nlp-datasets

Alphabetical list of free/public domain datasets with text data for use in Natural Language Processing (NLP). Most stuff here is just raw unstructured text data, if you are looking for annotated corpora or Treebanks refer to the sources at the bottom.

Datasets

- [Apache Software Foundation Public Mail Archives](): all publicly available Apache Software Foundation mail archives as of July 11, 2011 (200 GB)

- [Blog Authorship Corpus](): consists of the collected posts of 19,320 bloggers gathered from blogger.com in August 2004. 681,288 posts and over 140 million words. (298 MB)

- [Amazon Fine Food Reviews [Kaggle]](): consists of 568,454 food reviews Amazon users left up to October 2012. [Paper](). (240 MB)

- [Amazon Reviews](): Stanford collection of 35 million amazon reviews. (11 GB)

- [ArXiv](): All the Papers on archive as fulltext (270 GB) + sourcefiles (190 GB).

- [ASAP Automated Essay Scoring [Kaggle]](): For this competition, there are eight essay sets. Each of the sets of essays was generated from a single prompt. Selected essays range from an average length of 150 to 550 words per response. Some of the essays are dependent upon source information and others are not. All responses were written by students ranging in grade levels from Grade 7 to Grade 10. All essays were hand graded and were double-scored. (100 MB)

- [ASAP Short Answer Scoring [Kaggle]](): Each of the data sets was generated from a single prompt. Selected responses have an average length of 50 words per response. Some of the essays are dependent upon source information and others are not. All responses were written by students primarily in Grade 10. All responses were hand graded and were double-scored. (35 MB)

- [Classification of political social media](): Social media messages from politicians classified by content. (4 MB)

- **CLiPS Stylometry Investigation (CSI) Corpus**: a yearly expanded corpus of student texts in two genres: essays and reviews. The purpose of this corpus lies primarily in stylometric research, but other applications are possible. (on request)

- **ClueWeb09 FACC**: ClueWeb09 with Freebase annotations (72 GB)

- **ClueWeb11 FACC**: ClueWeb11 with Freebase annotations (92 GB)

- **Common Crawl Corpus**: web crawl data composed of over 5 billion web pages (541 TB)

- **Cornell Movie Dialog Corpus**: contains a large metadata-rich collection of fictional conversations extracted from raw movie scripts: 220,579 conversational exchanges between 10,292 pairs of movie characters, 617 movies (9.5 MB)

- **Corporate messaging**: A data categorization job concerning what corporations actually talk about on social media. Contributors were asked to classify statements as information (objective statements about the company or it's activities), dialog (replies to users, etc.), or action (messages that ask for votes or ask users to click on links, etc.). (600 KB)

- **Crosswikis**: English-phrase-to-associated-Wikipedia-article database. Paper. (11 GB)

- **DBpedia**: a community effort to extract structured information from Wikipedia and to make this information available on the Web (17 GB)

- **Death Row**: last words of every inmate executed since 1984 online (HTML table)

- **Del.icio.us**: 1.25 million bookmarks on delicious.com

- **Disasters on social media**: 10,000 tweets with annotations whether the tweet referred to a disaster event (2 MB).

- **Economic News Article Tone and Relevance**: News articles judged if relevant to the US economy and, if so, what the tone of the article was. Dates range from 1951 to 2014. (12 MB)

- **Enron Email Data**: consists of 1,227,255 emails with 493,384 attachments covering 151 custodians (210 GB)

- **Event Registry**: Free tool that gives real time access to news articles by 100.000 news publishers worldwide. Has API. (query tool)

- Examiner.com - Spam Clickbait News Headlines [Kaggle]: 3 Million crowdsourced News headlines published by now defunct clickbait website The Examiner from 2010 to 2015. (200 MB)

- Federal Contracts from the Federal Procurement Data Center (USASpending.gov): data dump of all federal contracts from the Federal Procurement Data Center found at USASpending.gov (180 GB)

- Flickr Personal Taxonomies: Tree dataset of personal tags (40 MB)

- Freebase Data Dump: data dump of all the current facts and assertions in Freebase (26 GB)

- Freebase Simple Topic Dump: data dump of the basic identifying facts about every topic in Freebase (5 GB)

- Freebase Quad Dump: data dump of all the current facts and assertions in Freebase (35 GB)

- GigaOM Wordpress Challenge [Kaggle]: blog posts, meta data, user likes (1.5 GB)

- Google Books Ngrams: available also in hadoop format on amazon s3 (2.2 TB)

- Google Web 5gram: contains English word n-grams and their observed frequency counts (24 GB)

- Gutenberg Ebook List: annotated list of ebooks (2 MB)

- Hansards text chunks of Canadian Parliament: 1.3 million pairs of aligned text chunks (sentences or smaller fragments) from the official records (Hansards) of the 36th Canadian Parliament. (82 MB)

- Harvard Library: over 12 million bibliographic records for materials held by the Harvard Library, including books, journals, electronic resources, manuscripts, archival materials, scores, audio, video and other materials. (4 GB)

- Hate speech identification: Contributors viewed short text and identified if it a) contained hate speech, b) was offensive but without hate speech, or c) was not offensive at all. Contains nearly 15K rows with three contributor judgments per text string. (3 MB)

- Hillary Clinton Emails [Kaggle]: nearly 7,000 pages of Clinton's heavily redacted emails (12 MB)

- [Home Depot Product Search Relevance \[Kaggle\]](#): contains a number of products and real customer search terms from Home Depot's website. The challenge is to predict a relevance score for the provided combinations of search terms and products. To create the ground truth labels, Home Depot has crowdsourced the search/product pairs to multiple human raters. (65 MB)

- [Identifying key phrases in text](#): Question/Answer pairs + context; context was judged if relevant to question/answer. (8 MB)

- [Jeopardy](#): archive of 216,930 past Jeopardy questions (53 MB)

- [200k English plaintext jokes](#): archive of 208,000 plaintext jokes from various sources.

- [Machine Translation of European Languages](#): (612 MB)

- [Material Safety Datasheets](#): 230,000 Material Safety Data Sheets. (3 GB)

- [Million News Headlines - ABC Australia \[Kaggle\]](#): 1.3 Million News headlines published by ABC News Australia from 2003 to 2017. (56 MB)

- [MCTest](#): a freely available set of 660 stories and associated questions intended for research on the machine comprehension of text; for question answering (1 MB)

- [NEGRA](#): A Syntactically Annotated Corpus of German Newspaper Texts. Available for free for all Universities and non-profit organizations. Need to sign and send form to obtain. (on request)

- [News Headlines of India - Times of India \[Kaggle\]](#): 2.7 Million News Headlines with category published by Times of India from 2001 to 2017. (185 MB)

- [News article / Wikipedia page pairings](#): Contributors read a short article and were asked which of two Wikipedia articles it matched most closely. (6 MB)

- [NIPS2015 Papers (version 2) \[Kaggle\]](#): full text of all NIPS2015 papers (335 MB)

- [NYTimes Facebook Data](#): all the NYTimes facebook posts (5 MB)

- [One Week of Global News Feeds \[Kaggle\]](#): News Event Dataset of 1.4 Million Articles published globally in 20 languages over one week of August 2017. (115 MB)

- **Objective truths of sentences/concept pairs**: Contributors read a sentence with two concepts. For example "a dog is a kind of animal" or "captain can have the same meaning as master." They were then asked if the sentence could be true and ranked it on a 1-5 scale. (700 KB)

- **Open Library Data Dumps**: dump of all revisions of all the records in Open Library. (16 GB)

- **Personae Corpus**: collected for experiments in Authorship Attribution and Personality Prediction. It consists of 145 Dutch-language essays by 145 different students. (on request)

- **Reddit Comments**: every publicly available reddit comment as of july 2015. 1.7 billion comments (250 GB)

- **Reddit Comments (May '15) [Kaggle]**: subset of above dataset (8 GB)

- **Reddit Submission Corpus**: all publicly available Reddit submissions from January 2006 - August 31, 2015). (42 GB)

- **Reuters Corpus**: a large collection of Reuters News stories for use in research and development of natural language processing, information retrieval, and machine learning systems. This corpus, known as "Reuters Corpus, Volume 1" or RCV1, is significantly larger than the older, well-known Reuters-21578 collection heavily used in the text classification community. Need to sign agreement and sent per post to obtain. (2.5 GB)

- **SaudiNewsNet**: 31,030 Arabic newspaper articles alongwith metadata, extracted from various online Saudi newspapers. (2 MB)

- **SMS Spam Collection**: 5,574 English, real and non-enconded SMS messages, tagged according being legitimate (ham) or spam. (200 KB)

- **SouthparkData**: .csv files containing script information including: season, episode, character, & line. (3.6 MB)

- **Stackoverflow**: 7.3 million stackoverflow questions + other stackexchanges (query tool)

- **Twitter Cheng-Caverlee-Lee Scrape**: Tweets from September 2009 - January 2010, geolocated. (400 MB)

- Twitter New England Patriots Deflategate sentiment: Before the 2015 Super Bowl, there was a great deal of chatter around deflated footballs and whether the Patriots cheated. This data set looks at Twitter sentiment on important days during the scandal to gauge public sentiment about the whole ordeal. (2 MB)

- Twitter Progressive issues sentiment analysis: tweets regarding a variety of left-leaning issues like legalization of abortion, feminism, Hillary Clinton, etc. classified if the tweets in question were for, against, or neutral on the issue (with an option for none of the above). (600 KB)

- Twitter Sentiment140: Tweets related to brands/keywords. Website includes papers and research ideas. (77 MB)

- Twitter sentiment analysis: Self-driving cars: contributors read tweets and classified them as very positive, slightly positive, neutral, slightly negative, or very negative. They were also prompted asked to mark if the tweet was not relevant to self-driving cars. (1 MB)

- Twitter Tokyo Geolocated Tweets: 200K tweets from Tokyo. (47 MB)

- Twitter UK Geolocated Tweets: 170K tweets from UK. (47 MB)

- Twitter USA Geolocated Tweets: 200k tweets from the US (45MB)

- Twitter US Airline Sentiment [Kaggle]: A sentiment analysis job about the problems of each major U.S. airline. Twitter data was scraped from February of 2015 and contributors were asked to first classify positive, negative, and neutral tweets, followed by categorizing negative reasons (such as "late flight" or "rude service"). (2.5 MB)

- U.S. economic performance based on news articles: News articles headlines and excerpts ranked as whether relevant to U.S. economy. (5 MB)

- Urban Dictionary Words and Definitions [Kaggle]: Cleaned CSV corpus of 2.6 Million of all Urban Dictionary words, definitions, authors, votes as of May 2016. (238 MB)

- Wesbury Lab Usenet Corpus: anonymized compilation of postings from 47,860 English-language newsgroups from 2005-2010 (40 GB)

- Wesbury Lab Wikipedia Corpus Snapshot of all the articles in the English part of the Wikipedia that was taken in April 2010. It was processed, as described in

detail below, to remove all links and irrelevant material (navigation text, etc) The corpus is untagged, raw text. Used by Stanford NLP (1.8 GB).

- Wikipedia Extraction (WEX): a processed dump of english language wikipedia (66 GB)

- Wikipedia XML Data: complete copy of all Wikimedia wikis, in the form of wikitext source and metadata embedded in XML. (500 GB)

- Yahoo! Answers Comprehensive Questions and Answers: Yahoo! Answers corpus as of 10/25/2007. Contains 4,483,032 questions and their answers. (3.6 GB)

- Yahoo! Answers consisting of questions asked in French: Subset of the Yahoo! Answers corpus from 2006 to 2015 consisting of 1.7 million questions posed in French, and their corresponding answers. (3.8 GB)

- Yahoo! Answers Manner Questions: subset of the Yahoo! Answers corpus from a 10/25/2007 dump, selected for their linguistic properties. Contains 142,627 questions and their answers. (104 MB)

- Yahoo! HTML Forms Extracted from Publicly Available Webpages: contains a small sample of pages that contain complex HTML forms, contains 2.67 million complex forms. (50+ GB)

- Yahoo! Metadata Extracted from Publicly Available Web Pages: 100 million triples of RDF data (2 GB)

- Yahoo N-Gram Representations: This dataset contains n-gram representations. The data may serve as a testbed for query rewriting task, a common problem in IR research as well as to word and sentence similarity task, which is common in NLP research. (2.6 GB)

- Yahoo! N-Grams, version 2.0: n-grams (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 (12 GB)

- Yahoo! Search Logs with Relevance Judgments: Annonymized Yahoo! Search Logs with Relevance Judgments (1.3 GB)

- Yahoo! Semantically Annotated Snapshot of the English Wikipedia: English Wikipedia dated from 2006-11-04 processed with a number of publicly-available NLP tools. 1,490,688 entries. (6 GB)

- Yelp: including restaurant rankings and 2.2M reviews (on request)

- Youtube: 1.7 million youtube videos descriptions (torrent)

Sources

- Awesome public datasets/NLP (includes more lists)
- AWS Public Datasets
- CrowdFlower: Data for Everyone (lots of little surveys they conducted and data obtained by crowdsourcing for a specific task)
- Kaggle 1, 2 (make sure though that the kaggle competition data can be used outside of the competition!)
- Open Library
- Quora (mainly annotated corpora)
- /r/datasets (endless list of datasets, most is scraped by amateurs though and not properly documented or licensed)
- rs.io (another big list)
- Stackexchange: Opendata
- Stanford NLP group (mainly annotated corpora and TreeBanks or actual NLP tools)
- Yahoo! Webscope (also includes papers that use the data that is provided)

https://github.com/niderhoff/nlp-datasets/blob/master/README.md