新评论 进入详细评论页>>

相关新闻

- 大数据：70多个网站让你免费获取大数据存储库
- 全球医疗卫生领域公共数据开放比较
- 日拟修个人信息保护法助大数据利用
- 我国首个大数据 交易行业规范发布
- 微软、IBM对垒大数据
- 北京警方利用大数据预测犯罪趋势
- 上海拟计划三年培养和引进近千名高端大数据人才
- 大数据如何改变传统防灾减灾模式?

热点专题更多>>

- 

2014 WOT全球软件
- 

第十届南京软博会
- 

大数据国家档案

- 云世界大会三年历

云资料下载更多>>

- 微软：预测2025年网络空间：大数据预见未
- 云华时代：2013年北京云基地白皮书
- 2013北京(国际）开源大会PPT下载
- 【淘金大数据】线下沙龙PPT下载
- 数据挖掘研究案例
- 工信部电信研究院：《中国公共云服务发展
- InfoWorld：深入探讨移动与BYOD
- 《针对性攻击与投机黑客：云安全状况》报

热点排行

1. 大数据：70多个网站让你免费获取大数据存
2. 全球医疗卫生领域公共数据开放比较
3. 日拟修个人信息保护法助大数据利用
4. 我国首个大数据 交易行业规范发布
5. 微软、IBM对垒大数据
6. 北京警方利用大数据预测犯罪趋势
7. 上海拟计划三年培养和引进近千名高端大数
8. 大数据如何改变传统防灾减灾模式？
9. 大数据时代，政府咋用数据？
10. 华为在法国建立大数据研发中心