Links to data sets for testing algorithms

Waldemar W. Koczkodaj
<wkoczkodaj@cs.laurentian.ca>

https://www.researchgate.net/profile/Waldemar_Koczkodaj2
https://www.researchgate.net/publication/271705962_The_highest_recognition_download_by_the_Nobel_laureate
http://www.cs.laurentian.ca/wkoczkodaj/info.html

Abstract

Most of the links were copied form Research Gate questions. They have been a bit cleaned for easier use. It may be further enhanced, commented and possibly published. Some of it may be good for homework assignments.

1. Links

World Health Organization:
http://apps.who.int/gho/data/?theme=main
Health data are not medical data!

http://vision.ucsd.edu/content/yaleface-database

http://vision.ucsd.edu/content/extended-yale-face-database-b-b

Links to face databases
http://www.face-rec.org/databases/
http://web.mit.edu/emeyers/www/face_databases.html

Multifunction Databases/datasets

http://ftp.ics.uci.edu/pub/machine-learning-databases/
http://lib.stat.cmu.edu/datasets/ (StatLib Datasets Archive)
http://www.statsci.org/datasets.html (StatSci Data sets)
http://physionet.org/physiobank/database/ (PhysioBank currently contains over 40,000 recordings of annotated, digitized physiologic signals and time series, organized in over 60 databases (collections of recordings). All are freely available from PhysioNet)
http://biostat.mc.vanderbilt.edu/wiki/Main/DataSets

Corpora for linguistics & NLP
http://corpus.byu.edu/ (home of COCA, BYU's BNC, GlobWbe, Wikipedia Corpus, etc.)
http://verbs.colorado.edu/verb-index/ (Unified Verb Index: links to VerbNet, PropBank, FrameNet, & OntoNotes)
http://wordnetweb.princeton.edu/perl/webwn (WordNet Online; downloadable from the home page links)

http://multiwordnet.fbk.eu/english/home.php (MultiWordNet)

Microarray databases:
http://smd.princeton.edu/
http://www.broadinstitute.org/scientific-community/data
http://www.tech.plym.ac.uk/spmc/links/bioinformatics/bioinformatics_data.html
Bioinformatics Data links to datasets (some are down)

Important Statistics Organizations with publications and datasets
http://ec.europa.eu/eurostat
http://www.bjs.gov/index.cfm?ty=dca (Bureau of Justice Statistics)
http://www.icpsr.umich.edu/icpsrweb/ICPSR/index.jsp (Inter-university Consortium forPolitical and Social Research)

Topics

Databases

Data Acquisition

Statistical Data Analysis

Data

WordNet

INK"https://www.researchgate.net/topic/physiological_signals?ev=tp_pst_dtl_xkey"Physiological Signals

Sanskrit

Coca

Bioinformatics

Data.gov **http://data.gov**

The US Government pledged last year to make all government data available freely online. This site is the first stage and acts as a portal to all sorts of amazing information on everything from climate to crime.

US Census Bureau **http://www.census.gov/data.html**

A wealth of information on the lives of US citizens covering population data, geographic data and education.

European Union Open Data Portal **http://open-data.europa.eu/en/data/**

As the above, but based on data from European Union institutions.

Data.gov.uk **http://data.gov.uk/**

Data from the UK Government, including the British National Bibliography – metadata on all UK books and publications since 1950.

The CIA World Factbook **https://www.cia.gov/library/publications/the-world-factbook/**

Information on history, population, economy, government, infrastructure and military of 267 countries.

Healthdata.gov **https://www.healthdata.gov/**

125 years of US healthcare data including claim-level Medicare data, epidemiology and population statistics.

NHS Health and Social Care Information Centre **http://www.hscic.gov.uk/home**

Health data sets from the UK National Health Service.

Amazon Web Services public datasets **http://aws.amazon.com/datasets**

Huge resource of public data, including the 1000 Genome Project, an attempt to build the most comprehensive database of human genetic information and NASA's database of satellite imagery of Earth.

Facebook Graph **https://developers.facebook.com/docs/graph-api**

Although much of the information on users' Facebook profile is private, a lot isn't – Facebook provide the Graph API as a way of querying the huge amount of information that its users are happy to share with the world (or can't hide because they haven't worked out how the privacy settings work).

Gapminder **http://www.gapminder.org/data/**

Compilation of data from sources including the World Health Organization and World Bank covering economic, medical and social statistics from around the world.

Google Trends **http://www.google.com/trends/explore**

Statistics on search volume (as a proportion of total search) for any given term, since 2004.

Google Finance **https://www.google.com/finance**

40 years' worth of stock market data, updated in real time.

Google Books Ngrams **http://storage.googleapis.com/books/ngrams/books/datasetsv2.html**

Search and analyze the full text of any of the millions of books digitised as part of the Google Books project.

National Climatic Data Center **http://www.ncdc.noaa.gov/data-access/quick-links#loc-clim**

Huge collection of environmental, meteorological and climate data sets from the US National Climatic Data Center. The world's largest archive of weather data.

DBPedia **http://wiki.dbpedia.org**

Wikipedia is comprised of millions of pieces of data, structured and unstructured on every subject under the sun. DBPedia is an ambitious project to catalogue and create a public, freely distributable database allowing anyone to analyze this data.

Topsy **http://topsy.com/**

Free, comprehensive social media data is hard to come by – after all their data is what generates profits for the big players (Facebook, Twitter etc) so they don't want to give it away. However

Topsy provides a searchable database of public tweets going back to 2006 as well as several tools to analyze the conversations.

Likebutton **http://likebutton.com/**

Mines Facebook's public data - globally and from your own network - to give an overview of what people "Like" at the moment.

New York Times **http://developer.nytimes.com/docs**

Searchable, indexed archive of news articles going back to 1851.

Freebase **http://www.freebase.com/**

A community-compiled database of structured data about people, places and things, with over 45 million entries.

Million Song Data Set **http://aws.amazon.com/datasets/6468931156960467**

http://tdrdata.com/

http://www.healthdata.gov/

http://www.who.int/research/en/

http://www.geohive.com/

http://www.ilo.org/global/statistics-and-databases/lang--en/index.htm

http://tdrdata.com/

http://www.healthdata.gov/

http://www.who.int/research/en/

http://www.geohive.com/

http://www.ilo.org/global/statistics-and-databases/lang--en/index.htm

http://archive.ics.uci.edu/ml/
The European Soil Data Centre (ESDAC) offers for free all the data for soil:

http://eusoils.jrc.ec.europa.eu/

http://snap.stanford.edu/data/web-Amazon.html

Images data:
ftp://medical.nema.org/medical/dicom/Multiframe/
http://marathon.csee.usf.edu/Mammography/Database.html

Another:
http://fimi.ua.ac.be/data/
http://archive.ics.uci.edu/ml/datasets.html

http://www.kaggle.com/competitions
http://www.kdnuggets.com/datasets/

Data.gov http://data.gov

The US Government pledged last year to make all government data available freely online. This site is the first stage and acts as a portal to all sorts of amazing information on everything from climate to crime.

US Census Bureau http://www.census.gov/data.html

A wealth of information on the lives of US citizens covering population data, geographic data and education.

European Union Open Data Portal http://open-data.europa.eu/en/data/

As the above, but based on data from European Union institutions.

Data.gov.uk http://data.gov.uk/

Data from the UK Government, including the British National Bibliography – metadata on all UK books and publications since 1950.

The CIA World Factbook https://www.cia.gov/library/publications/the-world-factbook/

Information on history, population, economy, government, infrastructure and military of 267 countries.

Healthdata.gov https://www.healthdata.gov/

125 years of US healthcare data including claim-level Medicare data, epidemiology and population statistics.

NHS Health and Social Care Information Centre http://www.hscic.gov.uk/home

Health data sets from the UK National Health Service.

Amazon Web Services public datasets http://aws.amazon.com/datasets

Huge resource of public data, including the 1000 Genome Project, an attempt to build the most comprehensive database of human genetic information and NASA's database of satellite imagery of Earth.

Facebook Graph https://developers.facebook.com/docs/graph-api

Although much of the information on users' Facebook profile is private, a lot isn't – Facebook provide the Graph API as a way of querying the huge amount of information that its users are happy to share with the world (or can't hide because they haven't worked out how the privacy settings work).

Gapminder http://www.gapminder.org/data/

Compilation of data from sources including the World Health Organization and World Bank covering economic, medical and social statistics from around the world.

Google Trends http://www.google.com/trends/explore

Statistics on search volume (as a proportion of total search) for any given term, since 2004.

Google Finance https://www.google.com/finance

40 years' worth of stock market data, updated in real time.

Google Books Ngrams http://storage.googleapis.com/books/ngrams/books/datasetsv2.html

Search and analyze the full text of any of the millions of books digitised as part of the Google Books project.

National Climatic Data Center http://www.ncdc.noaa.gov/data-access/quick-links#loc-clim

Huge collection of environmental, meteorological and climate data sets from the US National Climatic Data Center. The world's largest archive of weather data.

DBPedia http://wiki.dbpedia.org

Wikipedia is comprised of millions of pieces of data, structured and unstructured on every subject under the sun. DBPedia is an ambitious project to catalogue and create a public, freely distributable database allowing anyone to analyze this data.

Topsy http://topsy.com/

Free, comprehensive social media data is hard to come by – after all their data is what generates profits for the big players (Facebook, Twitter etc) so they don't want to give it away. However Topsy provides a searchable database of public tweets going back to 2006 as well as several tools to analyze the conversations.

Likebutton http://likebutton.com/

Facebook's public data - globally and from your own network - to give an overview of what people "Like" at the moment.

New York Times http://developer.nytimes.com/docs

Searchable, indexed archive of news articles going back to 1851.

Freebase http://www.freebase.com/

A community-compiled database of structured data about people, places and things, with over 45 million entries.

http://data.un.org/Explorer.aspx (UN datasets)

https://data.cityofnewyork.us/ (NYC data)

https://data.gov.in/ (Indian Govt Datasets)

Microarray data repositories

www.ncbi.nlm.nih.gov/geo/ (US NCBI)

http://www.ebi.ac.uk/arrayexpress/ (UK EMBL-EBI)

Some other curated databases

http://ctdbase.org/ (Comparative Toxicogenomics Database)

http://www.hprd.org/ (Human Protein Reference Database)

http://www.reactome.org/ (Reactome, Pathway Database)

http://www.mirbase.org/ (microRNA database)

References

[1] Research Gate (Q&A section), accessed 2015-03-15