





莫中文自然语言处理语料 Large Scale Chinese Corpus for NLP

 24 commits

 1 branch

 0 releases

 1 contributor

ch: master ▾

New pull request

Find file

Clone or download

brightmart update

Latest commit 9841a26 3 days ago

resources

update

3 days ago

README.md

update

3 days ago

README.md

为中文自然语言处理领域发展贡献语料

GitHub出现一个大型中文NLP资源，宣称要放出亿级语料库

 量子位 
已认证的官方帐号

已关注

248 人赞同了该文章


乾明 发自 凹非寺
量子位 报道 | 公众号 QbitAI


中文信息很多，但要找到合适的中文语料很难。


有人看不下去了，在GitHub上开了一个项目，专门贡献中文语料资源。


他说，要为解决中文语料难找贡献一份力量。

大规模中文自然语言处理语料 Large Scale Chinese Corpus for NLP

 24 commits

 1 branch

 0 releases


 1 contributor

Branch: master ▾


New pull request

Find file

Clone or download ▾


 brightmart update

Latest commit 9841a26 3 days ago

 resources


update


3 days ago

 README.md

update

3 days ago

 README.md

 量子位

为中文自然语言处理领域发展贡献语料

▲ 赞同 248 ▼ 9 条评论



目前，这个项目中一共有3种json版资源：

包含104万个词条的维基百科资源，包含250万篇新闻的新闻语料，以及包含150万个问答的百科类问答资源。

维基百科(wiki2019zh) ---- 新闻语料(news2016zh) ---- 百科问答(baike2018qa)

1. 维基百科json版(wiki2019zh)

104万个词条(1,043,224条; 原始文件大小1.6G, 压缩文件519M; 数据更新时间: 2019.2.7)

[点此下载](#)

2. 新闻语料json版(news2016zh)

250万篇新闻(原始数据9G, 压缩文件3.6G; 新闻内容跨度: 2014-2016年)

[点此下载](#), 密码: film

3. 百科类问答json版(baike2018qa)

150万个问答(原始数据1G多, 压缩文件663M; 数据更新时间: 2018年)

[点此下载](#), 密码: fu45



一般来说，这些资源可以作为通用的中文语料，用于预训练或者构建词向量等等。

不同的资源，用处也有不同，比如维基百科和问答百科，可以用来构建知识问答等等。

新闻语料资源，囊括了标题、关键词、描述和正文，也可以用来训练标题生成模型、关键词生成模型等等。

此外，在对数据集划分过的新闻语料和百科类问答资源中，只提供训练集和验证集，不提供测试集数据的下载。

是因为——

▲ 赞同 248



● 9 条评论

知乎

首发于
量子位

资源的贡献者表示，希望大家报告模型在验证集上的准确率，并提供模型信息、方法描述、运行方式，以及可运行的源代码（可选）。

这些信息都有的话，资源贡献者会在测试集上测试模型，并给出准确率。

他表示，项目中的语料库将会不断扩充，号召大家多多贡献资源，并给出了相应的目标：

到2019年5月1日，放出10个百万级中文语料&3个千万级中文语料。

到2019年12月31日，放出30个百万级中文语料 & 10个千万级中文语料 & 1个亿级中文语料。

从目前已经有的资源来看，一个语料可以是一个问答，也可以是一个词条等等。

这份资源的贡献者，名为徐亮，杭州实在智能的算法专家，主要关注文本分类、意图识别、问答和面向任务的对话。

如果你有兴趣，请收好资源传送门：

[github.com/brightmart/n...](https://github.com/brightmart/nlp_data)

此外，量子位之前也介绍过几份中文NLP资源，也一并附于此：

[有人收罗了40个中文NLP词库，放到了GitHub上](#)

[腾讯AI Lab开源800万中文词的NLP数据集](#)

[非正式汉语数据集资源上线，帮你训练网络语言处理](#)

—完—

量子位 · QbitAI

🔍 追踪AI技术和产品新动态

戳右上角「+关注」获取最新资讯 ↗ ↗

如果喜欢，请分享or点赞吧~比心♥

发布于 2019-02-14

[自然语言处理](#) [人工智能](#) [编程](#)

▲ 赞同 248 ▼

● 9 条评论

知乎



首发于
量子位

文章被以下专栏收录



量子位

关注专栏

推荐阅读

NLP领域最新文章

分词、词性标注、命名实体识别、细粒度情感分析、自动摘要（多文档摘要）、阅读理解、关系抽取、事件抽取、语义匹配

johnchenyl

北大开源全新中 准确率远超THU

选自GitHub，作者晶、孙栩，机器之北大开源了一个中它在多个分词数据的分词准确率。其巴分词误差率高达机器之心

9 条评论

⇌ 切换为时间排序

写下你的评论...



精选评论 (1)



mountain blue

1 个月前

文章说到：有人收罗了40个中文NLP词库，放到了GitHub上。那个人就是我[赞同]，欢迎关注我一下，谢谢！

👍 35 💬 查看回复

评论 (9)



mountain blue

1 个月前

文章说到：有人收罗了40个中文NLP词库，放到了GitHub上。关注我一下，谢谢！

▲ 赞同 248



💬 9 条评论

肝牛 回复 mountain blue

11 月前

Extract basic information from texts

```
>>> from cocohlp.extractor import extractor
>>> ex = extractor()
>>> text = '急需林明盛 男孩，于2018年11月27号11时在陕西省安康
# 抽取邮箱
>>> emails = ex.extract_email(text)
\\>> print(emails)
```

大佬膜拜了！

👍 2



bright 回复 mountain blue

28 天前

看头像就知道很👍

👍 赞

展开其他 3 条回复



夏思畅

26 天前

请问有没有开源可用的中文对话机器人呢？支持一些自定义问题的。

👍 赞



苏灵

6 天前

中文wiki在墙外啊

👍 赞

1 条评论被折叠（为什么？）