为中文自然语言处理领域发展贡献语料

贡献中文语料，请发送邮件: CLUEbenchmark@163.com

*** update ****

中文任务基准测评，10大任务 & 9个模型一键运行、详细测评：

Language Understanding Evaluation benchmark for Chinese(CLUE benchmark): run 10 tasks & 9 baselines with one line of code, performance comparision with details.

Releasing Pre-trained Model of ALBERT_Chinese:

Training with 30G+ Raw Chinese Corpus, xxlarge, small version and more, Target to match State of the Art performance in Chinese with 30% less parameters, 2019-Oct-7, During the National Day of China!

语料库将会不断扩充。。。

一期目标：10个百万级中文语料 & 3个千万级中文语料(2019年5月1号)

二期目标：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（2019年12月31日）

Update： 增加高质量社区问答json版(webtext2019zh)，可用于训练超大规模NLP模型；添加520万翻译语料(translation2019zh)。

1.维基百科(wiki2019zh)，100万个结构良好的中文词条

2.新闻语料(news2016zh)，250万篇新闻，含关键词、描述

3.百科问答(baike2018qa)，150万个带问题类型的问答

4.社区问答json版(webtext2019zh)，410万个高质量社区问答，适合训练超大模型

5.翻译语料(translation2019zh)，520万个中英文句子对

为什么需要这个项目

中文的信息无处不在，但如果想要获得大量的中文语料，却是不太容易，有时甚至非常困难。在2019年初这个时点上，

普通的从业者、研究人员或学生，并没有一个比较好的渠道获得极大量的中文语料。笔者想要训练一个中文的词向量，

在百度和github上上搜索了好久，收获却很少：要么语料的量级太小，要么数据过于成旧，或需要的处理太复杂。

不知道你是否也遇到了这样的问题？

我们这个项目，就是为了解决这一问题贡献微薄之力。

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)

Google Drive下载 或 百度云盘

可能的用途：

可以做为通用中文语料，做预训练的语料或构建词向量，也可以用于构建知识问答。

结构：

`{"id":<id>,"url":<url>,"title":<title>,"text":<text>}` 其中，`title`是词条的标题，`text`是正文；通过`"\n\n"`换行。

例子：


```
{"id": "53", "url": "https://zh.wikipedia.org/wiki?curid=53", "title": "经济学", "text": "经济学\n\n经济学是一门对产品和服务的生产、分配以及消费进行研究的社会科学。西方语言中的“经济学”一词源于古希腊的。\n\n经济学注重的是研究经济行为者在一个经济体系下的行为，以及他们彼此之间的互动。在现代，经济学的教材通常将这门领域的研究分为总体经济学和个体经济学。微观经济学检视一个社会里基本层次的行为，包括个体的行为者（例如个人、公司、买家或卖家）以及与市场的互动。而宏观经济学则分析整个经济体和其议题，包括失业、通货膨胀、经济成长、财政和货币政策等。..."}
```


效果：

经济学
经济学是一门对产品和服务的生产、分配以及消费进行研究的社会科学。西方语言中的“经济学”一词源于古希腊的。
经济学注重的是研究经济行为者在一个经济体系下的行为，以及他们彼此之间的互动。在现代，经济学的教材通常将这门领域的研究分为总体经济学和个体经济学。微观经济学检视一个社会里基本层次的行为，包括个体的行为者（例如个人、公司、买家或卖家）以及与市场的互动。而宏观经济学则分析整个经济体和其议题，包括失业、通货膨胀、经济成长、财政和货币政策等。
其他的对照还包括了实证经济学（研究「是什么」）以及规范经济学（研究「应该是什么」）、经济理论与实用经济学、行为经济学与理性选择经济学、主流经济学（研究理性-个体-均衡等）与非主流经济学（研究体制-历史-社会结构等）。
经济学的分析也被用在其他各种领域上，主要领域包括了商业、金融、和政府等，但同时也包括了如健康、犯罪、教育、法律、政治、社会架构、宗教、战争、和科学等等。到了21世纪初，经济学在社会科学领域各方面不断扩张影响力，使得有些学者讽刺地称其为「经济学帝国主义」。
在现代对于经济学的定义有数种说法，其中有许多说法因为发展自不同的领域或理论而有截然不同的定义，苏格兰哲学家和经济学家亚当·斯密在1776年将政治经济学定义为「国民财富的性质和原因的研究」，他说：
让-巴蒂斯特·赛伊在1803年将经济学从公共政策里独立出来，并定义其为对于财富之生产、分配、和消费的学问。另一方面，托马斯·卡莱尔则讽刺的称经济学为「忧郁的科学」（Dismal science），不过这一词最早是由马尔萨斯在1798年提出。约翰·斯图尔特·密尔在1844年提出了一个以社会科学定义经济学的角度：
.....

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)

Google Drive下载或 百度云盘下载，密码:k265

数据描述

包含了250万篇新闻。新闻来源涵盖了6.3万个媒体，含标题、关键词、描述、正文。

数据集划分：数据去重并分成三个部分。训练集：243万；验证集：7.7万；测试集，数万，不提供下载。

可能的用途：

可以做为【通用中文语料】，训练【词向量】或做为【预训练】的语料；

也可以用于训练【标题生成】模型，或训练【关键词生成】模型（选关键词内容不同于标题的数据）；

亦可以通过新闻渠道区分出新闻的类型。

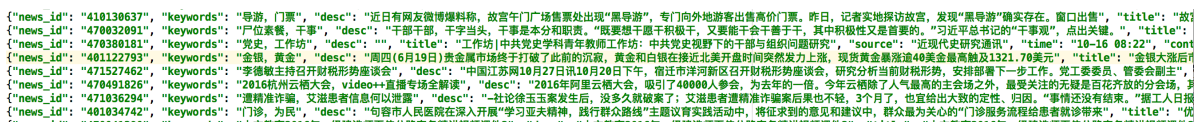
结构：

```
{'news_id': <news_id>,'title':<title>,'content':<content>,'source':
<source>,'time':<time>,'keywords': <keywords>,'desc': <desc>, 'desc': <desc>}
```

其中，title是新闻标题，content是正文，keywords是关键词，desc是描述，source是新闻的来源，time是发布时间

例子：


```
{"news_id": "610130831", "keywords": "导游,门票","title": "故宫淡季门票40元 “黑导游”卖外地客140元", "desc": "近日有网友微博爆料称，故宫午门广场售票处出现“黑导游”，专门向外地游客出售高价门票。昨日，记者实地探访故宫，发现“黑导游”确实存在。窗口出售", "source": "新华网", "time": "03-22 12:00", "content": "近日有网友微博爆料称，故宫午门广场售票处出现“黑导游”，专门向外地游客出售高价门票。昨日，记者实地探访故宫，发现“黑导游”确实存在。窗口出售40元的门票，被“黑导游”加价出售，最高加到140元。故宫方面表示，请游客务必通过正规渠道购买门票，避免上当受骗遭受损失。目前单笔门票购买流程不过几秒钟，耐心排队购票也不会等待太长时间。....再反弹”的态势，打击黑导游需要游客配合，通过正规渠道购买门票。"}
```

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)

Google Drive下载 或 百度云盘下载，密码:fu45

数据描述

含有150万个预先过滤过的、高质量问题和答案，每个问题属于一个类别。总共有492个类别，其中频率达到或超过10次的类别有434个。

数据集划分：数据去重并分成三个部分。训练集：142.5万；验证集：4.5万；测试集，数万，不提供下载。

可能的用途：

> 可以做为通用中文语料，训练词向量或做为预训练的语料；也可以用于构建百科类问答；其中类别信息比较有用，可以用于做监督训练，从而构建
>
> 更好句子表示的模型、句子相似性任务等。

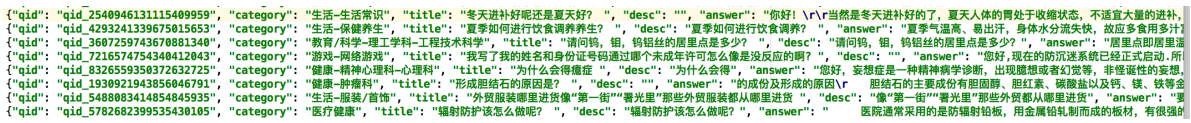
结构：

```
{"qid":<qid>,"category":<category>,"title":<title>,"desc":<desc>,"answer":
<answer>}
```

其中，category是问题的类型，title是问题的标题，desc是问题的描述，可以为空或与标题内容一致。

例子：


```
{"qid": "qid_2540946131115409959", "category": "生活知识", "title": "冬天进补好一些呢，还是夏天进步好啊？ ", "desc": "", "answer": "你好！\r\r当然是冬天进补好的了，夏天人体的胃处于收缩状态，不适宜大量的进补，所以我们有时候说：“夏天就要吃些清淡的，就是这个道理的。”\r\r不过，秋季进补要注意“四忌” 一忌多多益善。任何补药服用过量都有害。认为“多吃补药，有病治病，无病强身”是不的。过量进补会加重脾胃、肝脏负担。在夏季里，人们由于喝冷饮，常食冻品，多有脾胃功能减弱的现象，这时候如果突然大量进补，会骤然加重脾胃及肝脏的负担，使长期处于疲弱的消化器官难于承受，导致消化器官功能紊乱。 \r\r二忌以药代食。重药物轻食物的做法是不科学的，许多食物也是好的滋补品。如多吃荠菜可治疗高血压；多吃萝卜可健胃消食，顺气宽胸；多吃山药能补脾胃。日常食用的胡桃、芝麻、花生、红枣、扁豆等也是进补的佳品。\r\r三忌越贵越好。每个人的身体状况不同，因此与之相适应的补品也是不同的。价格昂贵的补品如燕窝、人参之类并非对每个人都适合。每种进补品都有一定的对象和适应症，应以实用有效为滋补原则，缺啥补啥。 \r\r四忌只补肉类。秋季适当食用牛羊肉进补效果好。但经过夏季后，由于脾胃尚未完全恢复到正常功能，因此过于油腻的食品不易消化吸收。另外，体内过多的脂类、糖类等物质堆积可能诱发心脑血管病。"}
```

{"qid": "qid_2540946131115409959", "category": "生活-生活常识", "title": "冬天进补好呢还是夏天好？ ", "desc": "", "answer": "你好！\r\r当然是冬天进补好的了，夏天人体的胃处于收缩状态，不适宜大量的进补，
{"qid": "qid_4293241339675015653", "category": "生活-保健养生", "title": "夏季如何进行饮食调养养生？ ", "desc": "夏季如何进行饮食调养？ ", "answer": "夏季气温高、易出汗，身体水分流失快，故应多食用多汁
{"qid": "qid_3607259743670881340", "category": "教育/科学-理工学科-工程技术科学", "title": "请问钨，钼，钨铝丝的居里点是多少？ ", "desc": "请问钨，钼，钨铝丝的居里点是多少？ ", "answer": "居里点即居里温
{"qid": "qid_721657475434041204 3", "category": "游戏-网络游戏", "title": "我写了我的姓名和身份证号码通过哪个未成年许可怎么像是没反应的啊？ ", "desc": "", "answer": "您好,现在的防沉迷系统已经正式启动.所
{"qid": "qid_8326559350372632725", "category": "健康-精神心理科-心理科", "title": "为什么会得癔症 ", "desc": "为什么会得", "answer": "您好，妄想症是一种精神病学诊断，出现臆想或者幻觉等，非怪诞性的妄想，
{"qid": "qid_1930921943856046791", "category": "健康-肿瘤科", "title": "形成胆结石的原因是？ ", "desc": "", "answer": "的成份及形成的原因\r　　胆结石的主要成份有胆固醇、胆红素、碳酸盐以及钙、镁、铁等金
{"qid": "qid_5488083414854845935", "category": "生活-服装/首饰", "title": "外贸服装哪里进货像“第一街”“曹光里”那些外贸服装都从哪里进货 ", "desc": "像“第一街”“曹光里”那些外贸都从哪里进货", "answer": "要
{"qid": "qid_5782682399535430105", "category": "医疗健康", "title": "辐射防护该怎么做呢？ ", "desc": "辐射防护该怎么做呢？ ", "answer": "　　医院通常采用的是防辐射铅板，用金属铅轧制而成的板材，有很强的

公开评测：

欢迎报告模型在验证集上的准确率。任务1： 类别预测。

报告包括：#1）验证集上准确率；#2）采用的模型、方法描述、运行方式，1页PDF；#3）可运行的源代码(可选)

基于#2和#3，我们会在测试集上做测试，并报告测试集上的准确率；只提供了#1和#2的队伍，验证集上的成绩依然可以被显示出来，但会被标记为未验证。

4.社区问答json版(webtext2019zh) ：大规模高质量数据集

410万个问答(过滤后数据3.7G，压缩文件1.7G；数据跨度：2015-2016年)

Google Drive下载

数据描述

含有410万个预先过滤过的、高质量问题和回复。每个问题属于一个【话题】，总共有2.8万个各式话题，话题包罗万象。

从1400万个原始问答中，筛选出至少获得3个点赞以上的的答案，代表了回复的内容比较不错或有趣，从而获得高质量的数据集。

除了对每个问题对应一个话题、问题的描述、一个或多个回复外，每个回复还带有点赞数、回复ID、回复者的标签。

数据集划分：数据去重并分成三个部分。训练集：412万；验证集：6.8万；测试集a：6.8万；测试集b，不提供下载。

可能的用途：

1）构建百科类问答：输入一个问题，构建检索系统得到一个回复或生产一个回复；或根据相关关键词从，社区问答库中筛选出你相关的领域数据

2）训练话题预测模型：输入一个问题(和或描述)，预测属于话题。

3）训练社区问答(cQA)系统：针对一问多答的场景，输入一个问题，找到最相关的问题，在这个基础上基于不同答案回复的质量、

问题与答案的相关性，找到最好的答案。

4）做为通用中文语料，做大模型预训练的语料或训练词向量。其中类别信息也比较有用，可以用于做监督训练，从而构建更好句子表示的模型、句子相似性任务等。

5）结合点赞数量这一额外信息，预测回复的受欢迎程度或训练答案评分系统。

结构：

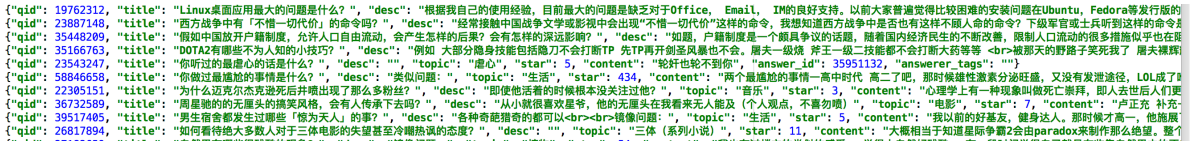
```
{"qid":<qid>,"title":<title>,"desc":<desc>,"topic":<topic>,"star":
<star>,"content":<content>,

"answer_id":<answer_id>,"answerer_tags":<answerer_tags>}
```

其中，qid是问题的id，title是问题的标题，desc是问题的描述，可以为空；topic是问题所属的话题，star是该回复的点赞个数，

content是回复的内容，answer_id是回复的ID,answerer_tags是回复者所携带的标签

例子：


```
{"qid": 65618973, "title": "AlphaGo只会下围棋吗？阿法狗能写小说吗？", "desc": "那么现在会不会有智能机器人能从事文学创作？<br>如果有，能写出什么水平的作品？", "topic": "机器人", "star": 3, "content": "AlphaGo只会下围棋，因为它的设计目的，架构，技术方案以及训练数据，都是围绕下围棋这个核心进行的。它在围棋领域的突破，证明了深度学习深度强化学习MCTS技术在围棋领域的有效性，并且取得了重大的PR效果。AlphaGo不会写小说，它是专用的，不会做跨出它领域的其它事情，比如语音识别，人脸识别，自动驾驶，写小说或者理解小说。如果要写小说，需要用到自然语言处理（NLP））中的自然语言生成技术，那是人工智能领域一个", "answer_id": 545576062, "answerer_tags": "人工智能@游戏业"}
```

```
{"qid": 19762312, "title": "Linux桌面应用最大的问题是什么？", "desc": "根据我自己的使用经验，目前最大的问题是缺乏对于Office， Email， IM的良好支持。以前大家普遍觉得比较困难的安装问题在Ubuntu，Fedora等发行版的
{"qid": 23887148, "title": "西方战争中有「不惜一切代价」的命令吗？", "desc": "经常接触中国战争文学或影视中会出现"不惜一切代价"这样的命令，我想知道西方战争中是否也有这样不顾人命的命令？下级军官或士兵听到这样的命令是
{"qid": 35448209, "title": "假如中国放开户籍制度，允许人口自由流动，会产生怎样的后果？会有怎样的深远影响？", "desc": "如题，户籍制度是一个颇具争议的话题，随着国内经济民生的不断改善，限制人口流动的很多措施似乎也在逐
{"qid": 35166763, "title": "DOTA2有哪些不为人知的小技巧？", "desc": "例如 大部分隐身技能包括隐刀不会打断TP 先TP再开剑圣风暴也不会。屠夫一级烧 斧王一级二技能都不会打断大药等等 <br>被那天的野路子笑死我了 屠夫裸辉
{"qid": 23543247, "title": "你听过的最虐心的话是什么？", "desc": "", "topic": "虐心", "star": 5, "content": "轮奸也轮不到你", "answer_id": 35951132, "answerer_tags": ""}
{"qid": 58846658, "title": "你做过最尴尬的事情是什么？", "desc": "类似问题：", "topic": "生活", "star": 434, "content": "两个最尴尬的事情一高中时代 高二了吧，那时候雄性激素分泌旺盛，又没有发泄途径，LOL成了
{"qid": 22305151, "title": "为什么迈克尔杰克逊死后并喷出现了那么多粉丝？", "desc": "即便他活着的时候根本没关注过他？", "topic": "音乐", "star": 3, "content": "心理学上有一种现象叫做死亡崇拜，即人去世后人们更
{"qid": 36732589, "title": "周星驰的无厘头的搞笑风格，会有人传承下去吗？", "desc": "从小就很喜欢星爷，他的无厘头在我看来无人能及（个人观点，不喜勿喷）", "topic": "电影", "star": 7, "content": "卢正充 补充
{"qid": 39517405, "title": "男生宿舍都发生过哪些「惊为天人」的事？", "desc": "各种奇葩猎奇的都可以<br><br>镜像问题：", "topic": "生活", "star": 5, "content": "我以前的好基友，健身达人。那时候才高一，他施展
{"qid": 26817894, "title": "如何看待绝大多数人对于三体电影的失望甚至冷嘲热讽的态度？", "desc": "", "topic": "三体（系列小说）", "star": 11, "content": "大概相当于知道星际争霸2会由paradox来制作那么绝望。整个
```


在该数据集上的公开评测和任务：

任务1： 话题预测。

报告包括：#1）验证集上准确率；#2）采用的模型、方法描述、运行方式，1页PDF；#3）可运行的源代码(可选)

基于#2和#3，我们会在测试集上做测试，并报告测试集上的准确率；只提供了#1和#2的队伍，验证集上的成绩依然可以被显示出来，但会被标记为未验证。

任务2：训练社区问答(cQA)系统。

要求：评价指标采用MAP，构建一个适合排序问题的测试集，并报告在该测试集上的效果。

任务3：使用该数据集（webtext2019zh)，参考OpenAI的GPT-2，训练中文的文本写作模型、测试在其他数据集上的zero-shot的效果，或测评语言模型的效果。

5.翻译语料(translation2019zh)

520万个中英文平行语料(原始数据1.1G，压缩文件596M)

Google Drive下载

数据描述

中英文平行语料520万对。每一个对，包含一个英文和对应的中文。中文或英文，多数情况是一句带标点符号的完整的话。

对于一个平行的中英文对，中文平均有36个字，英文平均有19个单词(单词如“she”)

数据集划分：数据去重并分成三个部分。训练集：516万；验证集：3.9万；测试集，数万，不提供下载。

可能的用途：

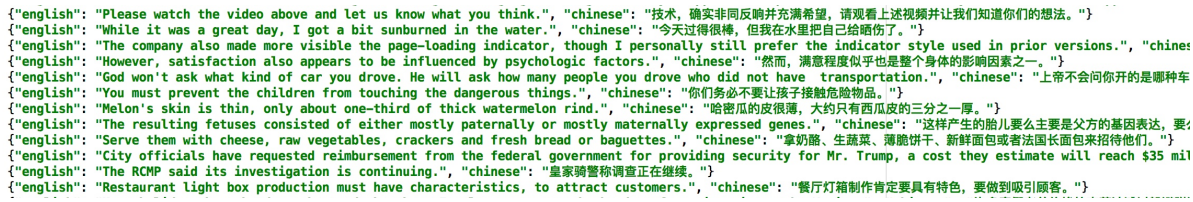
> 可以用于训练中英文翻译系统，从中文翻译到英文，或从英文翻译到中文；
>
> 由于有上百万的中文句子，可以只抽取中文的句子，做为通用中文语料，训练词向量或做为预训练的语料。英文任务也可以类似操作；

结构：

```
{"english": <english>, "chinese": <chinese>}

其中，english是英文句子，chinese是中文句子，中英文一一对应。
```

例子：

```
{"english": "In Italy, there is no real public pressure for a new, fairer tax system.", "chinese": "在意大利，公众不会真的向政府施压，要求实行新的、更公平的税收制度。"}
```



贡献语料/Contribution

贡献中文语料，请发送邮件至nlp_chinese_corpus@163.com

为了共同建立一个大规模开放共享的中文语料库，以促进中文自然语言处理领域的发展，凡提供语料并被采纳到该项目中，

除了会列出贡献者名单（可选）外，我们会根据语料的质量和量级，选出前20个同学，结合您的意愿，寄出键盘、鼠标、

显示屏、无线耳机、智能音箱或其他等值的物品，以表示对贡献者的感谢。

add your chinese corpus here by sending us an email

if there is any issue regarding the data, you can also contact with us, we will process it within one week.

thank you for your understanding.

项目贡献者或组织清单

1. ReactiveCJ

引用 Citation / How do I cite Us?

```
@misc{bright_xu_2019_3402023,
author       = {Bright Xu},
title        = {NLP Chinese Corpus: Large Scale Chinese Corpus for NLP },
month        = sep,
year         = 2019,
doi          = {10.5281/zenodo.3402023},
version      = {1.0},
publisher    = {Zenodo},
url          = {https://doi.org/10.5281/zenodo.3402023}
}
```

DOI 10.5281/zenodo.3402023

也请发邮件告知我们您的论文名称或在这个项目的数据集上的工作

Reference

1. 利用Python构建Wiki中文语料词向量模型试验
2. A tool for extracting plain text from Wikipedia dumps
3. Open Chinese convert (OpenCC) in pure Python:開放中文轉換
4. dumps of wiki, latest in chinese