数据集网址集合

http://archive.ics.uci.edu/ml/index.php
http://aws.amazon.com/publicdatasets/
http://www.kaggle.com/competitions
http://www.kdnuggets.com/datasets/index.html
https://mp.weixin.qq.com/s?
__biz=MzI4ODU5NjQ3OQ==&mid=2247483972&idx=1&sn=c7f7bbb3312934468912705d74d7c07f&chksm=ec3d4ad4db4ac3c2
http://mp.weixin.qq.com/s/4eDan-7KNnwgVP0DT96gzQ
http://mp.weixin.qq.com/s/tKc72xnqu4R4wkrVbK_bXA （偏国内，包含工具）
https://www.quandl.com/

http://archive.ics.uci.edu/ml/datasets.html

20G的金融行业数据集

http://mp.weixin.qq.com/s/_NS0UUDr84yq0rLg7jfr5g

图片数据

http://labelme.csail.mit.edu/Release3.0/index.php?message=1
http://www.image-net.org/index

吴恩达医学数据

http://mp.weixin.qq.com/s/M3s3z3YnEBvUxpDVGFVKHw

影像数据

http://www.91weitu.com/

气象

http://172.16.14.141:9100/

爬虫工具

https://www.oschina.net/p/beanbun
https://mp.weixin.qq.com/s/5rtoVnhYcVZpuRszr88diQ
https://gitee.com/xiyouMc/pornhubbot
https://gitee.com/l-weiwei/spiderman
https://gitee.com/flashsword20/webmagic

古诗

https://github.com/chinese-poetry/chinese-poetry

Datasets

Neural Networks used for supervised learning are notoriously data hungry. That’ s why open datasets are an incredibly important contribution to the research community. The following are a few datasets that stood out this year:

- Youtube Bounding Boxes
- Google QuickDraw Data
- DeepMind Open Source Datasets
- Google Speech Commands Dataset
- Atomic Visual Actions
- Several updates to the Open Images data set
- Nsynth dataset of annotated musical notes
- Quora Question Pairs

Public Data Sets on Amazon Web Services (AWS)

http://aws.amazon.com/datasets
Amazon从2008年开始就为开发者提供几十TB的开发数据。

Yahoo! Webscope

http://webscope.sandbox.yahoo.com/index.php

Konect is a collection of network datasets

http://konect.uni-koblenz.de/

Stanford Large Network Dataset Collection

http://snap.stanford.edu/data/index.html

安全相关的数据集

http://www.secrepo.com/

几个跟互联网有关的数据集：

1、Dataset for "Statistics and Social Network of YouTube Videos"
http://netsg.cs.sfu.ca/youtubedata/

2、1998 World Cup Web Site Access Logs
http://ita.ee.lbl.gov/html/contrib/WorldCup.html
这个是1998年世界杯期间的数据集。从1998/04/26 到 1998/07/26 的92天中，发生了 1,352,804,107次请求。

3、Page view statistics for Wikimedia projects
http://dammit.lt/wikistats/

4、AOL Search Query Logs - RP
http://www.researchpipeline.com/mediawiki/index.php?title=AOL_Search_Query_Logs

5、livedoor gourmet
http://blog.livedoor.jp/techblog/archives/65836960.html

海量图像数据集：

1、ImageNet
http://www.image-net.org/
包含1400万的图像。

2、Tiny Images Dataset
http://horatio.cs.nyu.edu/mit/tiny/data/index.html
包含8000万的32x32图像。

3、 MirFlickr1M
http://press.liacs.nl/mirflickr/
Flickr中的100万的图像集。

4、 CoPhIR
http://cophir.isti.cnr.it/whatis.html
Flickr中的1亿600万的图像

5、SBU captioned photo dataset
http://dsl1.cewit.stonybrook.edu/~vicente/sbucaptions/
Flickr中的100万的图像集。

6、Large-Scale Image Annotation using Visual Synset(ICCV 2011)
http://cpl.cc.gatech.edu/projects/VisualSynset/
包含2亿图像

7、NUS-WIDE
http://lms.comp.nus.edu.sg/research/NUS-WIDE.htm
Flickr中的27万的图像集。

8、SUN dataset
http://people.csail.mit.edu/jxiao/SUN/
包含13万的图像

9、MSRA-MM
http://research.microsoft.com/en-us/projects/msrammdata/
包含100万的图像，23000视频

10、TRECVID
http://trecvid.nist.gov/

Stack Overflow Dump Files
7.3G stackoverflow.com-Posts.7z
573.1K stackoverflow.com-Tags.7z
153.0M stackoverflow.com-Users.7z
2.2G stackoverflow.com-Comments.7z

截止目前好像还没有国内的企业或者组织开放自己的数据集。希望也能有企业开发自己的数据集给研究人员使用，从而推动海量数据处理在国内的发展！

2014/07/07 雅虎发布超大Flickr数据集 1亿的图片+视频

http://yahoolabs.tumblr.com/post/89783581601/one-hundred-million-creative-commons-flickr-images-for

100多个有趣的数据集

http://www.csdn.net/article/2014-06-06/2820111-100-Interesting-Data-Sets-for-Statistics