# 45个小众而实用的NLP开源字典和工具

原创 夕小瑶 夕小瑶的卖萌屋 2月14日

来自专辑

卖萌屋@自然语言处理



一只小狐狸带你解锁NLP/ML/DL秘籍

前言

随着BERT、ERNIE、XLNet等预训练模型的流行,解决NLP问题时不用上预训练模型似乎总显得自己有点过时。<mark>但是这显然是不对</mark>的。

众所周知,无论训练还是推理,预训练模型都会消耗大量的算力,且高度依赖GPU计算资源。然而,有很多的NLP问题实际上仅仅 靠<mark>字典+规则</mark>就可以做到够用,那么这时候强行上笨重的模型无异于高射炮打蚊子,性价比是非常低的。

于是小夕就从一个比较疯狂的github repo里为大家<mark>精挑细选了45个</mark>比较实用的开源小工具和字典,让大家在搭建NLP系统、辅助炼 丹的过程中少一些模型和算力的依赖,多一些小而美的代码。

# repo地址:

https://github.com/fighting41love/funNLP

注:这是一个非常丧心病狂的repo,里面收录了300多个项目,但是比较鱼龙混杂,要记得多多横向对比哦

来,感受一下m(\_\_)m

涉及内容包括但不限于:中英文敏感词、语言检测、中外手机/电话归属地/运营商查询、名字推断性别、手机号抽取、身份证 抽取、邮箱抽取、中日文人名库、中文缩写库、拆字词典、词汇情感值、停用词、反动词表、暴恐词表、繁简体转换、英文模 拟中文发音、汪峰歌词生成器、职业名称词库、同义词库、反义词库、否定词库、汽车品牌词库、汽车零件词库、连续英文切 割、各种中文词向量、公司名字大全、古诗词库、IT词库、财经词库、成语词库、地名词库、历史名人词库、诗词词库、医学 词库、饮食词库、法律词库、汽车词库、动物词库、中文聊天语料、中文谣言数据、百度中文问答数据集、句子相似度匹配算 法集合、bert资源、文本生成&摘要相关工具、cocoNLP信息抽取工具、国内电话号码正则匹配、清华大学XLORE:中英文跨 语言百科知识图谱、清华大学人工智能技术系列报告、自然语言生成、NLU太难了系列、自动对联数据及机器人、用户名黑名 单列表、罪名法务名词及分类模型、微信公众号语料、cs224n深度学习自然语言处理课程、中文手写汉字识别、中文自然语 言处理 语料/数据集、变量命名神器、分词语料库+代码、任务型对话英文数据集、ASR 语音数据集 + 基于深度学习的中文语 音识别系统、笑声检测器、Microsoft多语言数字/单位/如日期时间识别包、中华新华字典数据库及api(包括常用歇后语、成 语、词语和汉字)、文档图谱自动生成、SpaCy 中文模型、Common Voice语音识别数据集新版、神经网络关系抽取、基于 bert的命名实体识别、关键词(Keyphrase)抽取包pke、基于医疗领域知识图谱的问答系统、基于依存句法与语义角色标注的 事件三元组抽取、依存句法分析4万句高质量标注数据、cnocr: 用来做中文OCR的Python3包、中文人物关系知识图谱项 目、中文nlp竞赛项目及代码汇总、中文字符数据、speech-aligner: 从"人声语音"及其"语言文本"产生音素级别时间对齐标 注的工具、AmpliGraph: 知识图谱表示学习(Python)库: 知识图谱概念链接预测、Scattertext 文本可视化(python)、语言/ 知识表示工具: BERT & ERNIE、中文对比英文自然语言处理NLP的区别综述、Synonyms中文近义词工具包、HarvestText 领域自适应文本挖掘工具(新词发现-情感分析-实体链接等)、word2word: (Python)方便易用的多语言词-词对集: 62种 语言/3,564个多语言对、语音识别语料生成工具:从具有音频/字幕的在线视频创建自动语音识别(ASR)语料库、构建医疗实 体识别的模型(包含词典和语料标注)、单文档非监督的关键词抽取、Kashgari中使用gpt-2语言模型、开源的金融投资数据 提取工具、文本自动摘要库TextTeaser: 仅支持英文、人民日报语料处理工具集、一些关于自然语言的基本模型、基于14W歌 曲知识库的问答尝试--功能包括歌词接龙and已知歌词找歌曲以及歌曲歌手歌词三角关系的问答、基于Siamese bilstm模型的 相似句子判定模型并提供训练数据集和测试数据集、用Transformer编解码模型实现的根据Hacker News文章标题自动生成评 论、用BERT进行序列标记和文本分类的模板代码、LitBank: NLP数据集——支持自然语言处理和计算人文学科任务的100部 带标记英文小说语料、百度开源的基准信息抽取系统、虚假新闻数据集、Facebook: LAMA语言模型分析,提供 Transformer-XL/BERT/ELMo/GPT预训练语言模型的统一访问接口、CommonsenseQA: 面向常识的英文QA挑战、中文知 识图谱资料、数据及工具、各大公司内部里大牛分享的技术文档 PDF 或者 PPT、自然语言生成SQL语句(英文)、中文NLP 数据增强(EDA)工具、英文NLP数据增强工具 、基于医药知识图谱的智能问答系统、京东商品知识图谱、基于mongodb存

鬼知道我是怎么把这300个repo看完一遍的

1. textfilter: 中英文敏感词过滤

repo: observerss/textfilter

>>> f = DFAFilter()
>>> f.add("sexy")
>>> f.filter("hello sexy baby")
hello \*\*\*\* baby

敏感词包括政治、脏话等话题词汇。其原理主要是基于词典的查找(项目中的keyword文件),内容很不清真

2. langid: 97种语言检测

repo: saffsd/langid.py

pip install langid

>>> import langid>>> langid.classify("This is a test") ('en', -54.41310358047485)

3. langdetect: 另一个语言检测

地址: https://code.google.com/archive/p/language-detection

pip install langdetect

from langdetect import detectfrom langdetect import detect\_langs

s1 = "本篇博客主要介绍两款语言探测工具,用于区分文本到底是什么语言,"s2 = 'We are pleased to introduce today a new technology'p rint(detect(s1))

print(detect(s2))

print(detect\_langs(s3)) # detect\_langs()输出探测出的所有语言类型及其所占的比例

输出结果如下:注:语言类型主要参考的是ISO 639-1语言编码标准,详见ISO 639-1百度百科

跟上一个语言检测比较,准确率低,效率高。

### 4. phone 中国手机归属地查询:

repo: Is0f/phone

已集成到 python package cocoNLP中

from phone import Phone

p = Phone()

p.find(18100065143)#return {'phone': '18100065143', 'province': '上海', 'city': '上海', 'zip\_code': '200000', 'area\_code': '021', 'phone\_type': '电信'}

支持号段: 13,15,18\*,14[5,7],17[0,6,7,8]

记录条数: 360569 (updated:2017年4月)

作者提供了数据phone.dat 方便非python用户Load数据。

### 5. phone国际手机、电话归属地查询:

repo: AfterShip/phone

npm install phone

import phone from 'phone';

phone('+852 6569-8900'); // return ['+85265698900', 'HKG'] phone('(817) 569-8900'); // return ['+18175698900, 'USA']

# 6. ngender 根据名字判断性别:

repo: observerss/ngender

# 基于朴素贝叶斯计算的概率

pip install ngender

>>> import ngender>>> ngender.guess('赵本山') ('male', 0.9836229687547046)>>> ngender.guess('宋丹丹') ('female', 0.9759486128949907)

# 7. 抽取email的正则表达式

已集成到 python package cocoNLP中

email\_pattern =  $'^[*#\u4e00-\u9fa5 a-zA-Z0-9_.-]+@[a-zA-Z0-9-]+(\.[a-zA-Z0-9-]+)*\.[a-zA-Z0-9]{2,6}$'emails = re.findall(email_pattern, t ext, flags=0)$ 

# 8. 抽取phone\_number的正则表达式

 $cellphone\_pattern = \ '^{((13[0-9])(14[0-9])(15[0-9])(17[0-9])(18[0-9]))} \\ (18[0-9])) \\ (18[0$ 

# 9. 抽取身份证号的正则表达式

### 10. 人名语料库:

repo: wainshine/Chinese-Names-Corpus

人名抽取功能已加入 python package cocoNLP

中文(现代、古代)名字、日文名字、中文的姓和名、称呼(大姨妈、小姨妈等)、英文->中文名字(李约翰)、成语词典

(可用于中文分词、姓名识别)

### 11. 中文缩写库:

repo: zhangyics/Chinese-abbreviation-dataset

全国人大: 全国/n 人民/n 代表大会/n

中国: 中华人民共和国/ns 女网赛: 女子/n 网球/n 比赛/vn

### 12. 汉语拆字词典:

repo: kfcd/chaizi

漢字 拆法 (一) 拆法 (二) 拆法 (三) 拆 手斥 扌斥 才斥

### 13. 词汇情感值:

repo: rainarch/SentiBridge

山泉水 充沛 0.400704566541 0.370067395878 视野 宽广 0.305762728932 0.325320747491 大峡谷 惊险 0.312137906517 0.378594957281

## 14. 中文词库、停用词、敏感词

repo: dongxiexidian/Chinese

# 此package的敏感词库分类更细:

反动词库, 敏感词库表统计, 暴恐词库, 民生词库, 色情词库

# 15. 汉字转拼音:

repo: mozillazg/python-pinyin

# 文本纠错会用到

## 16. 中文繁简体互转:

repo: skydark/nstools

## 17. 英文模拟中文发音引擎

repo: tinyfool/ChineseWithEnglish

say wo i ni #说:我爱你

相当于用英文音标,模拟中文发音。

18. 同义词库、反义词库、否定词库:

repo: guotong1988/chinese\_dictionary

19. 中文字符数据

repo: skishore/makemeahanzi

- 简/繁体汉字笔顺
- 矢量笔画
- 20. 无空格英文串分割、抽取单词:

repo: keredson/wordninja

>>> import wordninja>>> wordninja.split('derekanderson') ['derek', 'anderson']>>> wordninja.split('imateapot') ['im', 'a', 'teapot']

# 21. IP地址正则表达式:

 $(25[0-5]|2[0-4] \land |[0-1] \land |$ 

22. 腾讯QQ号正则表达式:

[1-9]([0-9]{5,11})

23. 国内固话号码正则表达式:

[0-9-() () ]{7,18}

24. 用户名正则表达式:

 $[A-Za-z0-9_\-\u4e00-\u9fa5]+$ 

25. g2pC: 基于上下文的汉语读音自动标记模块

repo: Kyubyong/g2pC

26. 时间抽取:

已集成到 python package cocoNLP中

# 在2016年6月7日9:44执行測試,结果如下 Hi, all。下周一下午三点开会 >> 2016-06-13 15:00:00-false 周一开会 >> 2016-06-13 00:00:00-true 下下周一开会 >> 2016-06-20 00:00:00-true

java version:

https://github.com/shinyke/Time-NLP

python version:

https://github.com/zhanzecheng/Time\_NLP

# 27. 快速转化「中文数字」和「阿拉伯数字」

repo: HaveTwoBrush/cn2an

- 中文、阿拉伯数字互转
- 中文与阿拉伯数字混合的情况,在开发中

### 28. 公司名字大全

repo: wainshine/Company-Names-Corpus

# 29. 古诗词库

repo: panhaiqi/AncientPoetry

## 更全的古诗词库:

https://github.com/chinese-poetry/chinese-poetry

### 30. THU整理的词库

repo: http://thuocl.thunlp.org/

已整理到本repo的data文件夹中.

IT词库、财经词库、成语词库、地名词库、历史名人词库、诗词词库、医学词库、饮食词库、法律词库、汽车词库、动物词库

# 31. PDF表格数据提取工具

repo: camelot-dev/camelot

# 32. 国内电话号码正则匹配(三大运营商+虚拟等)

repo: VincentSit/ChinaMobilePhoneNumberRegex

# 33.用户名黑名单列表:

repo: marteinn/The-Big-Username-Blacklist

# 包含了用户名禁用列表,比如:

administrator administration autoconfig autodiscover broadcasthost domain editor guest host hostmaster info keybase.txt localdomain localhost master mail mail0 mail

### 34. Microsoft多语言数字/单位/如日期时间识别包:

repo: Microsoft/Recognizers-Text

35. chinese-xinhua 中华新华字典数据库及api,包括常用歇后语、成语、词语和汉字

repo: pwxcoo/chinese-xinhua

## 36. 文档图谱自动生成

repo: liuhuanyong/TextGrapher

• TextGrapher - Text Content Grapher based on keyinfo extraction by NLP method。输入一篇文档,将文档进行关键信息提取,进行结构化,并最终组织成图谱组织形式,形成对文章语义信息的图谱化展示

### 37. 186种语言的数字叫法库

repo: google/UniNum

# 38. 繁简体转换

repo: berniey/hanziconv

39. 汉字字符特征提取器 (featurizer),提取汉字的特征(发音特征、字形特征)用做深度学习的特征

repo: howl-anderson/hanzi\_char\_featurizer

### 40. 中文缩写数据集

repo: zhangyics/Chinese-abbreviation-dataset

41. 无道词典 - 有道词典的命令行版本,支持英汉互查和在线查询

repo: ChestnutHeng/Wudao-dict

42. 最好的汉字数字(中文数字)-阿拉伯数字转换工具

repo: Wall-ee/chinese2digits

43. LineFlow: 面向所有深度学习框架的NLP数据高效加载器

repo: tofunlp/lineflow

### 44. 将自然语言数字串解析转换为整数和浮点数

repo: jaidevd/numerizer

### 45. 英文脏话大列表

repo: zacanger/profane-words

此外,这个repo里还收录了很多数据集,不过也是比较杂乱,小夕这里都pass掉了,有需要的小伙伴可以去repo里翻一翻。

另外,Stanford CS224n-2020-winter追剧计划 的github repo也建设好了哦,做作业、追视频有困难的小伙伴可以在这个repo里找到 其他小伙伴整理的笔记、总结感悟、作业和project。



同时这也是一个开源开放的作业批改平台,将自己的作业和课程理解提交到该repo下,会有助教和其他小伙伴帮忙审阅修改哦。<mark>快快来提交&star吧</mark>

https://github.com/xixiaoyao/CS224n-winter-together



- Stanford CS224n追剧计划详细攻略
- 深度网络为何会有灾难性遗忘? 如何进行有效的持续学习?
- 训练太慢? 显存不够用? 混合精度训练了解一下
- NLP最佳入门书籍与必刷paper list



### 夕小瑶的卖萌屋

关注&星标小夕,带你解锁AI秘籍 内容过于专业,胆小者慎入 声明:pdf仅供学习使用,一切版权归原创公众号所有;建议持续关注原创公众号获取最新文章,学习愉快!