350 NLP Projects

with Code

The Most Powerful NLP-Weapon Arsenal

Himanshu Ramchandani
M.Tech | Data Science

NLP Migrant Workers' Paradise: Almost the most complete Chinese NLP resource library

In the process of getting started and getting familiar with NLP, I used a lot of packages on github, so I sorted it out and shared it here.

Many bags are very interesting and worth collecting, satisfying everyone's collection addiction! If you find it useful, please share and star⭐,thanks!

Long-term irregular updates, welcome to watch and fork!❤️❤️❤️

🍆 🍒 🍐 🍊	🌻 🍓 🍈 🍅 🍍
* Corpus	* Document Processing
* Thesaurus and lexical tools	* Table Processing
* Pre-trained language model	* Text Matching
* Extraction	* Text Data Enhancement
* Knowledge map	* Text Retrieval
* Text generation	* Reading Comprehension
* Text summarization	* Sentiment Analysis
* Intelligent question answering	* Common Regular Expressions
* Text error correction	* Speech Processing

* Common regular expressions

* Event extraction

* Machine translation

* Digital transformation

* Anaphora resolution

* Text clustering

* Text classification

* Knowledge reasoning

* Explainable NLP

* Text adversarial attack

* Text visualization

* Text annotation tool

* Comprehensive tool

* Funny and funny tool

* Course report interview, etc.

* Competition

* Financial NLP

* Medical NLP

* Legal NLP

* Text generation image

* Others

---

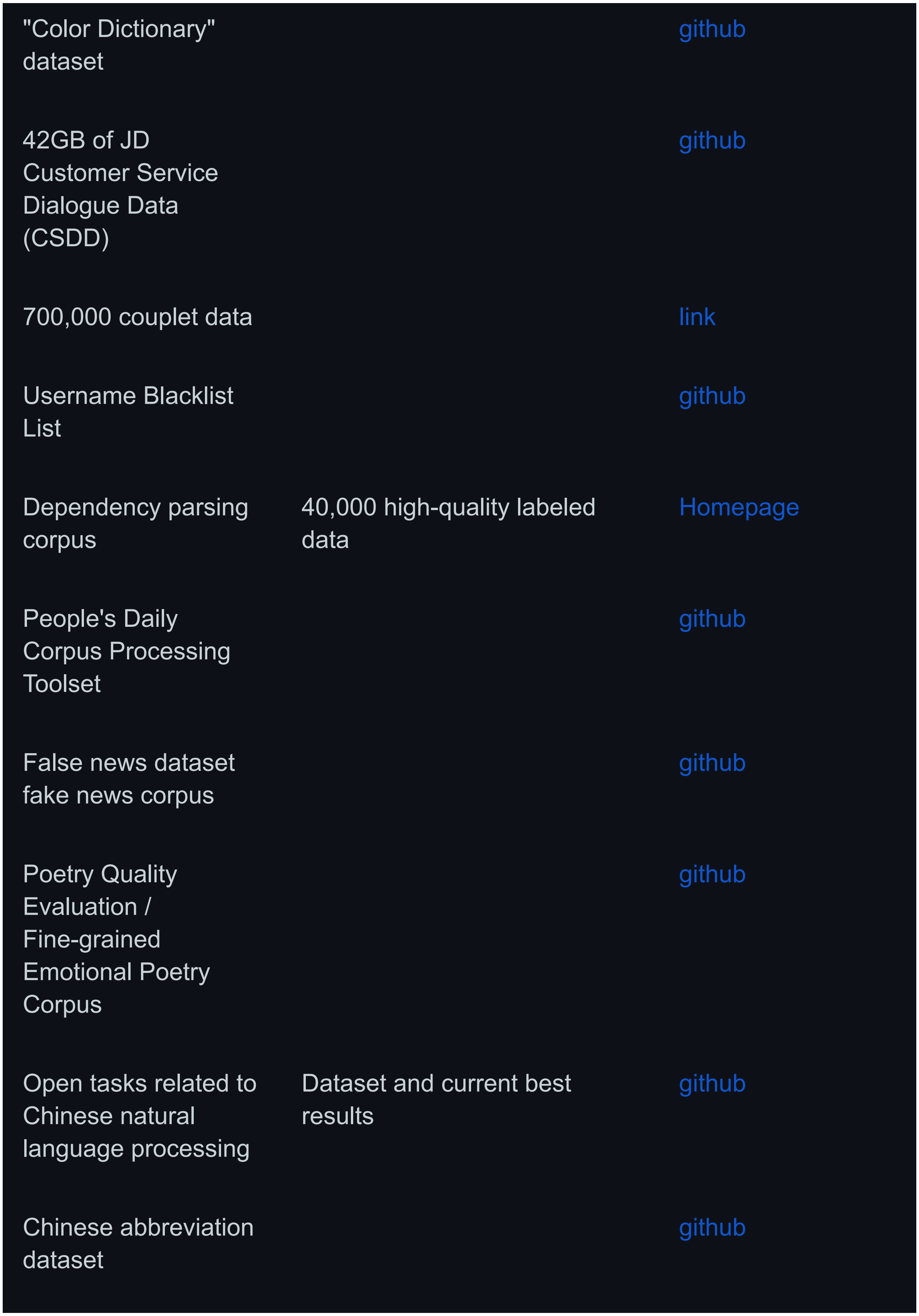
corpus

Resource name (Name)	Description	Link
Corpus of names		wainshine/Chinese-Names-Corpus
Chinese-Word-Vectors	Various Chinese word vectors	github repo

Chinese Chat Corpus	The library includes Douban Duolun, PTT gossip corpus, Qingyun corpus, TV drama dialogue corpus, Tieba forum reply corpus, Weibo corpus, little yellow chicken corpus	link
Chinese rumor data	In this data file, each line contains a rumor data in json format	github
Chinese Question Answering Dataset		link extract code 2dva
WeChat official account corpus	The 3G corpus, which includes some articles from WeChat official accounts captured from the web, has removed HTML and only contains plain text. One article per line, in JSON format, name is the name of the WeChat official account, account is the ID of the WeChat official account, title is the title, and content is the text	github
Chinese natural language processing corpus, data set		github

Task-based dialogue English dataset	[The most complete task-based dialogue data set] mainly introduces a complete task-based dialogue data set, which covers the main information of all commonly used data sets in the field of task-based dialogue. In addition, in order to help researchers better grasp the context of field progress, we present the State-of-the-art experimental results on several datasets in the form of Leaderboard.	github
Speech Recognition Corpus Generation Tool	Create an Automatic Speech Recognition (ASR) corpus from online videos with audio/subtitles	github
LitBank NLP dataset	A corpus of 100 labeled English novels supporting natural language processing and computational humanities tasks	github
ChineseULMFiT	Sentiment Analysis Text Classification Corpus and Model	github
The administrative division data of provinces, municipalities and towns are marked with pinyin		github

Automated Summarization Corpus of Education Industry News		github
Chinese Natural Language Processing Dataset		github
Baidu Zhizhi Q&A Corpus	More than 5.8 million questions, 9.38 million answers, 5800 classification labels. Based on the question and answer corpus, it can support a variety of applications, such as chat question and answer, logic mining	github
Wikipedia Massively Parallel Text Corpus	85 languages, 1620 language pairs, 135M contrasting sentences	github
Ancient Poetry Thesaurus		github repo more complete ancient poetry lexicon
Low memory loading Wikipedia data	Use the new version of nlp library to load 17GB+ English Wiki corpus and only occupy 9MB of memory Traversal speed 2-3 Gbit/s	github
couplet data	700,000 couplets, more than 700,000 couplets	github



"Color Dictionary" dataset		github
42GB of JD Customer Service Dialogue Data (CSDD)		github
700,000 couplet data		link
Username Blacklist List		github
Dependency parsing corpus	40,000 high-quality labeled data	Homepage
People's Daily Corpus Processing Toolset		github
False news dataset fake news corpus		github
Poetry Quality Evaluation / Fine-grained Emotional Poetry Corpus		github
Open tasks related to Chinese natural language processing	Dataset and current best results	github
Chinese abbreviation dataset		github

Chinese task benchmarking	Representative dataset - benchmark (pretrained) model - corpus - baseline - toolkit - leaderboard	github
Chinese Rumor Database		github
CLUEDatasetSearch	Chinese and English NLP datasets Search all Chinese NLP datasets, with commonly used English NLP datasets attached	github
Multi-Document Summarization Dataset		github
Make Everyone "Courteous" Polite Migration Quest	Transform impolite sentences into polite ones while preserving meaning, providing a dataset with 139M+ instances	paper and code
Cantonese/English Conversational Bilingual Corpus		github
List of Chinese NLP datasets		github
Nomenclature recognition data set of person-like names/place		github

names/organization names		
Chinese Language Comprehension Benchmark	Includes representative datasets & benchmark models & corpora & leaderboards	github
OpenCLaP multi-domain open source Chinese pre-trained language model warehouse	Civil documents, criminal documents, Baidu Encyclopedia	github
Chinese full word coverage BERT and two reading comprehension data	DRCD dataset: Released by Delta Research Institute of Taiwan, China, it has the same form as SQuAD, and is an extractive reading comprehension dataset based on traditional Chinese. CMRC 2018 dataset: Chinese machine reading comprehension data released by the Xunfei Joint Laboratory of Harbin Institute of Technology. According to a given question, the system needs to extract fragments from the text as answers, in the same form as SQuAD.	github
Dakshina dataset	Latin/native script parallel dataset for twelve South Asian languages	github

OPUS-100	Multilingual (100 kinds) parallel corpus centered on English	github
Chinese Reading Comprehension Dataset		github
Chinese natural language processing vector collection		github
Chinese Language Comprehension Benchmark	Includes representative datasets, benchmark (pretrained) models, corpora, leaderboards	github
Large list of NLP datasets/benchmark tasks		github
LitBank NLP dataset	A corpus of 100 labeled English novels supporting natural language processing and computational humanities tasks	github
700,000 couplet data		github
Parallel Corpus of Classical Chinese (Ancient Chinese)-Modern Chinese	The short chapters include "The Analects of Confucius", "Mencius", "Zuo Zhuan" and other short ancient books, which have been merged with "Zi Zhi Tong Jian"	github

COLDDateset, Chinese Offensive Language Detection Dataset	Covers topics such as race, gender, and region, and the data will be released after the paper is published	paper

Thesaurus and Lexical Tools

Resource name (Name)	Description	Link
textfilter	Sensitive word filtering in Chinese and English	observerss/textfilter
Name extraction function	Chinese (modern, ancient) names, Japanese names, Chinese surnames and first names, titles (big aunt, little aunt, etc.), English -> Chinese name (John Lee), idiom dictionary	cocoNLP
Chinese Abbreviation Library	National People's Congress: National People's Congress; China: People's Republic of China; Women's Tennis: Women/n Tennis/n Game/vn	github

Chinese Dictionaries	How to dismantle Chinese characters (1) How to dismantle (2) How to dismantle (3)	kfcd/chaizi
Lexical Sentiment Value	Mountain spring water: 0.400704566541 Sufficient : 0.37006739587	rainarch/SentiBridge
Chinese thesaurus, stop words, sensitive words		dongxiexidian/Chinese
python-pinyin	Convert Chinese characters to Pinyin	mozillazg/python-pinyin
zhtools	Conversion between Traditional and Simplified Chinese	skydark/nstools
English simulation Chinese pronunciation engine	say wo i ni #say: I love you	tinyfool/ChineseWithEnglish
chinese_dictionary	Thesaurus, antonym, negative thesaurus	guotong1988/chinese_dictionary

wordninja	English string segmentation and word extraction without spaces	wordninja
Vocabulary related to automobile brand and automobile parts		data
Thesaurus organized by THU	IT thesaurus, financial thesaurus, idiom thesaurus, place names, historical celebrity thesaurus, poetry thesaurus, medical thesaurus, diet thesaurus, legal thesaurus, automobile thesaurus, animal thesaurus	link

Crime Legal Terms and Classification Model	Contains 856 crime knowledge graphs, crime prediction based on 2.8 million crime training database, 13 types of question classification and legal information question and answer function based on 20W legal question and answer pairs	github
Word segmentation corpus + code		Baidu network disk link - extraction code pea6
Chinese word segmentation + part-of-speech tagging based on Bi-LSTM + CRF	keras implementation	link
Chinese word segmentation and part-of-speech tagging based on Universal Transformer + CRF		link
Fast Neural Network Word Segmentation Package	java version	

chinese-xinhua	Zhonghua Xinhua dictionary database and api, including commonly used Xiehouyu, idioms, words and Chinese characters	github
SpaCy Chinese model	Contains Parser, NER, syntax tree and other functions. Some English packages use spacy's English model. If you want to adapt to Chinese, you may need to use spacy's Chinese model.	github
Chinese character data		github
Synonyms Chinese Synonym Toolkit		github
Harvest Text	Domain adaptive text mining tools (new word discovery-sentiment analysis-entity linking, etc.)	github

word2word	Easy-to-use multilingual word-word pair set 62 languages/3,564 multilingual pairs	github
Polyphone dictionary data and codes		github
Chinese characters, words, idioms query interface		github
103976 English vocabulary packs	(sql version, csv version, Excel version)	github
Big list of swear words in English		github
word pinyin data		github
Number calling library in 186 languages		github
Large-scale name database of countries around the world		github

Chinese character feature extractor (featurizer)	Extract the features of Chinese characters (pronunciation features, font features) for deep learning features	github
char_featurizer - Chinese character feature extraction tool		github
Python interface library of mecab, the CJK word segmentation library		github
g2pC context-based Chinese pronunciation automatic marking module		github
ssc, Sound Shape Code	Phonetic code - Chinese character string similarity calculation method based on "phonetic code"	version 1 version 2 blog/introduction

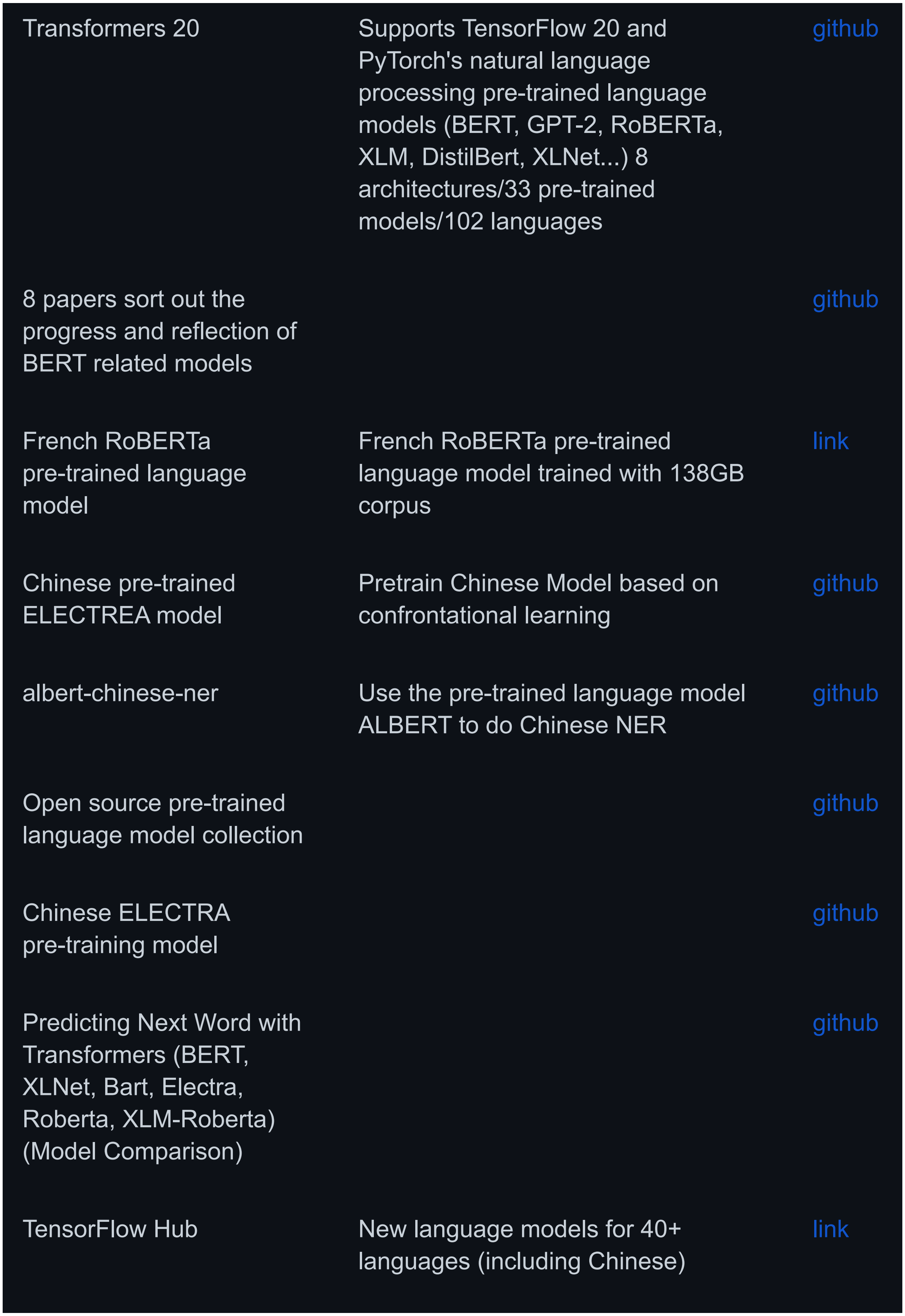
Acquisition of multiple meanings/sense items of Chinese words and semantic disambiguation of specific sentences based on the encyclopedia knowledge base		github
Tokenizer is a fast and customizable text tokenization library		github
Tokenizers	State-of-the-art tokenizer with a focus on performance and versatility	github
Realize text "face changing" through synonym replacement		github
token2index is a powerful lightweight term index library compatible with PyTorch/Tensorflow		github
Traditional and Simplified Conversion		github

Cantonese NLP Tools		github
domain dictionary	Professional dictionary knowledge base covering 68 fields with a total of 9.16 million words	github

Pre-trained language model & large model

Resource name (Name)	Description	Link
BMList	Big Model Big List	github
Chinese translation of bert papers		link
The slides of the original author of bert		link
Text Classification Practice		github
bert tutorial text classification tutorial		github
Bert pytorch implementation		github
Bert pytorch implementation		github

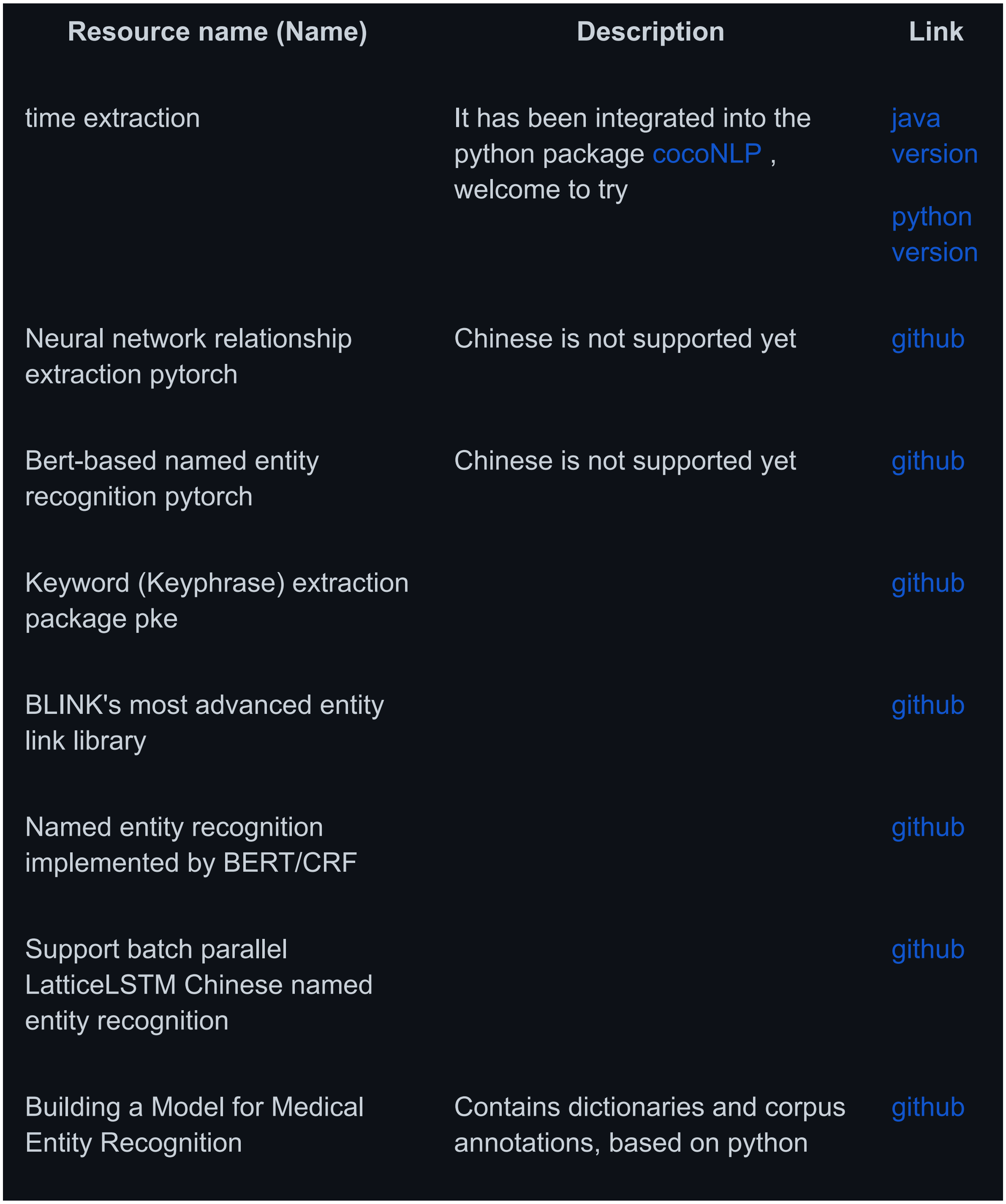
BERT generates sentence vectors, BERT does text classification and text similarity calculation		github
Diagram of bert and ELMO		github
BERT Pre-trained models and downstream applications		github
Language/Knowledge Representation Tool BERT & ERNIE		github
Using the gpt-2 language model in Kashgari		github
Facebook LAMA	Probes for analyzing factual and commonsense knowledge contained in pretrained language models. Language model analysis, providing a unified access interface for Transformer-XL/BERT/ELMo/GPT pre-trained language models	github
Chinese GPT2 training code		github
XLMFacebook's cross-language pre-trained language model		github
Massive Chinese pre-trained ALBERT model		github



Transformers 20	Supports TensorFlow 20 and PyTorch's natural language processing pre-trained language models (BERT, GPT-2, RoBERTa, XLM, DistilBert, XLNet...) 8 architectures/33 pre-trained models/102 languages	github
8 papers sort out the progress and reflection of BERT related models		github
French RoBERTa pre-trained language model	French RoBERTa pre-trained language model trained with 138GB corpus	link
Chinese pre-trained ELECTREA model	Pretrain Chinese Model based on confrontational learning	github
albert-chinese-ner	Use the pre-trained language model ALBERT to do Chinese NER	github
Open source pre-trained language model collection		github
Chinese ELECTRA pre-training model		github
Predicting Next Word with Transformers (BERT, XLNet, Bart, Electra, Roberta, XLM-Roberta) (Model Comparison)		github
TensorFlow Hub	New language models for 40+ languages (including Chinese)	link

UER	Chinese pre-trained model warehouses based on different corpora, encoders, and target tasks (including BERT, GPT, ELMO, etc.)	github
Open source pre-trained language model collection		github
Multilingual sentence vector package		github
Language Model as a Service (LMaaS)	Language Model as a Service	github
Open source language model GPT-NeoX-20B	20 billion parameters, currently the largest publicly accessible pre-trained general autoregressive language model	github
Chinese Science Literature Dataset (CSL)	Contains 396,209 meta-information (titles, abstracts, keywords, disciplines, categories) of papers in Chinese core journals. The CSL dataset can be used as a pre-training corpus, and can also be used to construct many NLP tasks, such as text summarization (title prediction), keyword generation, and text classification.	github
Large model development artifact		github

extract



Resource name (Name)	Description	Link
time extraction	It has been integrated into the python package cocoNLP , welcome to try	java version python version
Neural network relationship extraction pytorch	Chinese is not supported yet	github
Bert-based named entity recognition pytorch	Chinese is not supported yet	github
Keyword (Keyphrase) extraction package pke		github
BLINK's most advanced entity link library		github
Named entity recognition implemented by BERT/CRF		github
Support batch parallel LatticeLSTM Chinese named entity recognition		github
Building a Model for Medical Entity Recognition	Contains dictionaries and corpus annotations, based on python	github

Pipeline Entity and Relationship Extraction Based on TensorFlow and BERT	- Entity and Relation Extraction Based on TensorFlow and BERT Pipeline entity and relationship extraction based on TensorFlow and BERT, the solution to the information extraction task of the 2019 Language and Intelligence Technology Competition. Schema based Knowledge Extraction, SKE 2019	github
Chinese named entity recognition NeuroNER vs BertNER		github
Chinese Named Entity Recognition Based on BERT		github
Chinese key phrase extraction tool		github
bert	tensorflow version for Chinese named entity recognition	github
bert-Kashgari	Kashgari, a keras-based encapsulation classification and labeling framework, can build a classification or sequence labeling model in a few minutes	github
cocoNLP	Extraction of information such as name, address, email address, mobile phone number, mobile phone attribution, etc., rake phrase extraction algorithm.	github

Microsoft Multilingual Number/Unit/Eg Date Time Recognition Package		github
Baidu open source benchmark information extraction system		github
Chinese address word segmentation (identification and extraction of address elements), NER through sequence annotation		github
Open Domain Text Knowledge Triple Extraction and Knowledge Base Construction Based on Dependency Syntax		github
Chinese keyword extraction method based on pre-training model		github
chinese_keyphrase_extractor (CKPE)	A tool for chinese keyphrase extraction A tool for quickly extracting and identifying keyphrases from natural language text	github
Simple resume parser to extract key information from resumes		github
BERT-NER-Pytorch three different modes of BERT Chinese NER experiments		github

knowledge map

Resource name (Name)	Description	Link
Tsinghua University XLORE Chinese-English cross-language encyclopedia knowledge map	Baidu, Chinese Wiki, English Wiki	link
Automatic generation of document maps		github
Question answering system based on knowledge graph in medical field		github This repo refers to github
Chinese character relationship knowledge map project		github
AmpliGraph Knowledge Graph Representation Learning (Python) Library Knowledge Graph Concept Link Prediction		github
Chinese knowledge map materials, data and tools		github
Chinese Knowledge Graph Based on Baidu Encyclopedia	Extract triplet information and build a Chinese knowledge map	github

Zincbase Knowledge Graph Construction Toolkit		github
Question answering system based on knowledge graph		github
Collation of knowledge map deep learning related materials		github
Southeast University "Knowledge Graph" graduate course (data)		github
Knowledge map car audio work project		github
"One Piece" Knowledge Graph		github
A dataset of 132 knowledge graphs	Covers common sense, city, finance, agriculture, geography, weather, social networking, Internet of Things, medical care, entertainment, life, business, travel, science and education	link
Large-scale, structured, Chinese-English bilingual COVID-19 Knowledge Graph (COKG-19)		link

Event Triple Extraction Based on Dependency Syntax and Semantic Role Labeling		github
Abstract Knowledge Graph	The current scale is 500,000, supporting the abstraction of nominal entities, state descriptions, and event actions	github
Large-scale Chinese knowledge map data 1.4 billion entities		github
Jiagu natural language processing tool	Based on models such as BiLSTM, it provides functions such as knowledge graph relationship extraction, Chinese word segmentation, part-of-speech tagging, named entity recognition, sentiment analysis, new word discovery, keyword text summarization, text clustering, etc.	github
medical_NER - Chinese Medical Knowledge Graph Named Entity Recognition		github
A large list of learning materials/datasets/tool resources related to knowledge graphs		github

LibKGE is a knowledge graph embedding library for reproducible research		github
Military field knowledge map question answering project based on mongodb storage	Including aircraft, space equipment, etc. 8 categories, more than 100 subcategories, a total of 5,800 items of military weapons knowledge base, the project does not use a graph database for storage, through jieba to analyze questions, identify entity items in questions, and complete based on query templates The query of multiple types of questions is mainly to provide a demo of the question-and-answer thinking in the industry.	github
Jingdong Commodity Knowledge Graph		github
Chinese Relation Extraction Based on Distant Supervision		github
Intelligent Question Answering System Based on Medical Knowledge Graph		github
BLINK's most advanced entity link library		github

A small securities knowledge graph/knowledge base		github
dstlr unstructured text scalable knowledge map construction platform		github
Baidu Encyclopedia character entry attribute extraction	Using BERT-based fine-tuning and feature extraction methods for knowledge graphs	github
Data related to COVID-19	New crown and other types of pneumonia Chinese medical dialogue dataset; open data sources of institutions such as Tsinghua University (COVID-19)	github github
DGL-KE Graph Embedding Representation Learning Algorithm		github
causality map		method data
Causal Event Pairs Based on Multi-Domain Text Datasets		link

text generation

Resource name (Name)	Description	Link

Texar	Toolkit for Text Generation and Beyond	github
Prof. Ehud Reiter's Blog		link Professor Wan Xiaojun of Peking University strongly recommends this blog, which conducts in-depth discussions and reflections on NLG technology, evaluation and application.
Large list of resources related to text generation		github
Open Domain Dialogue Generation and Its Practice in Microsoft Xiaoice	Natural language generation allows machines to master the ability of automatic creation	link
Text Generation Control		github
A large list of natural language generation related resources		github
Evaluating Natural Language Generation with BLEURT		link
Automatic couplet data and robots		Code link

		700,000 couplet data
Automatically generate comments	Generating comments based on Hacker News article titles using Transformer codec model	github
Natural language generation SQL statement (English)		github
Natural Language Generation Resource Collection		github
Benchmarking Chinese Generation Tasks		github
Topic-specific text generation/text augmentation based on GPT2		github
Encoding, Tokenization, and Implementation of a Controlled and Efficient Text Generation Methodology		github
TextFooler's adversarial text generation module for text classification/inference		github

SimBERT	BERT model based on UniLM idea, integrating retrieval and generation	github
New word generation and sentence making	Non-existing words generate new words from scratch with GPT-2 variants, their definitions, and example sentences	github
Automatically generate multiple choice questions from text		github
Synthetic Data Generation Benchmark		github

text summary

Resource name (Name)	Description	Link
Chinese text summarization/keyword extraction		github
Automatic Summarization of Resume Based on Named Entity Recognition		github

Automatic text summarization library TextTeaser	English only	github
Extractive summary extraction based on the latest language models such as BERT		github
A Comprehensive Guide to Text Summarization with Deep Learning in Python		link
(Colab) Abstract Text Summary Implementation Highlights (Tutorial		github

Smart Q&A

Resource name (Name)	Description	Link
Chinese chatbot	Train the chatbot you want according to your own corpus, which can be used in scenarios such as intelligent customer service, online question and answer, intelligent chat, etc.	github
Interesting robot qingyun	Chinese chatbot trained by qingyun	github
Open dialogue robots, knowledge graphs, semantic understanding, natural language processing tools and data		github

qa right robot	Amodel-for-Retrivalchatbot - customer service robot, Chinese Retreival chatbot (Chinese retrieval robot)	git
ConvLab open source multi-domain end-to-end dialogue system platform		github
A dialog system based on the latest version of rasa		github
Chatbots based on the financial-judicial domain (with the nature of small talk)		github
End-to-end closed-domain dialogue system		github
MiningZhiDaoQACorpus	5.8 million Baidu Zhizhi Q&A data mining project, Baidu Zhizhi Q&A corpus, including more than 5.8 million questions, each with a question label. Based on this question and answer corpus, it can support a variety of applications, such as logic mining	github
GPT2 model GPT2-chitchat for Chinese chatting		github
Selection of relevant resource lists (Leaderboards, Datasets, Papers) based on multiple		github

rounds of responses from retrieval chatbots		
Microsoft Conversational Bot Framework		github
chatbot-list	Application and architecture of intelligent customer service and chatbots, algorithm sharing and introduction in the industry	github
Chinese medical dialogue data Chinese medical dialogue data set		github
A Large-Scale Medical Dialogue Dataset	Contains 1.1 million medical consultations and 4 million doctor-patient dialogues	github
Large-scale cross-domain Chinese task-oriented multi-round dialogue dataset and model CrossWOZ		paper & data
Open source conversational information search platform		github
Contextual Interaction Multimodal Dialogue Challenge 2020 (DSTC9 2020)		github
Use Quora questions to paraphrase the trained T5 questions (Paraphrase)		github

Google releases Taskmaster-2 natural language task dialogue dataset		github
Haystack's flexible, powerful, and extensible Question Answering (QA) framework		github
End-to-end closed-domain dialogue system		github
Amazon releases knowledge-based human-human open domain dialogue dataset		github
Albert Large QA model trained based on Baidu webqa and dureader dataset		github
CommonsenseQA Commonsense-Oriented English QA Challenge		link
MedQuAD (English) Medical Question Answering Dataset		github
A Q&A engine using Wikipedia text as context, based on Albert and Electra		github
A question answering attempt based on the 14W song knowledge base	Functions include Lyrics Solitaire, Finding Songs with Known Lyrics, and Questions and Answers about the	github

	Triangular Relationship of Song Artists Lyrics	

text error correction

Resource name (Name)	Description	Link
Chinese text error correction module code		github
English spell checking library		github
Python spell checking library		github
GitHub Typo Corpus Large-Scale GitHub Multilingual Spelling/Grammar Error Dataset		github
BertPunc BERT-based state-of-the-art punctuation repair model		github
Chinese writing proofreading tool		github
Text Error Correction Literature List	Chinese Spell Checking (CSC) and Grammatical Error Correction (GEC)	github
Winner of Text Smart Proofreading Contest	It has been applied, from the team of Soochow University and Dharma Academy	link

multimodal

Resource name (Name)	Description	Link
Chinese Multimodal Dataset "Wukong"	Huawei's Noah's Ark Laboratory open source large-scale, including 100 million text pairs	github
Chinese graphic representation pre-training model Chinese-CLIP	The Chinese version of the CLIP pre-training model, open source multiple model scales, and a few lines of code can handle Chinese image-text representation extraction & image-text retrieval	github

speech processing

Resource name (Name)	Description	Link
ASR Speech Dataset + Chinese Speech Recognition System Based on Deep Learning		github

Tsinghua University THCHS30 Chinese Speech Dataset	data_thchs30tgz-OpenSLR domestic image data_thchs30tgz test-noisetgz-OpenSLR domestic image test-noisetgz resourcetgz-OpenSLR domestic image resourcetgz Free ST Chinese Mandarin Corpus Free ST Chinese Mandarin Corpus AIShell-1 open source version dataset-OpenSLR domestic image AIShell-1 open source version dataset Primewords Chinese Corpus Set 1-OpenSLR Domestic Mirror Primewords Chinese Corpus Set 1
laughter detector	github

Common Voice Speech Recognition Dataset New Version	Includes over 1,400 hours of speech samples from 42,000 contributors, covering github	link
speech-aligner	A tool for generating phoneme-level time-aligned annotations from "human voice speech" and its "language text"	github
ASR Speech Dictionary/Dictionary		github
Speech Sentiment Analysis		github
masr	Chinese speech recognition, providing pre-training model, high recognition rate	github
Chinese Text Normalization for Speech Recognition		github
Voice quality evaluation indicators (MOSNet, BSSEval, STOI, PESQ, SRMR)		github
Chinese/English Pronunciation Dictionary for Speech Recognition		github

Multilingual speech-text translation corpus released by CoVoSTEFacebook	Includes audio, text transcription and English translation in 11 languages (French, German, Dutch, Russian, Spanish, Italian, Turkish, Persian, Swedish, Mongolian and Chinese)	github
Parakeet text-to-speech synthesis based on PaddlePaddle		github
(Java) Accurate Speech Natural Language Detection Library		github
Multilingual speech-text translation corpus released by CoVoSTEFacebook		github
Text-to-Speech Synthesis Implemented in TensorFlow 2		github
Python audio feature extraction package		github
ViSQOL audio quality perception is objective and complete reference index, divided into two modes: audio and voice		github

zhrtvc	Easy-to-use Chinese voice clone and Chinese speech synthesis system	github
aukit	An easy-to-use speech processing toolbox, including speech noise reduction, audio format conversion, feature spectrum generation and other modules	github
phkit	An easy-to-use phoneme processing toolbox, including Chinese phonemes, English phonemes, text-to-pinyin, text regularization and other modules	github
zhvoice	Chinese speech corpus, the speech is clearer and more natural, including 8 open source data sets, 3200 speakers, 900 hours of speech, 13 million words	github
audio for speech behavior detection	, binarization, speaker recognition, automatic speech recognition, emotion recognition and other audio annotation tools	github

Deep Learning Emotional Text-to-Speech Synthesis	github
Python audio data augmentation library	github
Audio Enhancement Based on Large-Scale Audio Dataset Audioset	github
voice transfer	github

document processing

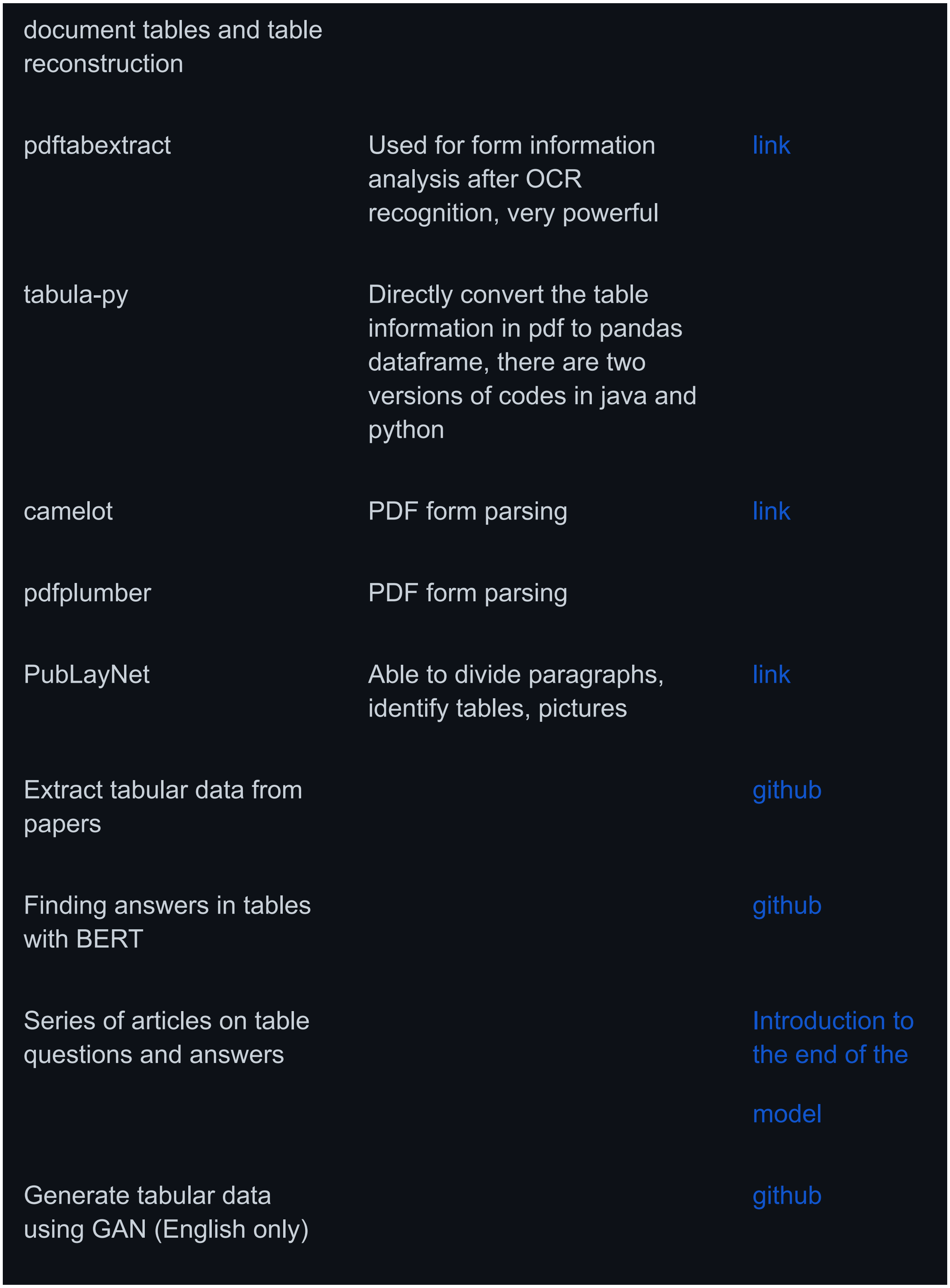
Resource name (Name)	Description	Link
LayoutLM-v3 Document Understanding Model		github
PyLaia Deep Learning Toolkit for Handwritten Document Analysis		github
Single-document unsupervised keyword extraction		github

DocSearch Free Documentation Search Engine		github
fdfgen	Ability to automatically create pdf documents and fill in information	link
pdfx	Automatically extract cited references and download the corresponding pdf file	link
invoice2data	Invoice pdf information extraction	invoice2data
PDF document information extraction		github
PDFMiner	PDFMiner can get the exact position of the text in the page, as well as other information such as font or line. It also has a PDF converter that can convert PDF files to other text formats such as HTML. There is also an extensible parser PDF that can be used for other purposes than text analysis.	link
PyPDF2	PyPDF 2 is a python PDF library capable of splitting, merging, cropping and converting pages of PDF files. It can also add custom data, viewing options and passwords to PDF files. It can retrieve text and metadata from PDFs, and can also merge entire files together.	link

PyPDF2	PyPDF 2 is a python PDF library capable of splitting, merging, cropping and converting pages of PDF files. It can also add custom data, viewing options and passwords to PDF files. It can retrieve text and metadata from PDFs, and can also merge entire files together.	link
ReportLab	ReportLab can quickly create PDF documents. A time-proven, super-easy-to-use open source project for creating complex, data-driven PDF documents and custom vector graphics. It's free, open source, and written in Python. With more than 50,000 downloads per month, the package is part of standard Linux distributions, embedded in many products, and was chosen to power Wikipedia's print/export functionality.	link
Simple PDF file text editor written by SIMPdfPython		github
pdf-diff	PDF file diff tool can display the difference between two pdf documents	github

form processing

Resource name (Name)	Description	Link
Use unet to realize automatic detection of		github



document tables and table reconstruction		
pdftabextract	Used for form information analysis after OCR recognition, very powerful	link
tabula-py	Directly convert the table information in pdf to pandas dataframe, there are two versions of codes in java and python	
camelot	PDF form parsing	link
pdfplumber	PDF form parsing	
PubLayNet	Able to divide paragraphs, identify tables, pictures	link
Extract tabular data from papers		github
Finding answers in tables with BERT		github
Series of articles on table questions and answers		Introduction to the end of the model
Generate tabular data using GAN (English only)		github

carefree-learn (PyTorch)	Automated Machine Learning (AutoML) Package for Tabular Datasets	github
Closed domain fine-tuning table detection		github
PDF form data extraction tool		github
TaBERT A New Model for Understanding Tabular Data Queries		paper
form processing	Awesome-Table-Recognition	github

text match

Resource name (Name)	Description	Link
Sentence, QA similarity matching MatchZoo	A collection of text similarity matching algorithms, including multiple deep learning methods, worth trying.	github
Chinese Question Sentence Similarity Calculation Competition and Scheme Summary		github
similarity similarity calculation toolkit	Written in java, it is used for similarity calculations related to words, phrases, sentences, lexical analysis,	github

	sentiment analysis, semantic analysis, etc.	
Chinese word similarity calculation method	Combined with the word similarity calculation method of Synonyms Cilin Extended Edition and Hownet, the vocabulary coverage is more and the results are more accurate.	gihtub
Python string similarity algorithm library		github
Similar sentence judgment model based on Siamese bilstm model, providing training data set and test data set	100,000 training samples provided	github

Text Data Augmentation

Resource name (Name)	Description	Link
Chinese NLP Data Augmentation (EDA) Tool		github
English NLP data enhancement tool		github
One-click Chinese data enhancement tool		github
The application and effect of data enhancement in machine translation and other nlp tasks		link
NLP Data Augmentation Resource Collection		github

Common regular expressions

<table>
<tr><th>Resource name (Name)</th><th>Description</th><th>Link</th></tr>
<tr><td>Regular expression to extract email</td><td></td><td>It has been integrated into the python package cocoNLP , welcome to try</td></tr>
<tr><td>Extract phone_number</td><td></td><td>It has been integrated into the python package cocoNLP , welcome to try</td></tr>
<tr><td>Regular expression for extracting ID number</td><td>IDCards_pattern = r'^([1-9]\d{5}[12]\d{3}(0[1-9]|1[012])(0[1-9]|[12][0-9]|3[01])\d{3}[0-9xX])
IDs = re.findall(IDCards_pattern, text, flags=0)</td><td></td></tr>
<tr><td>IP address regular expression</td><td>(25[0-5]| 2[0-4]\d| [0-1]\d{2}| [1-9]?\d).(25[0-5]| 2[0- 4]\d| [0-1]\d{2}| [1-9]?\d).(25[0-5]| 2[0-4]\d| [0-1]\d {2}| [1-9]?\d).(25[0-5]| 2[0-4]\d| [0-1]\d{2}| [1-9]?\d)</td><td></td></tr>
<tr><td>Tencent QQ number regular expression</td><td>[1-9]([0-9]{5,11})</td><td></td></tr>
</table>

Domestic fixed-line number regular expression	[0-9-()()]{7,18}	
username regex	[A-Za-z0-9_-\u4e00-\u9fa5]+	
Regular matching of domestic phone numbers (three major operators + virtual, etc.)		github
Regular Expression Tutorial		github

text search

Resource name (Name)	Description	Link
Efficient Fuzzy Search Tool		github
Large list/search engine of BERT models for various languages/tasks		link
Deepmatch's deep matching model library for recommendation, advertising and search		github

wwsearch is a full-text search engine developed by the enterprise WeChat background		github
aili - the fastest in-memory index in the East		github
Efficient string matching tool RapidFuzz	a fast string matching library for Python and C++, which is using the string similarity calculations from FuzzyWuzzy	github

reading comprehension

Resource name (Name)	Description	Link
Efficient Fuzzy Search Tool		github
Large list/search engine of BERT models for various languages/tasks		link
Deepmatch's deep matching model library for recommendation, advertising and search		github
allennlp reading comprehension supports a variety of data and models		github

emotion analysis

Resource name (Name)	Description	Link

aspect sentiment analysis package		github
awesome-nlp-sentiment-analysis	Sentiment analysis, emotional cause identification, evaluation object and evaluation word extraction	github
Sentiment analysis technology enables intelligent customer service to better understand human emotions		github

event extraction

Resource name (Name)	Description	Link
Chinese event extraction		github
List of Literature Resources for NLP Event Extraction		github
BERT event extraction implemented by PyTorch (ACE 2005 corpus)		github
News Event Clue Extraction		github

machine translation

Resource name (Name)	Description	Link

no way dictionary	The command line version of Youdao Dictionary supports English-Chinese mutual search and online search	github
NLLB	Language model NLLB that supports arbitrary inter-translation of 200+ languages	link
Easy-Translate	Script to translate large text files locally, based on Facebook/Meta AI's M2M100 model and NLLB200 model, supports 200+ languages	github

digital conversion

Resource name (Name)	Description	Link
The best Chinese character number (Chinese number)-Arabic number conversion tool		github
Quickly convert "Chinese numerals" and "Arabic numerals"		github
Parse and convert natural language numeric strings to integers and floating point numbers		github

anaphora resolution

Resource name (Name)	Description	Link
Chinese reference to digestion data		github

baidu ink code a0qq

text clustering

Resource name (Name)	Description	Link
TextCluster short text clustering preprocessing module Short text cluster		github

Text Categorization

Resource name (Name)	Description	Link
NeuralNLP-NeuralClassifier Tencent open source deep learning text classification tool		github

knowledge reasoning

Resource name (Name)	Description	Link
GraphbrainAI is an open source software library and research tools designed to facilitate automatic meaning extraction and text understanding as well as knowledge exploration and inference		github
(Harvard) free book on causal reasoning		pdf

Interpretable Natural Language Processing

Resource name (Name)	Description	Link
State-of-the-art interpreter library for textual machine learning models		github

text attack

Resource name (Name)	Description	Link
TextAttack natural language processing model adversarial attack framework		github
OpenBackdoor: Text backdoor attack and defense toolkit	OpenBackdoor is developed based on Python and PyTorch, which can be used to reproduce, evaluate and develop related algorithms for text backdoor attack and defense	github

text visualization

Resource name (Name)	Description	Link
Scattertext text visualization (python)		github

whatlies word vector interactive visualization		spacytools
PySS3 machine visualization tool for SS3 text classifiers for explainable AI		github
Render 3D images with Notepad		github
attnvisGPT2, BERT and other transformer language models attention interactive visualization		github
Texthero text data efficient processing package	Including preprocessing, keyword extraction, named entity recognition, vector space analysis, text visualization, etc.	github

text annotation tool

Resource name (Name)	Description	Link
Overview of NLP annotation platform		github
brat rapid annotation tool sequence annotation tool		link
Poplar web version natural language annotation tool		github

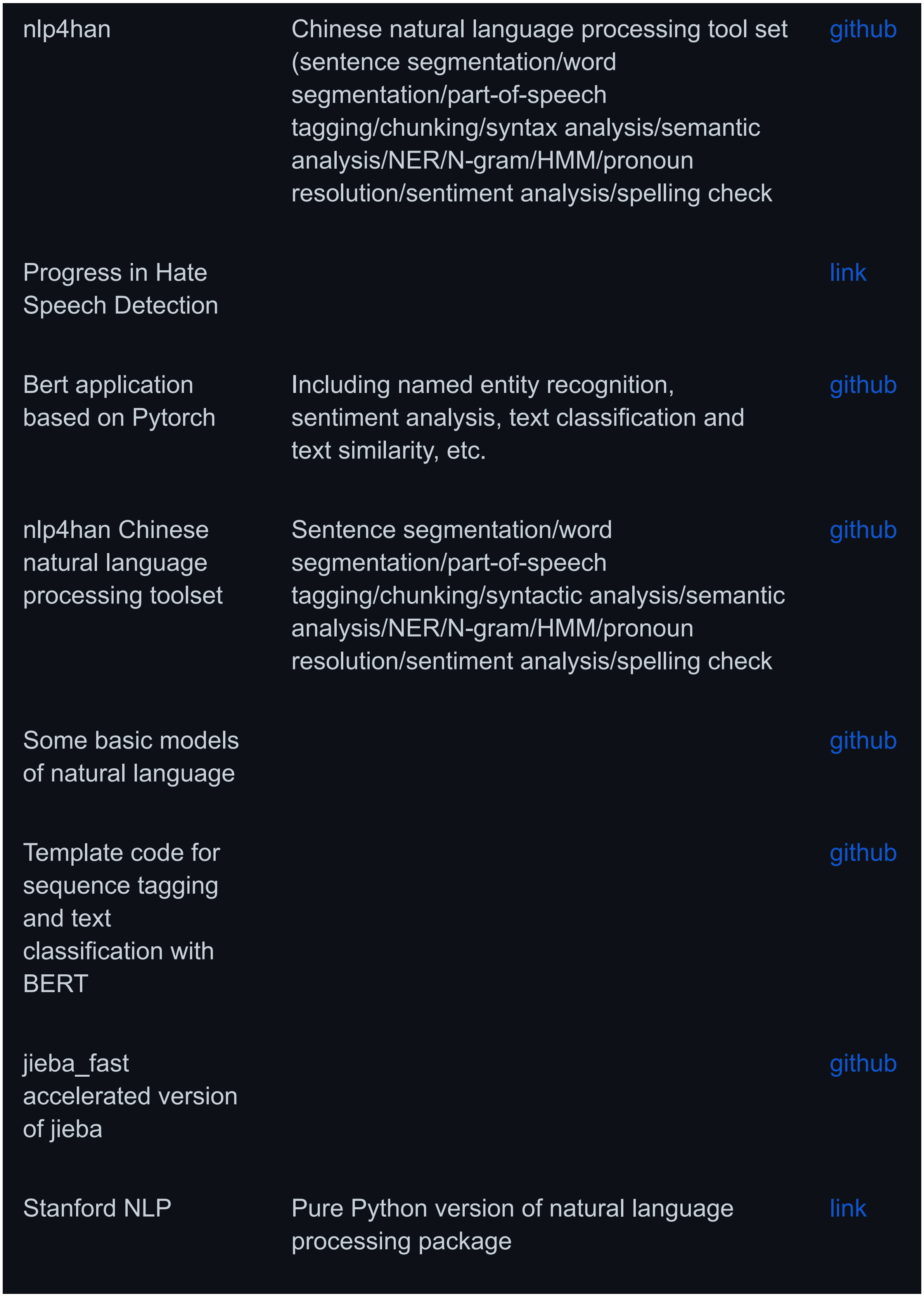
LIDA is a lightweight interactive dialogue annotation tool	github
doccano is a web-based open source collaborative multilingual text annotation tool	github
Datasaurai online data labeling workflow management tool	link

language detection

Resource name (Name)	Description	Link
langid	97 languages detected	https://github.com/saffsd/langid.py
langdetect	language detection	https://code.google.com/archive/p/language-detection/

comprehensive tool

Resource name (Name)	Description	Link
jieba		jieba
hanlp		hanlp



nlp4han	Chinese natural language processing tool set (sentence segmentation/word segmentation/part-of-speech tagging/chunking/syntax analysis/semantic analysis/NER/N-gram/HMM/pronoun resolution/sentiment analysis/spelling check	github
Progress in Hate Speech Detection		link
Bert application based on Pytorch	Including named entity recognition, sentiment analysis, text classification and text similarity, etc.	github
nlp4han Chinese natural language processing toolset	Sentence segmentation/word segmentation/part-of-speech tagging/chunking/syntactic analysis/semantic analysis/NER/N-gram/HMM/pronoun resolution/sentiment analysis/spelling check	github
Some basic models of natural language		github
Template code for sequence tagging and text classification with BERT		github
jieba_fast accelerated version of jieba		github
Stanford NLP	Pure Python version of natural language processing package	link

Python Spoken Natural Language Processing Toolset (English)		github
PreNLP natural language preprocessing library		github
Some papers and codes related to nlp	Including topic model, word vector (Word Embedding), named entity recognition (NER), text classification (Text Classificatin), text generation (Text Generation), text similarity (Text Similarity) calculation, etc., involving various nlp-related Algorithm, based on keras and tensorflow	github
Python text mining/NLP practical example		github
Forte's flexible and powerful natural language processing pipeline toolset		github
stanza Stanford team NLP tools	Can handle more than sixty languages	github
Fancy-NLP is a text knowledge mining tool for building product portraits		github

Comprehensive and easy Chinese NLP toolkit		github
Recurrence of vectorized recall pipelines commonly used in the industry based on DSSM		github
Texthero text data efficient processing package	Including preprocessing, keyword extraction, named entity recognition, vector space analysis, text visualization, etc.	github
nlpgnn graph neural network natural language processing toolbox		github
Macadam	Based on Tensorflow (Keras) and bert4keras, a natural language processing toolkit focusing on text classification, sequence labeling and relation extraction	github
LineFlow is an efficient NLP data loader for all deep learning frameworks		github
Arabica: Python text data exploratory analysis toolkit		github
Python stress testing tool: SMSBoom		github

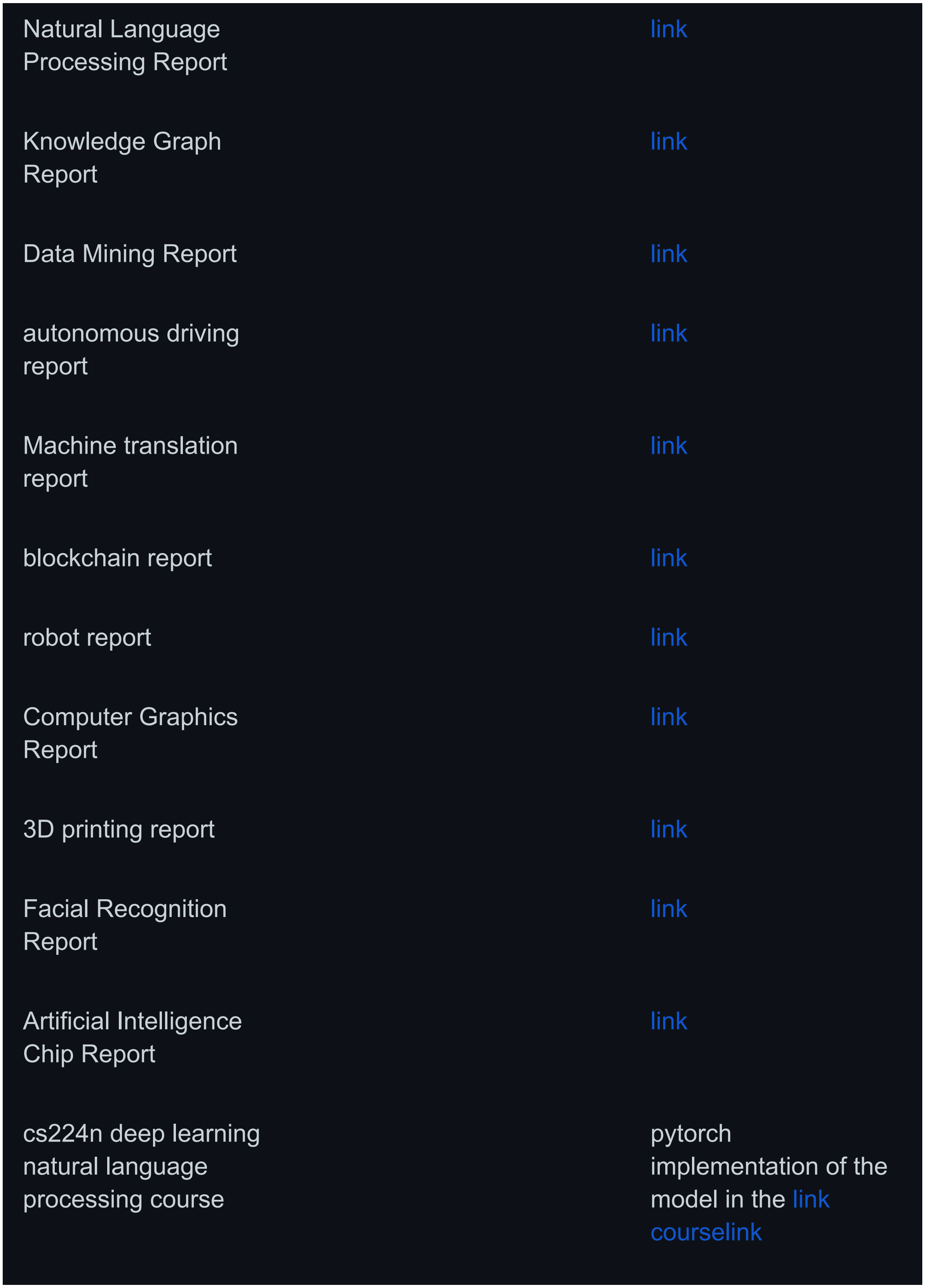
funny tool

Resource name (Name)	Description	Link
Wang Feng Lyric Generator		phunterlau/wangfeng-rnn
Analysis of girlfriend's emotional fluctuations		github
NLP is too difficult series		github
Variable naming artifact		github link
Image text removal, can be used for manga translation		github
CoupletAI - couplet generation	Automatic couplet system based on CNN+Bi-LSTM+Attention	github
Solving Complex Mathematical Equations Using Neural Network Symbolic Reasoning		github

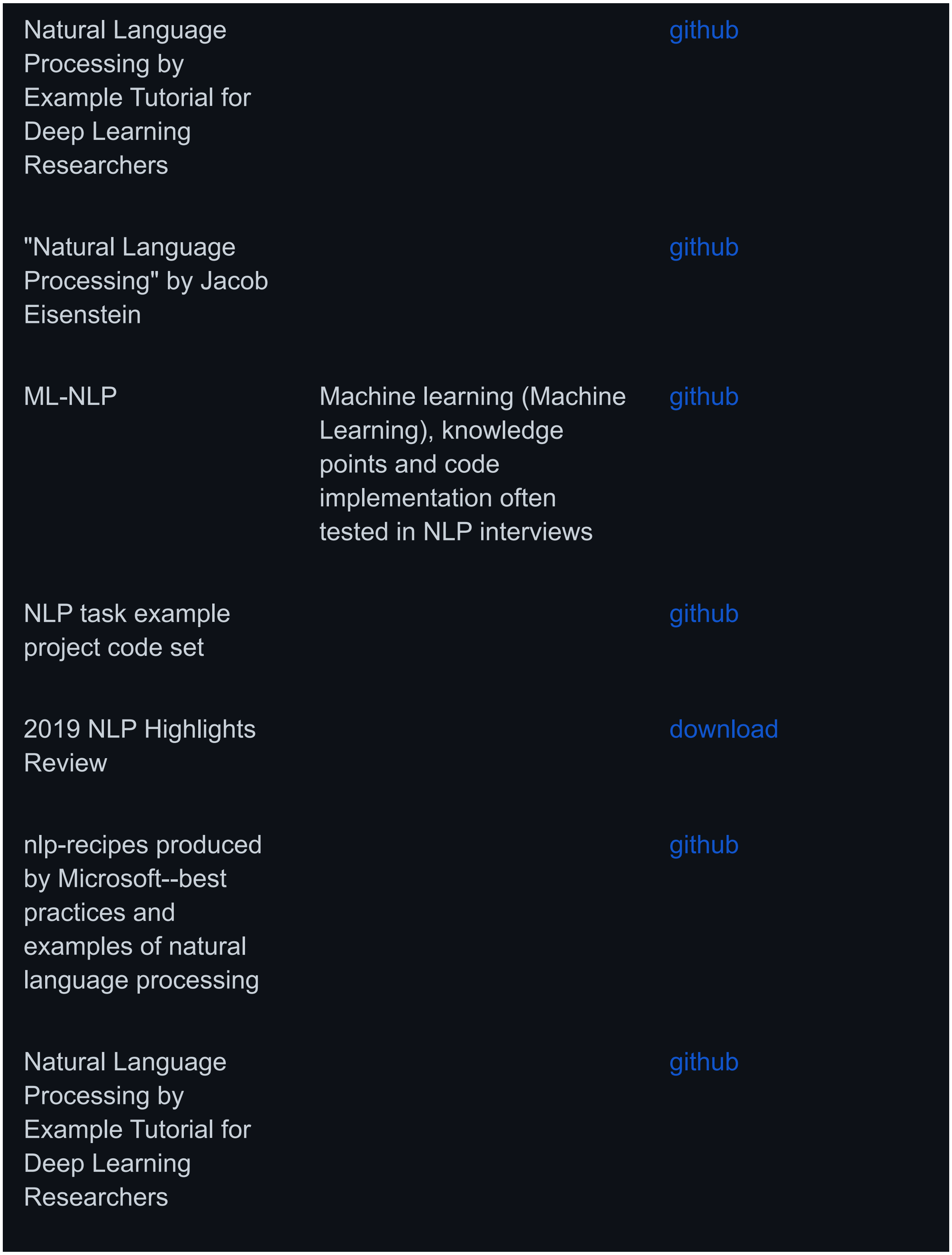
Question answering robot based on 14W song knowledge base	Functions include Lyrics Solitaire, Finding Songs with Known Lyrics, and Questions and Answers about the Triangular Relationship of Song Artists Lyrics	github
COPE - Metric Poem Editor		github
Paper2GUI	An AI desktop APP toolbox for ordinary people. It can be used immediately without installation. It already supports 18+ AI models, covering speech synthesis, video frame complementing, video super-resolution, target detection, image stylization, OCR recognition, etc.	github
Politeness estimator (trained using Sina Weibo data)		github paper
Grass python (Python Chinese version) getting started guide	Chinese programming language	homepage gitee

course report interview

Resource name (Name)	Description	Link



Natural Language Processing Report	link
Knowledge Graph Report	link
Data Mining Report	link
autonomous driving report	link
Machine translation report	link
blockchain report	link
robot report	link
Computer Graphics Report	link
3D printing report	link
Facial Recognition Report	link
Artificial Intelligence Chip Report	link
cs224n deep learning natural language processing course	pytorch implementation of the model in the link courselink



Natural Language Processing by Example Tutorial for Deep Learning Researchers		github
"Natural Language Processing" by Jacob Eisenstein		github
ML-NLP	Machine learning (Machine Learning), knowledge points and code implementation often tested in NLP interviews	github
NLP task example project code set		github
2019 NLP Highlights Review		download
nlp-recipes produced by Microsoft--best practices and examples of natural language processing		github
Natural Language Processing by Example Tutorial for Deep Learning Researchers		github

Transfer Learning in Natural Language Processing (NLP)	youtube
Machine Learning Systems book	link github

Contest

Resource name (Name)	Description	Link
Review the TOP solutions of all NLP competitions		github
2019 Baidu Triple Extraction Competition, "Scientific Space Team" source code (7th place)		github

Financial Natural Language Processing

Resource name (Name)	Description	Link
BDCI2019 Financial Negative Information Judgment		github
Open source financial investment data extraction tool		github
A large list of natural language processing research resources in the financial field		github
Chatbots based on the financial-judicial domain (with the nature of small talk)		github

Demonstration of small-scale financial knowledge graph construction process	github

Medical Natural Language Processing

Resource name (Name)	Description	Link
Chinese medical NLP public resources arrangement		github
spaCy Medical Text Mining and Information Extraction		github
Building a Model for Medical Entity Recognition	Contains dictionaries and corpus annotations, based on python	github
Question answering system based on knowledge graph in medical field		github This repo refers to github
Chinese medical dialogue data Chinese medical dialogue data set		github
A Large-Scale Medical Dialogue Dataset	Contains 1.1 million medical consultations and 4 million doctor-patient dialogues	github

Data related to COVID-19	New crown and other types of pneumonia Chinese medical dialogue dataset; open data sources of institutions such as Tsinghua University (COVID-19)	github github

Legal Natural Language Processing

Resource name (Name)	Description	Link
Blackstone's spaCy pipeline and NLP model for unstructured legal text		github
List of Forensic Intelligence Literature Resources		github
Chatbots based on the financial-judicial domain (with the nature of small talk)		github
Crime Legal Terms and Classification Model	Contains 856 crime knowledge graphs, crime prediction based on 2.8 million crime training database, 13 types of question classification and legal information question and answer function based on 20W legal question and answer pairs	github

text to image

Resource name (Name)	Description	Link
Dalle-mini	A mini version of DALL·E that generates pictures based on text prompts	github

other

Resource name (Name)	Description	Link
phone	China mobile phone attribution query	ls0f/phone
phone	International mobile phone and telephone attribution inquiry	AfterShip/phone
ngender	gender based on name	observers/ngender
A summary of the differences between Chinese and English natural language processing NLP		link
Technical documents PDF or PPT shared by Daniel in each major company		github
comparxiv is used to compare the difference between two submitted versions on arXiv		pypi
Meta-architecture of CHAMELEON deep learning news recommendation system		github
Automatic Resume Screening System		github

A variety of text readability evaluation indicators implemented by Python	github

Data Science ML Full Stack Roadmap
https://github.com/hemansnation/Data-Science-ML-Full-Stack-2022

Join the Data Science & ML Full Stack WhatsApp Group Community here:
If the group is full, please join another one.

https://chat.whatsapp.com/B7Mdp6QTMJ0KZYGWrziT3Y
https://chat.whatsapp.com/HWDSJU4KXrXJIcn5Npp3Gm
https://chat.whatsapp.com/DmATV5uaVY7IKrTMHDiHnr
https://chat.whatsapp.com/Blz2n8QYSgdKWfQbJZxHtJ

Join Telegram for Data Science ML AI Resources:
https://t.me/+sREuRiFssMo4YWJI
Join Community on LinkedIn:
https://www.linkedin.com/groups/12540639/

Connect with me on these platforms:
LinkedIn: https://www.linkedin.com/in/hemansnation/
Twitter: https://twitter.com/hemansnation
GitHub: https://github.com/hemansnation
Instagram: https://www.instagram.com/masterdexter.ai/

Are you a professional?
DM for One-on-One sessions for Python, Data Science, Machine Learning, and Data Engineering.
Here: https://bit.ly/3U6zQvQ

Python Notion Template
https://hemansnation.gumroad.com/l/god-level-python-with-himanshu-ramchandani