Yahoo! Webscope™ Datasets Catalog
April 2009, 21 Datasets Available

The "Yahoo! Webscope™ Program" is a reference library of interesting and scientifically useful datasets for non-commercial use by academics and other scientists. All datasets have been reviewed to conform to Yahoo!'s data protection standards, including strict controls on privacy. Data may be used only for academic use by faculty and other University researchers who agree to and sign the Data Sharing Agreement.

Yahoo! is pleased to make these datasets available to researchers who are advancing the state of knowledge and understanding in web sciences. The current Webscope catalog is listed below, along with the process to request a dataset. We ask that researchers recognize the value of Webscope data in their work, at a minimum, by acknowledging Yahoo! as source of the data when researchers publish their findings in technical journals, papers, conferences, keynote speeches, etc.

Process to request a dataset:
To request a dataset, simply send an email to research-data-requests@yahoo-inc.com. The request should specify the name of the dataset, the name of the principal investigator(s), the expected study period, the general research area, and technical conferences where results may be published. Multiple datasets may be requested, but we ask that each dataset have an intended research purpose in mind prior to sending the request.

Academic Researchers are required to sign the Webscope Data Sharing Agreement to access the datasets. Once we've received your data request, the Data Agreement will be sent to you for completion and signatures. Completed agreements may be scanned and emailed to kimcapps@yahoo-inc.com, or faxed to 408-336-0782. The researcher will receive a copy of the data on DVD in the mail by 7-10 days after the completed Data Sharing Agreement has been received by Yahoo!.

Webscope Datasets by Category

Ratings Data – These types of datasets can be utilized to research collaborative filtering, recommender systems and machine learning algorithms. Yahoo! is interested in various types of research in providing recommendations for users and personalized and more relevant content to users.

R1……Yahoo! Music User Ratings of Musical Artists, version 1.0
R2……Yahoo! Music User Ratings of Songs with Artist, Album, and Genre Meta Information, version 1.0
R3……Yahoo! Music ratings for User Selected and Randomly Selected Songs, version 1.0
R4……Yahoo! Movies User Ratings of Movies, and Movie Descriptive Content Information, version 1.0
R5……Yahoo! Delicious Popular URLs and Tags, version 1.0

Language Data – These types of datasets can be utilized to research information retrieval and natural language processing algorithms. Yahoo! is interested in improved search and information retrieval.

L1……..Yahoo! N-Grams, version 2.0
L2……..Metadata Extracted from Publicly Available Web Pages, version 1.0
L3……..Yahoo! Semantically Annotated Snapshot of the English Wikipedia, version 1.0

L4…….Yahoo! Answers Manner Questions, version 1.0
L5…….Yahoo! Answers Manner Questions, version 2.0
L6…….Yahoo! Answers Comprehensive Questions and Answers version 1.0
L7…….Yahoo! Answers Question Types, version 1.0
L8…….Yahoo! Search Query Logs for Nine Languages, version 1.0
L9…….Yahoo! Answers Question Types Sample of 1000, version 1.0

Graph Data - These types of datasets can be utilized to research matrix, graph, clustering, and machine learning algorithms. Yahoo! is interested in better understanding of social networks.

G1…….Yahoo! Search Marketing Advertiser-Phrase Bipartite Graph, version 1.0
G2…….Yahoo! AltaVista Web Page Hyperlink Connectivity Graph, circa 2002
G3…….Yahoo! Groups User-Group Membership Bipartite Graph, version 1.0
G4…….Yahoo! Network Flows Data, version 1.0

Advertising & Markets Data - These types of datasets can be utilized to research behavior and incentives in auctions markets. Yahoo! is interested in design of advertising systems and Yahoo! marketplaces.

A1…….Yahoo! Search Marketing Advertiser Bidding Data, version 1.0
A2…….Yahoo! Buzz Game Transactions with Buzz Scores, version 1.0
A3…….Yahoo! Search Marketing Advertiser Bid-Impression-Click data on competing Keywords, version 1.0

See also: **G1…….Yahoo! Search Marketing Advertiser-Phrase Bipartite Graph, v1.0**

Appendix: Yahoo! Webscope Global ReadMe Document

Ratings Data:

R1. Yahoo! Music User Ratings of Musical Artists, version 1.0 Size: 423 MB, 1 CD

This dataset represents a snapshot of the Yahoo! Music community's preferences for various musical artists. The dataset contains over ten million ratings of musical artists given by Yahoo! Music users over the course of a one month period sometime prior to March 2004. Users are represented as meaningless anonymous numbers so that no identifying information is revealed. The dataset may be used by researchers to validate recommender systems or collaborative filtering algorithms. The dataset may serve as a testbed for matrix and graph algorithms including PCA and clustering algorithms.

R2. Yahoo! Music User Ratings of Songs with Artist, Album, and Genre Meta Information, version 1.0 Size: 2.4 GB, 1 DVD

This dataset represents a snapshot of the Yahoo! Music community's preferences for various songs. The dataset contains over 717 million ratings of 136 thousand songs given by 1.8 million users of Yahoo! Music services.

The data was collected between 2002 and 2006. Each song in the dataset is accompanied by artist, album, and genre attributes. The users, songs, artists, and albums are represented by randomly assigned numeric id's so that no identifying information is revealed. The mapping from

genre id's to genre, as well as the genre hierarchy, is given.

R3. Yahoo! Music ratings for User Selected and Randomly Selected songs, version 1.0 Size: 4 MB, 1 CD

This dataset contains ratings for songs collected from two different sources. The first source consists of ratings supplied by users during normal interaction with Yahoo! Music services. The second source consists of ratings for randomly selected songs collected during an online survey conducted by Yahoo! Research. The rating data includes 15,400 users, and 1000 songs. The data contains at least ten ratings collected during normal use of Yahoo! Music services for each user, and exactly ten ratings for randomly selected songs for each of the first 5400 users in the dataset. The dataset includes approximately 300,000 user-supplied ratings, and exactly 54,000 ratings for randomly selected songs. All users and items are represented by randomly assigned numeric identification numbers. In addition, the dataset includes responses to seven multiple-choice survey questions regarding rating-behavior for each of the first 5400 users. The survey data and ratings for randomly selected songs were collected between August 22, 2006 and September 7, 2006. The normal-interaction data was collected between 2002 and 2006.

R4. Yahoo! Movies User Ratings of Movies, and Movie Descriptive Content Information, version 1.0 Size: 23 MB, 1 CD

This dataset contains a small sample of the Yahoo! Movies community's preferences for various movies, rated on a scale from A+ to F. Users are represented as meaningless anonymous numbers so that no identifying information is revealed. The dataset also contains a large amount of descriptive information about many movies released prior to November 2003, including cast, crew, synopsis, genre, average ratings, awards, etc. The dataset may be used by researchers to validate recommender systems or collaborative filtering algorithms, including hybrid content and collaborative filtering algorithms. The dataset may serve as a testbed for relational learning and data mining algorithms as well as matrix and graph algorithms including PCA and clustering algorithms.

R5. Yahoo! Delicious Popular URLs and Tags, version 1.0 Size: 6.39 MB, 1 CD

This dataset represents 100,000 URLs that were bookmarked on Delicious by users of the service. Each URL has been saved at least 100 times. For each URL, the date that it was first bookmarked by a Delicious user is indicated, along with the total number of saves. Also indicated are the ten most commonly used tags for each URL, along with the number of times each tag was used. This dataset provides a view into the nature of popular content in the Delicious social bookmarking system, including how users apply tags to individual items.

Language Data:

L1. Yahoo! N-Grams, version 2.0 Size: 3 Dual DVDs: 3.5GB, 4.2GB & 4.3GB

This dataset contains n-grams (contiguous sets of words of size n), n = 1 to 5, extracted from a corpus of 14.6 million documents (126 million unique sentences, 3.4 billion running words) crawled from over 12000 news-oriented sites. The documents were published on these sites between February 2006 and December 2006. The dataset does not contain the documents themselves, but only the n-grams that occur at least twice. It provides statistics such as frequency of occurrence, number and entropy of different left (right) single-token contexts of each n-gram. This dataset may be used by researchers to build statistical language models for speech or handwriting recognition or machine translation.

L2. Metadata Extracted from Publicly Available Web Pages, version 1.0 Size: 2.5 GB, 1 DVD

The dataset contains about 100 million triples of RDF data obtained by extracting metadata from publicly available webpages. Three forms of embedded metadata are extracted: microformats (hCard, hCalendar and hReview), RDFa metadata and RDF documents linked to webpages. All metadata extracted from a webpage is converted to RDF. The data is made available in the WARC format, version 0.9.
The dataset may serve as a testbed for research in scalability in the Semantic Web area and also for developing methods to deal with metadata that is incomplete, erroneous or biased in some way.

L3. Yahoo! Semantically Annotated Snapshot of the English Wikipedia, version 1.0 Size: 2 DVDs: 3.3 GB & 2.7 GB

This SW1 dataset contains a snapshot of the English Wikipedia dated from 2006-11-04 processed with a number of publicly-available NLP tools. In order to build SW1, we started from the XML-ized Wikipedia dump distributed by the University of Amsterdam. This snapshot of the English Wikipedia contains 1,490,688 entries (excluding redirects). First, the text is extracted from the XML entry and split into sentences using simple heuristics. Then we ran several syntactic and semantic NLP taggers on it and collected their output. Raw Data (Multitag format)
The multitag format contains all the Wikipedia text plus all the semantic tags. All other data files can be reconstructed from this. A multitag file contains several Wikipedia entries. The Wikipedia snapshot was cut into 3000 multitag files each containing roughly 500 entries.

L4. Yahoo! Answers Manner Questions, version 1.0 Size: 352 MB, 1 CD

Yahoo! Answers is a website where people post questions and answers, all of which are public to any web user willing to browse or download them. The data we have collected is a subset of the Yahoo! Answers corpus from a 10/25/2007 dump. It is a small subset of the questions, selected for their linguistic properties (for example they all start with "how {to|do|did|does|can|would|could|should}"). Additionally, we removed questions and answers of obvious low quality, i.e., we kept only questions and answers that have at least four words, out of which at least one is a noun and at least one is a verb. The final subset contains 142,627 questions and their answers.

In addition to question and answer text, the corpus contains a small amount of metadata, i.e., which answer was selected as the best answer, and the category and sub-category that was assigned to this question. No personal information is included in the corpus. The question URIs were replaced with locally-generated identifiers.
This dataset may be used by researchers to learn and validate answer-extraction models for manner questions. An example of such work was published by Surdeanu et al. (2008).

L5. Yahoo! Answers Manner Questions, version 2.0 Size: 103 MB, 1 CD

Yahoo! Answers is a website where people post questions and answers, all of which are public to any web user willing to browse or download them. The data we have collected is a subset of the Yahoo! Answers corpus from a 10/25/2007 dump. It is a small subset of the questions, selected for their linguistic properties (for example they all start with "how {to|do|did|does|can|would|could|should}"). Additionally, we removed questions and answers of obvious low quality, i.e., we kept only questions and answers that have at least four words, out of which at least one is a noun and at least one is a verb. The final subset contains 142,627 questions and their answers.

In addition to question and answer text, the corpus contains a small amount of metadata, i.e., which answer was selected as the best answer, and the category and sub-category that was assigned to this question. No personal information is included in the corpus. The question URIs

were replaced with locally-generated identifiers.
This dataset may be used by researchers to learn and validate answer-extraction models for manner questions. An example of such work was published by Surdeanu et al. (2008).

L6. Yahoo! Answers Comprehensive Questions and Answers version 1.0 Size: 3.45 GB, 1 DVD

Yahoo! Answers is a web site where people post questions and answers, all of which are public to any web user willing to browse or download them. The data we have collected is the Yahoo! Answers corpus as of 10/25/2007. It includes all the questions and their corresponding answers. The corpus distributed here contains 4,483,032 questions and their answers.

In addition to question and answer text, the corpus contains a small amount of metadata, i.e., which answer was selected as the best answer, and the category and sub-category that was assigned to this question. No personal information is included in the corpus. The question URIs and all user ids were anonymized so that no identifying information is revealed.
This dataset may be used by researchers to learn and validate answer extraction models. An example of such work was published by Surdeanu et al. (2008).

L7. Yahoo! Answers Question Types, version 1.0 Size: 900KB, 1 CD

This dataset contains URLs of questions posted to Yahoo! Answers, along with the question types assigned to these questions by human judges. The question types are "informational", "advice", "opinion", and "polling".

L8. Yahoo! Search Query Logs for Nine Languages, version 1.0 Size: 24MB, 1 CD

This dataset contains the 1000 most frequent web search queries issued to Yahoo! Search for nine different languages. The languages covered are Chinese, English, French, German, Italian, Japanese, Korean, Portuguese, and Spanish. The dataset may be useful for various information retrieval and data mining research investigations, especially those involving cross- or multi-lingual search tasks.

L9. Yahoo! Answers Question Types Sample of 1000, version 1.0 Size: 900KB, 1 CD

This dataset contains URLs of questions posted to Yahoo! Answers, along with the question types assigned to these questions by human judges. The question types are "informational", "advice", "opinion", and "polling".

Graph Data:

G1. Yahoo! Search Marketing Advertiser-Phrase Bipartite Graph, Version 1.0 Size: 14MB, 1 CD

Yahoo! Search Marketing operates Yahoo!'s auction-based platform for selling advertising space next to Yahoo! Search results. Advertisers bid for the right to appear alongside the results of particular search queries. For example, a travel vendor might bid for the right to appear alongside the results of the search query "Las Vegas travel." An advertiser's bid is the price the advertiser is willing to pay whenever a user actually clicks on their ad. Yahoo! Search Marketing auctions are continuous and dynamic: advertisers may alter their bids at any time, for example raising their bid for the query "buy flowers" during the week before Valentines Day.

This is a small completely anonymized graph reflecting the pattern of connectivity between some Yahoo! Search Marketing advertisers and some of the search keyword phrases that they bid on. Both advertisers and keyword phrases are represented as meaningless anonymous numbers so that no identifying information is revealed. The dataset contains 459,678 anonymous phrases ids, 193,582 anonymous advertiser ids, and 2,278,448 edges, representing the act of an advertiser bidding on a phrase.

G2. Yahoo! AltaVista Web Page Hyperlink Connectivity Graph, circa 2002 Size: 3 Dual DVDs: 4.5GB, 5.3GB & 4.4GB

This dataset contains URLs and hyperlinks for over 1.4 billion public web pages indexed by the Yahoo! AltaVista search engine in 2002. The dataset encodes the graph or map of links among web pages, where nodes in the graph are URLs. The Yahoo! AltaVista web graph is an example of a large real-world graph. The dataset may serve as a testbed for matrix, graph, clustering, data mining, and machine learning algorithms.

G3. Yahoo! Groups User-Group Membership Bipartite Graph, version 1.0 Size: 97 MB, 1 CD

Millions of communities and groups use Yahoo! Groups as a meeting place and forum to discuss mutual interests on nearly any topic. This dataset contains a sample of the "membership graph" of Yahoo! Groups, where both users and groups are represented as meaningless anonymous numbers so that no identifying information is revealed. Users and groups are nodes in the membership graph, with edges indicating that a user is a member of a group. The dataset consists only of the anonymous bipartite membership graph, and does not contain any information about users, groups, or discussions. The Yahoo! Groups membership graph is an example of a large real-world power-law graph. The dataset may serve as a testbed for matrix and graph algorithms including PCA and graph clustering algorithms, as well as machine learning algorithms.

G4. Yahoo! Network Flows Data, version 1.0 Size: 5 Dual DVDs or requestor provides one mini portable hard drive

Yahoo! network flows data contains communication patterns between end-users in the large Internet and Yahoo servers. A netflow record includes timestamp, source IP address, destination IP address, source port, destination port, protocol, number of packets, and number of bytes transferred from the source to the destination. The record does not include the content of the data communication.

Each Nntflow data file consists of sampled netflow records exported from routers in 15-minute intervals. The dataset includes netflow data files collected from three border routers in October 11 2007. All IP addresses in the dataset are anonymized using a random permutation algorithm. **Note:** due to the size of this dataset, we can provide five dual layer DVDs or ten 4.7 GB DVDs or you can send a mini portable hard drive that we'll transfer the data onto and return to you.

Advertising & Markets Data:

A1. Yahoo! Search Marketing Advertiser Bidding Data, version 1.0 Size: 80.1 MB, 1 CD

Yahoo! Search Marketing operates Yahoo!'s auction-based platform for selling advertising space next to Yahoo! Search results. Advertisers bid for the right to appear alongside the results of particular search queries. For example, a travel vendor might bid for the right to appear alongside the results of the search query "Las Vegas travel". An advertiser's bid is the price the advertiser is

willing to pay whenever a user actually clicks on their ad. Yahoo! Search Marketing auctions are continuous and dynamic.

Advertisers may alter their bids at any time, for example raising their bid for the query "buy flowers" during the week before Valentines Day. This dataset contains the bids over time of all advertisers participating in Yahoo! Search Marketing auctions for the top 1000 search queries during the period from June 15, 2002, to June 14, 2003. Advertisers' identities and query phrases are represented as meaningless anonymous numbers so that no identifying information about advertisers is revealed. The data may be used by economists or other researchers to investigate the behavior of bidders in this unique real-time auction format, responsible for roughly two billion dollars in revenue in 2005 and growing.

A2. Yahoo! Buzz Game Transactions with Buzz Scores, version 1.0 Size: 85 MB, 1 CD

The Yahoo!/O'Reilly Tech Buzz Game tests the theory that a free electronic market can predict trends in technology. In the game, users buy and sell fantasy "stocks" in various technologies. The prices of the stocks fluctuate according to supply and demand using a market mechanism invented at Yahoo! called a dynamic pari-mutuel market. Weekly dividends are paid out to stockholders based on the current "buzz score," or percentage of Yahoo! searches associated with each technology. This dataset shows all the transactions over the course of the Tech Buzz Game, divided into two periods. The first period spans April 1, 2005 to July 31, 2005, after which all markets were closed and cashed out according to final buzz scores at that time. The second period spans August 22, 2005 to the present, at the beginning of which new markets were established and the market reset to a starting point. Traders are represented as meaningless anonymous numbers so that no identifying information is revealed. In addition, the dataset contains buzz score data showing the daily percentages of search volume for each stock. Researchers may use the data to test the predictive value of the market or to test market behavioral theories.

A3. Yahoo! Search Marketing Advertiser Bid-Impression-Click data on competing Keywords, version 1.0 Size: 845 MB, 1 DVD

This dataset contains a small sample of advertiser's bid and revenue information over a period of 4 months. Bid and revenue information is aggregated with a granularity of a day over advertiser account id, keyphrase and rank. Apart from bid and revenue, impressions and clicks information is also included. Sequence of keywords make a keyphrase, keyphrase category is a keyword. A keyphrase can belong to one or more keyphrase categories. Keyphrase categories used for this dataset have been listed in a separate file. Advertiser account id is represented as a meaningless string. Keyphrase is represented as sequence of meaningless strings, where each string represents a keyword or keyphrase category.

Appendix: Yahoo! Webscope Global ReadMe Document: Information provided with each dataset

The data included herein is provided as part of the Yahoo! Webscope program for use solely under the terms of a signed Yahoo! Data Sharing Agreement.

Any publication using this data should attribute Yahoo!, ideally in the bibliography of the paper, unless Yahoo! explicitly requests no attribution. Please include the phrase, “Yahoo! Webscope,” the web address http://research.yahoo.com/Academic_Relations and the name of the specific dataset used, including version number if applicable. For example:

Yahoo! Webscope dataset ydata-ymusic-user-artist-ratings-v1_0
[http://research.yahoo.com/Academic_Relations]

Please send a copy of each paper and its full citation information to research-data-requests@yahoo-inc.com upon publication.

This data may be used only for academic research purposes and may not be used for any commercial purposes, by any commercial entity, or by any party not under a signed Data Sharing Agreement. The data may not be reproduced in whole or in part, may not be posted on the web, on internal networks, or in networked data stores, and may not be archived offsite. The data must be returned to Yahoo! at the end of the research project or in three years, whichever comes first.

This dataset was produced from Yahoo!'s records and has been reviewed by an internal board to assure that no personally identifiable information is revealed. You may not perform any analysis, reverse engineering or processing of the data or any correlation with other data sources that could be used to determine or infer personally identifiable information.

Please refer to the Data Sharing Agreement for complete terms. Contact research-data-requests@yahoo-inc.com with questions.