Data Sources

Aus DIMA Wiki

Data/source	Description	Link
Various data for Mining frequent items/itemsets	Data used in KDD cup tasks/competitions and other frequent itemset repository	http://www.sigkdd.org/kddcup/index.php http://fimi.ua.ac.be/data/
Reddit comments	challenging and interesting large dataset, that might serve very well for benchmarks	http://thenextweb.com/insider/2015/07/10/you-can-now-download-a-dataset-of-1-65-billion-reddit-comments-beware-the-redditor-ai/
Evolutionary Computation for Big Data and Big Learning Workshop. Data Mining competition 2014	Protein Structure Prediction. The dataset has 32 million instances, 631 attributes, 2 classes (56GB). Related paper: http://bioinformatics.oxfordjournals.org/content/28/19/2441	http://cruncher.ncl.ac.uk/bdcomp/ Training set (3723MB): http://cruncher.ncl.ac.uk/bdcomp/TrainingSet.arff.gz Test set (347MB): http://cruncher.ncl.ac.uk/bdcomp/TestSet.arff.gz
DEBS Challenge seriers provide real (relatively big) sets	2012: monitoring of large hi-tech manufacturing equipment 2013: Sports football monitoring systems 2015: Taxi trips in NY	http://www.csw.inf.fu-berlin.de/debs2012/grandchallenge.html http://www.orgs.ttu.edu/debs2013/index.php?goto=cfchallengedetails http://www.debs2015.org/call-grand-challenge.html
CancerData.org	Several sets of cancer related data	https://www.cancerdata.org/data?q=collections https://www.cancerdata.org
The Koblenz Network Collection	Several data sets, various sizes, topics	http://konect.uni-koblenz.de
Big data resources U. Stanford	Stanford Large Network Dataset Collection	http://snap.stanford.edu/data/index.html http://snap.stanford.edu http://mmds.org
Text, several languages	A collection of sentences and translations, 3874893 (17.04.2015) sentences and growing	http://tatoeba.org/eng/
Knime	Several, not so big sets, used in their tutorials and white paper examples	http://www.knime.org/white-papers
Used in Knime and DEBS2015 Challenge: http://www.debs2015.org/call-grand-challenge.html	NY Taxi trips	DEBS2015 link (12GB): https://drive.google.com/file/d/0B4zFfvIVhcMzcWV5SEQtSUdtMWc/view?usp=sharing Chris Whong FOIL request (48GB): http://chriswhong.com/open-data/foil_nyc_taxi/ http://www.andresmh.com/nyctaxitrips/ NYC Taxi & Limousine Commission (TLC) (about 175 GB)

		http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml
Capital Bike Share	Data (400 MB) from bike sharing system of Washington D.C.	http://www.capitalbikeshare.com/trip-history-data
Criteo	Criteo released a 1TB click log dataset containing several billion data points	http://labs.criteo.com/2015/03/criteo-releses-its-new-dataset/
USA National Climatic Data Center (NCDC)	Climate and historical weather	http://www.ncdc.noaa.gov
CERN releases actual Big Data from its CMS	In total ~24PB of experimental measurements from CMS (http://home.web.cern.ch/about/experiments/cms), including data set documentations, tools and (to some extent) visualizations of the measurements. Other experiments will follow.	http://cms.web.cern.ch/news/cms-releases-first-batch-high-level-lhc-open-data
Kaggle.com Academic Machine Learning Competitions	Various links to data sources, among other amazon open data sources	https://www.kaggle.com/wiki/DataSources https://inclass.kaggle.com/ https://aws.amazon.com/datasets/
Helix Nebula Cloud technology	(it seems they have data available, we have asked...)	http://helix-nebula.eu
arXiv	800K documents	http://arxiv.org
Wikipedia	25 TB	http://www.wikipedia.org
Amazon	varying sources	http://aws.amazon.com/datasets
Datasets for Data Mining and Data Science	varying sources	http://www.kdnuggets.com/datasets/index.html
Big data sets available for free	various sources	http://www.datasciencecentral.com/profiles/blogs/big-data-sets-available-for-free
Github list of public data sets	various sources	http://www.datasciencecentral.com/profiles/blogs/great-github-list-of-public-data-sets
Link to H20 list of public data sets	various sources	http://docs.h2o.ai/resources/publicdata.html
Various sources	Social graphs	http://snap.stanford.edu/data/index.html http://law.di.unimi.it/datasets.php
Various sources	collaborative filtering (user-item-ratings)	http://labrosa.ee.columbia.edu/millionsong/ http://grouplens.org/datasets/movielens/ https://www.kddcup2012.org/ http://www.cise.ufl.edu/research/sparse/matrices/
		http://stat-computing.org/dataexpo/2009/

Various sources	Classification	http://missionlocal.org/san-francisco-restaurant-health-inspections/
Amazon	+ amazon aws public datasets	http://aws.amazon.com/datasets
UC Irvine Machine Learning Repository	Most of the sets are small, but some sets have a "large" number of features	ordered buy number of instances: http://archive.ics.uci.edu/ml/datasets.html?format=&task=&att=&area=&numAtt=&numIns=&type=&sort=instDown&view=table ordered by number of features/attributes: http://archive.ics.uci.edu/ml/datasets.html?format=&task=&att=&area=&numAtt=&numIns=&type=&sort=attDown&view=table
LIBSVM Data: Classification, Regression, and Multi-label	Various sets (various sizes), several included in competitions/challenges	http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/
Dresden Web Table Corpus	Millions of html web tables in JSON data structure. Collected with web crawl and manually tagged.	https://wwwdb.inf.tu-dresden.de/misc/dwtc/
PhysioNet	PhysioNet offers free web access to large collections of recorded physiologic signals (PhysioBank), 75 collection of recordings, 4 TB of data (total amount)	http://physionet.org/physiobank/database/#ecg
ISSDA: Irish Social Science Data Archive	Various datasets for scientific and teaching purposes. Including Smart metering data (utilized in HP IoTAbench). N.B. Some data requires to fill in an invasive request form.	http://www.ucd.ie/issda/data/
Datasets for Geeks	Various machine learning datasets	http://www.datasets.co/