Conversation with Kaggle Recommendation System

User: Geben Sie mir die Datensätze zur Verarbeitung natürlicher Sprache

Kaggle Recommender Engine: 1. Dataset Name: rtatman/corpus-of-bilingual-childrens-speech Title: Corpus of bilingual children's speech Description: Corpus of bilingual children's speech by Rachael Tatman. Transcribed natural speech from 25 bilingual children Author: Rachael Tatman Link: https://www.kaggle.com/datasets/rtatman/corpus-of-bilingual-childrens-speech Last updated: 21/07/2017 21:59 Size: 269KB Usability rating: 0.82 View count: 19973 Licence: CC BY-NC-SA 4.0 license Tags: languages, education, social science, linguistics 2. Dataset Name: toponowicz/spoken-language-identification Title: Spoken Language Identification Description: Spoken Language Identification by Tomasz. Speech samples of English, German and Spanish languages. Author: Tomasz Link: https://www.kaggle.com/datasets/toponowicz/spoken-language-identification Last updated: 05/07/2018 16:32 Size: 15GB Usability rating: 0.69 View count: 34071 Licence: CC0: Public Domain license Tags: languages, arts and entertainment, music 3. Dataset Name: toygarr/datasets-for-natural-language-processing Title: datasets for natural language processing Description: datasets for natural language processing by Toygar Tanyel. 6 processed and in different context public datasets for nlp tasks Author: Toygar Tanyel Link: https://www.kaggle.com/datasets/toygarr/datasets-for-natural-language-processing Last updated: 04/01/2022 14:07 Size: 153MB Usability rating: 0.81 View count: 8308 Licence: Unknown license Tags: sampling, earth and nature, social science, nlp, classification, text 4. Dataset Name: abhishek/10k-german-news-articles Title: 10k German News Articles Description: 10k German News Articles by Abhishek Thakur. A dataset for topic extraction and NLP for German language Author: Abhishek Thakur Link: https://www.kaggle.com/datasets/abhishek/10k-german-news-articles Last updated: 07/11/2019 08:50 Size: 123MB Usability rating: 0.82 View count: 16653 Licence: CC BY-NC-SA 4.0 license Tags: earth and nature, social science, nlp, classification, news 5. Dataset Name: mediatechlab/gneutralspeech Title: g Neutral Speech Male Description: g Neutral Speech Male by Pedro Lopes. Globo's audio and text dataset for male single speaker in portuguese. Author: Pedro Lopes Link: https://www.kaggle.com/datasets/mediatechlab/gneutralspeech Last updated: 22/09/2022 17:43 Size: 5GB Usability rating: 0.71 View count: 3641 Licence: Other (specified in description) license Tags: brazil, internet, signal processing, deep learning, text, audio 6. Dataset Name: samdeeplearning/deepnlp Title: Deep-NLP Description: Deep-NLP by samdeeplearning. natural language processing Author: samdeeplearning Link: https://www.kaggle.com/datasets/samdeeplearning/deepnlp Last updated: 01/03/2017 05:43 Size: 234KB Usability rating: 0.82 View count: 108954 Licence: Other (specified in description) license Tags: languages, earth and nature, education, linguistics, psychology 7. Dataset Name: averkij/wals-dataset Title: WALS dataset Description: WALS dataset by Sergei Averkiev. World languages with features and coordinates Author: Sergei Averkiev Link: https://www.kaggle.com/datasets/averkij/wals-dataset Last updated: 17/01/2020 12:30 Size: 297KB Usability rating: 0.97 View count: 3932 Licence: Attribution 4.0 International (CC BY 4.0) license Tags: languages, geography, linguistics, geospatial analysis 8. Dataset Name: franciscopuentes/castilian-language-corpus-from-wikipedia Title: Castilian language corpus from Wikipedia. Description: Castilian language corpus from Wikipedia. by Francisco Puentes. SQlite database with pages, words, biwords and its relations. Author: Francisco Puentes Link: https://www.kaggle.com/datasets/franciscopuentes/castilian-language-corpus-from-wikipedia Last updated: 26/10/2021 11:14 Size: 2GB Usability rating: 0.81 View count: 1184 Licence: CC BY-SA 4.0 license Tags: europe, linguistics, bigquery 9. Dataset Name: rtatman/120-million-word-spanish-corpus Title: 120 Million Word Spanish Corpus Description: 120 Million Word Spanish Corpus by Rachael Tatman. The Spanish Language portion of the Wikicorpus (v 1.0) Author: Rachael Tatman Link: https://www.kaggle.com/datasets/rtatman/120-million-word-spanish-corpus Last updated: 08/08/2017 20:43 Size: 489MB Usability rating: 0.62 View count: 31165 Licence: CC BY-SA 3.0 license Tags: languages, europe, south america, linguistics 10. Dataset Name: rtatman/world-atlas-of-language-structures Title: World Atlas of Language Structures Description:

World Atlas of Language Structures by Rachael Tatman. Information on the linguistic structures in 2,679 languages Author: Rachael Tatman Link: https://www.kaggle.com/datasets/rtatman/world-atlas-of-language-structures Last updated: 07/09/2017 23:40 Size: 2MB Usability rating: 0.82 View count: 31931 Licence: Other (specified in description) license Tags: languages, social science, linguistics 11. Dataset Name: stanfordu/stanford-natural-language-inference-corpus Title: Stanford Natural Language Inference Corpus Description: Stanford Natural Language Inference Corpus by Sohier Dane. A collection of 570k labeled human-written English sentence pairs Author: Sohier Dane Link: https://www.kaggle.com/datasets/stanfordu/stanford-natural-language-inference-corpus Last updated: 21/07/2017 17:35 Size: 44MB Usability rating: 0.82 View count: 26929 Licence: CC BY-SA 4.0 license Tags: languages, education, linguistics 12. Dataset Name: dhruvildave/en-fr-translation-dataset Title: English-French Translation Dataset Description: English-French Translation Dataset by Dhruvil Dave. 22.5 Million translations of English and French Author: Dhruvil Dave Link: https://www.kaggle.com/datasets/dhruvildave/en-fr-translation-dataset Last updated: 09/02/2021 19:27 Size: 3GB Usability rating: 1.00 View count: 30099 Licence: Database: Open Database, Contents: © Original Authors license Tags: nlp, nltk 13. Dataset Name: ramakrishnan1984/785-million-language-translation-database-ai-ml Title: 785 Million Language Translation Database for AI Description: 785 Million Language Translation Database for AI by Ramakrishnan Lakshmanan. Translation Dataset with 785 million records spanning across 548 languages Author: Ramakrishnan Lakshmanan Link: https://www.kaggle.com/datasets/ramakrishnan1984/785-million-language-translation-database-ai-ml Last updated: 28/08/2023 10:22 Size: 6GB Usability rating: 0.81 View count: 1417 Licence: GNU Lesser General Public License 3.0 license Tags: languages, tabular, text, bigquery 14. Dataset Name: team-ai/japaneseenglish-bilingual-corpus Title: Japanese-English Bilingual Corpus Description: Japanese-English Bilingual Corpus by VenaliSonone. Japanese-English Bilingual Corpus of Wikipedia's Kyoto Articles Author: VenaliSonone Link: https://www.kaggle.com/datasets/team-ai/japaneseenglish-bilingual-corpus Last updated: 15/09/2017 15:30 Size: 159MB Usability rating: 0.71 View count: 29072 Licence: Other (specified in description) license Tags: languages, linguistics, psychology, artificial intelligence 15. Dataset Name: sudalairajkumar/tamil-nlp Title: Tamil NLP Description: Tamil NLP by SRK. Datasets for Natural Language Processing in Tamil Author: SRK Link: https://www.kaggle.com/datasets/sudalairajkumar/tamil-nlp Last updated: 11/03/2019 06:29 Size: 3MB Usability rating: 1.00 View count: 38795 Licence: CC BY-SA 4.0 license Tags: languages, literature, nlp, text I hope this was a helpful response. Now you can talk with the recommended data.

User: Geben Sie mir die Datensätze zur Verarbeitung natürlicher Sprache basierend auf dem Text.

Kaggle Recommender Engine: Ich habe einige Datensätze ausgewählt, die speziell für Natural Language Processing (NLP) mit Schwerpunkt auf Textdaten geeignet sind: 1. Dataset Name: tomoni-hirano/fake-news-detection-by-nlp Titel: Fake News Detection Beschreibung: Dieser Datensatz enthält 1 Mio. Nachrichten zum Training von Modellen zur Erkennung von Falschnachrichten. Autor: Tomoni Hirano Link: https://www.kaggle.com/datasets/tomoni-hirano/fake-news-detection-by-nlp Letzte Aktualisierung: 08.07.2019 Lizenz: Unbekannt Stichworte: NLP, Text, Nachrichten 2. Dataset Name: yiweiliu/1980-2019-senate-speech Titel: Senate Speech 1980-2019 Beschreibung: Ein Datensatz von Reden im US-Senat von 1980 bis 2019. Es gibt ungefähr 1,2 Millionen Reden in diesem Datensatz. Autor: Yiwei Liu Link: https://www.kaggle.com/datasets/yiweiliu/1980-2019-senate-speech Letzte Aktualisierung: 29.02.2020 Lizenz: Unbekannt Stichworte: NLP, Text, Politik, Geschichte 3. Dataset Name: kazanova/sentiment140 Titel: Sentiment140 Beschreibung: Ein Datensatz mit 1,6 Millionen Tweets, die positiv oder negativ eingestuft sind. Gut geeignet für sentiment-Analysen. Autor: Kazanova Link: https://www.kaggle.com/datasets/kazanova/sentiment140 Letzte Aktualisierung: 15.07.2017 Lizenz: Datenbank: Open Database, Inhalte: © Original Authors Stichworte: NLP, Text, Soziale Medien, Stimmungsanalyse 4. Dataset Name: expressvpn/censored-in-10-countries-news-dataset-in-english Titel: Censored News Dataset Beschreibung: Sammlung von aktuellen Nachrichten-Feeds von

Organisationen, die in mindestens einem von zehn ausgewählten Ländern zensiert sind. Autor: ExpressVPN Link: https://www.kaggle.com/datasets/expressvpn/censored-in-10-countries-news-dataset-in-english Letzte Aktualisierung: 02.07.2021 Lizenz: CC0: Public Domain Stichworte: NLP, Text, Nachrichten, Zensur 5. Dataset Name: rtatman/5day-data-challenge-twitter-us-airline-sentiment Titel: US Airline Sentiment Beschreibung: Ein Datensatz, der aus einer Sammlung von Tweets besteht, die sich auf US-Fluggesellschaften beziehen, mit Sentiment-Labels. Autor: Rachael Tatman Link: https://www.kaggle.com/datasets/rtatman/5day-data-challenge-twitter-us-airline-sentiment Letzte Aktualisierung: 28.07.2017 Lizenz: CC0: Public Domain Stichworte: NLP, Text, Soziale Medien, Stimmungsanalyse Dies sind nur einige Beispiele für die vielen Textdatensätze, die für NLP-Projekte geeignet sind. In der Praxis hängt die Auswahl des idealen Datensatzes von den spezifischen Anforderungen Ihres Projekts ab.