Title: Spark NLP

URL: https://en.wikipedia.org/wiki/Spark_NLP

PageID: 60360004

Categories: Category:2017 software, Category:Free software programmed in Python, Category:Free software programmed in Scala, Category:Natural language processing toolkits, Category:Open-source artificial intelligence, Category:Software using the Apache license

Source: Wikipedia (CC BY-SA 4.0). Content may require attribution.

Spark NLP is an open-source text processing library for advanced natural language processing for the Python , Java and Scala programming languages. The library is built on top of Apache Spark and its Spark ML library.

Its purpose is to provide an API for natural language processing pipelines that implement recent academic research results as production-grade, scalable, and trainable software. The library offers pre-trained neural network models, pipelines, and embeddings, as well as support for training custom models.

Features

The design of the library makes use of the concept of a pipeline which is an ordered set of text annotators. Out of the box annotators include, tokenizer , normalizer, stemming , lemmatizer , regular expression , TextMatcher, chunker , DateMatcher, SentenceDetector, DeepSentenceDetector, POS tagger , ViveknSentimentDetector, sentiment analysis , named entity recognition , conditional random field annotator, deep learning annotator, spell checking and correction, dependency parser, typed dependency parser, document classification , and language detection .

The Models Hub is a platform for sharing open-source as well as licensed pre-trained models and pipelines. It includes pre-trained pipelines with tokenization, lemmatization, part-of-speech tagging, and named entity recognition that exist for more than thirteen languages; word embeddings including GloVe , ELMo , BERT , ALBERT , XLNet , Small BERT, and ELECTRA; sentence embeddings including Universal Sentence Embeddings (USE) and Language Agnostic BERT Sentence Embeddings (LaBSE). It also includes resources and pre-trained models for more than two hundred languages. Spark NLP base code includes support for East Asian languages such as tokenizers for Chinese, Japanese, Korean; for right-to-left languages such as Urdu, Farsi, Arabic, Hebrew and pre-trained multilingual word and sentence embeddings such as LaUSE and a translation annotator.

Usage in healthcare

Spark NLP for Healthcare is a commercial extension of Spark NLP for clinical and biomedical text mining. It provides healthcare-specific annotators, pipelines, models, and embeddings for clinical entity recognition, clinical entity linking, entity normalization, assertion status detection, de-identification, relation extraction, and spell checking and correction.

The library offers access to several clinical and biomedical transformers: JSL-BERT-Clinical, BioBERT, ClinicalBERT, GloVe-Med, GloVe-ICD-O. It also includes over 50 pre-trained healthcare models, that can recognize the entities such as clinical, drugs, risk factors, anatomy, demographics, and sensitive data.

Spark OCR

Spark OCR is another commercial extension of Spark NLP for optical character recognition (OCR) from images, scanned PDF documents, and DICOM files. It is a software library built on top of Apache Spark . It provides several image pre-processing features for improving text recognition results such as adaptive thresholding and denoising , skew detection & correction, adaptive scaling, layout analysis and region detection, image cropping , removing background objects.

Due to the tight coupling between Spark OCR and Spark NLP, users can combine NLP and OCR pipelines for tasks such as extracting text from images, extracting data from tables, recognizing and highlighting named entities in PDF documents or masking sensitive text in order to de-identify images.

Several output formats are supported by Spark OCR such as PDF , images, or DICOM files with annotated or masked entities, digital text for downstream processing in Spark NLP or other libraries, structured data formats (JSON and CSV), as files or Spark data frames.

Users can also distribute the OCR jobs across multiple nodes in a Spark cluster .

License and availability

Spark NLP is licensed under the Apache 2.0 license. The source code is publicly available on GitHub as well as documentation and a tutorial. Prebuilt versions of Spark NLP are available in PyPi and Anaconda Repository for Python development, in Maven Central for Java & Scala development, and in Spark Packages for Spark development.

Award

In March 2019, Spark NLP received Open Source Award for its contributions in natural language processing in Python, Java, and Scala.

References

Sources

Thomas, Alex (21 July 2020). Natural Language Processing with Spark NLP: Learning to Understand Text at Scale . O'Reilly Media. ISBN 978-1492047766 .

Quinto, Butch (2020). Next-Generation Machine Learning with Spark . Berkeley, California: Apress. doi : 10.1007/978-1-4842-5669-5 . ISBN 978-1-4842-5668-8 . S2CID 211234215 .

External links

Spark NLP

Teams■■