lionbridge.ai

The Best 25 Datasets for Natural Language Processing

Article by Meiryum Ali | July 09, 2019

7-8 minutos

Natural language processing is a massive field of research. With so many areas to explore, it can sometimes be difficult to know where to begin – let alone start searching for data.

With this in mind, we've combed the web to create the ultimate collection of free online datasets for NLP. Although it's impossible to cover every field of interest, we've done our best to compile datasets for a broad range of NLP research areas, from sentiment analysis to audio and voice recognition projects. Use it as a starting point for your experiments, or check out our specialized collections of datasets if you already have a project in mind.

Datasets for Sentiment Analysis

Where can I download datasets for sentiment analysis?

Machine learning models for sentiment analysis need to be trained with large, specialized datasets. The following list should hint at some of the ways that you can improve your sentiment analysis algorithm.

Multidomain Sentiment Analysis Dataset: This is a slightly older dataset that features a variety of product reviews taken from Amazon.

IMDB Reviews: Featuring 25,000 movie reviews, this relatively small dataset was compiled primarily for binary sentiment classification use cases.

Stanford Sentiment Treebank: Also built from movie reviews, Stanford's dataset was designed to train a model to identify sentiment in longer phrases. It contains over 10,000 snippets taken from Rotten Tomatoes.

Sentiment140: This popular dataset contains 160,000 tweets formatted with 6 fields: polarity, ID, tweet date, query, user, and the text. Emoticons have been pre-removed.

Twitter US Airline Sentiment: Scraped in February 2015, these tweets about US airlines are classified as classified as positive, negative, and neutral. Negative tweets have also been categorized by reason for complaint.

Datasets for Text

Where can I download text datasets for natural language processing?

Natural language processing is a massive field of research, but the following list includes a broad range of datasets for different natural language processing tasks, such as voice recognition and chatbots.

20 Newsgroups: This collection of approximately 20,000 documents covers 20 different newsgroups, from baseball to religion.

Reuters News Dataset: The documents in this dataset appeared on Reuters in 1987. They have since been assembled and indexed for use in machine learning.

The WikiQA Corpus: This corpus is a publicly-available collection of question and answer pairs. It was originally assembled for use in research on open-domain question answering.

UCI's Spambase: Originally created by a team at Hewlett-Packard, this large spam email dataset is useful for developing personalized spam filters.

Yelp Reviews: This open dataset released by Yelp contains more than 5 million reviews.

WordNet: Compiled by researchers at Princeton University, WordNet is essentially a large lexical database of English 'synsets', or groups of synonyms that each describe a different, distinct concept.

Audio Speech Datasets for Natural Language Processing

Where can I download audio datasets for natural language processing?

Audio speech datasets are useful for training natural language processing applications such as virtual assistants, in-car navigation, and any other sound-activated systems.

2000 HUB5 English: This dataset contains transcripts derived from 40 telephone conversations in English. The corresponding speech files are also available through this page.

LibriSpeech: This corpus contains roughly 1,000 hours of English speech, comprised of audiobooks read by multiple speakers. The data is organized by chapters of each book.

Spoken Wikipedia Corpora: Containing hundreds of hours of audio, this corpus is composed of spoken articles from Wikipedia in English, German, and Dutch. Due to the nature of the project, it also contains a diverse set of readers and topics.

Free Spoken Digit Dataset: This is a collection of 1,500 recordings of spoken digits in English.

TIMIT: This data is designed for research in acoustic-phonetic studies and the development of automatic speech recognition systems. It contains recordings of 630 speakers of American English reading ten 'phonetically rich' sentences.

Datasets for Natural Language Processing (General)

Where can I download open datasets for natural language processing?

Still can't find what you need? Here are a few more datasets for natural language processing tasks.

Enron Dataset: Containing roughly 500,000 messages from the senior management of Enron, this dataset was made as a resource for those looking to improve or understand current email tools.

Amazon Reviews: This dataset contains around 35 million reviews from Amazon spanning a period of 18 years. It includes product and user information, ratings, and the plaintext review.

Google Books Ngrams: A Google Books corpora of n-grams, or ‘fixed size tuples of items’, can be found at this link. The ‘n’ in ‘n-grams’ specifies the number of words or characters in that specific tuple.

Blogger Corpus: Gathered from blogger.com, this collection of 681,288 blog posts contains over 140 million words. Each blog included here contains at least 200 occurrences of common English words.

Wikipedia Links Data: Containing approximately 13 million documents, this dataset by Google consists of web pages that contain at least one hyperlink pointing to English Wikipedia. Each Wikipedia page is treated as an entity, while the anchor text of the link represents a mention of that entity.

Gutenberg eBooks List: This annotated list of ebooks from Project Gutenberg contains basic information about each eBook, organized by year.

Hansards Text Chunks of Canadian Parliament: This corpus contains 1.3 million pairs of aligned text chunks from the records of the 36th Canadian Parliament.

Jeopardy: The archive linked here contains more than 200,000 questions and answers from the quiz show Jeopardy. Each data point also contains a range of other information, including the category of the question, show number, and air date.

SMS Spam Collection in English: This dataset consists of 5,574 English SMS messages that have been tagged as either legitimate or spam. 425 of the texts are spam messages that were manually extracted from the Grumbletext website.

Still can’t find what you need? Lionbridge AI creates and annotates customized datasets for a wide variety of NLP projects, including everything from chatbot variations to entity annotation. With over 20 years of experience in managing a crowd of over 500,000+ linguistic specialists, Lionbridge AI is perfectly placed to provide your model with a solid foundation. Contact us to find out how custom data can take your machine-learning project to the next level.