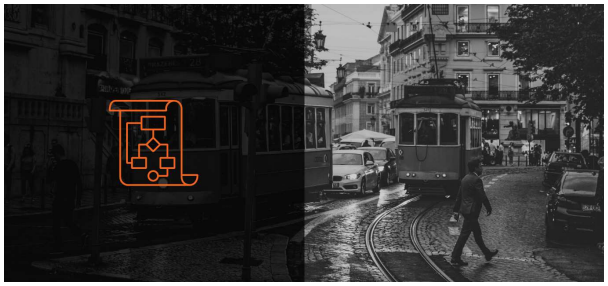
lionbridge.ai

12 Best Portuguese Language Datasets for Machine Learning

Article by Alex Nguyen | August 19, 2019

3-4 minutos



To build any multilingual natural language processing (NLP) application, a large base of high-quality training data is key. Although there are many sources for annotated datasets available on the web, most of them are only available in English.

To help, we at Lionbridge have compiled a list of high-quality Portuguese datasets that covers a wide spectrum of AI use cases, from speech recognition to machine translation. Let's jump right in!

Portuguese Text Datasets

O Corpus do Português: This corpus contains about one billion words of lexical, semantic, and syntactic text data in Portuguese.

NILC / San Carlos Corpora: This website features a collection of contemporary Portuguese corpora. All text is POS-tagged.

CIPM Medieval Portuguese: This dataset contains a large collection of medieval Portuguese text.

NER in Brazilian Portuguese tweets: A dataset for named entity recognition for tweets in Brazilian Portuguese. The dataset contains about 3.6k training and 900 test samples.

CRPC Comparative Portuguese corpus: A large collection of corpora containing texts from several varieties of Portuguese (European, Brazil, Angola, Cape Verde and more).

Brazilian Portuguese Literature Corpus: Featuring 81 distinct works, this dataset contains over 3.7 million words of Brazilian literature written between 1840 and 1908.

Portuguese Tweets for Sentiment Analysis: For sentiment polarity classification, this dataset contains 800k tweets in Portuguese divided into positive, negative, and neutral classes.

BlogSet-BR: An extensive Brazilian Portuguese corpus containing 2.1 billion words extracted from 7.4 million posts over 808 thousand different Brazilian blogs.

Portuguese Parallel Text Datasets

Spanish-Portuguese website parallel corpus: From the EU Open Data Portal, this is a parallel corpus of bilingual texts in Spanish and Portuguese crawled from the web.

European Parliament Proceedings Parallel Corpus: Extracted from the proceedings of the European Parliament from 1996-2011, this

corpus includes parallel text for statistical machine translation systems in 21 European languages, including Portuguese.

Portuguese Audio Datasets

NURC-RJ Spoken Portuguese Corpus: This corpus includes over 350 hours of recorded interviews in Brazilian Portuguese.

Brazilian Portuguese Phonemes: Created for voice transcription use cases, this is a small dataset containing recorded voice samples from a Brazilian Portuguese speaker.

Still can't find what you need? Lionbridge AI provides custom multilingual datasets for over 300 languages and dialects. Whether you need hundreds or millions of data points, our 500,000+ certified contributors can ensure your algorithm has a solid ground truth.

Interested? Get high-quality data now