

本项目旨在从各大论坛、贴吧、微博平台爬取数据构造高质量闲聊对话语料，用于聊天机器人训练语料。
主要包含天涯问答、可可英语、微博、贴吧数据

1、天涯问答

1.1、数据来源:<http://bbs.tianya.cn/list.jsp?item=907&sub=2> 天涯问答模块

默认	最新	精品	问答	版务	更多	版内搜索	🔍	屏
标题			作者			点击		
🔹 怎样通过一张照片找到一个人呢？			無边無际			4712		
☹ 荆州洛园万达有吗？？ 问			18271017889			0		
☹ 汽车平台哪家可信度高？ 问			kinkiliu5460			6		
☹ 喜欢的域名被人抢先买走了，想想还是很生气的 问			柚子不知海			861		

1.2、目标：爬取所有问答板块内容、回复、评论 由于天涯的帖子回复和评论会有互相嵌套（也就是有针对回复的评论、针对评论的再评论...，具体可参见：

<http://bbs.tianya.cn/post-907-63967-1-1.shtml#216819>

下的第四个林浦老者 2018-05-26 16:57:29 的回复；该评论的结果下面会呈现）本项目把所有的内容解析为一对一的“问答对”（即：帖子——第一级回复，回复——针对回复的评论，评论——评论的评论，评论——评论的回复...）

1.3、流程：先解析获得所有问答详情页面 url（回复数不足 1 的放弃）；再获取该页面所有内容并（一特殊分隔符区分帖子、回复及评论）；最后解析页面内容生成问答对

具体页面一如下分隔保存到文件待解析, 具体分隔方式如下：帖子作者@@ucontent@@帖子内容，回复 1@@block@@回复 2，评论作者@@ucomment@@（at 的作者 @@alt_ucomment@@评论正文），评论 1@@comment@@评论 2，回复@@concom@@回复下面的评论，帖子@@headblock@@所有回复。更详细见：tianya 问答/tainya_crawl.py/get_wenda_detail 函数

1.4 quick start:

```
cd tianya 问答
python get_urllist.py 1000 #option 1000 表示获取 1000 个页面的 url，生成文件
                                detail_url.txt #已经上传 detail_url.txt 文件，该步骤可以省略
python tainya_crawl.py #该程序获得爬取得数据并解析成问答对，一旦某个页面爬取出
                        错，直接跳过，继续下一个页面，生成 tianya.txt
```

1.5 原贴及结果截图

社会病了吗，还是我们根本就没有底线？

楼主：流云如水sakura 时间：2018-03-24 14:58:29 点击：54 回复：6

可能我还小，网络真的是言论自由的 社会，难怪这个社会舆论会把一个人杀的魂飞魄散，想想都觉得好笑，也会觉得背后发冷，我看到别人提到一个正常的问题，下面评论才真的是惊天地，泣鬼神，你要是看全了没准世界大战都爆发了。真的是搞不清出现在人的心里，是社会病了吗，还是我们原本早就无药可救。看到别人提出的一些问题，我现在都会自动忽略下面评论人的年龄，想想如果一个40多岁的人如果还是想一个喷子、水军一样回答着评论，想想都觉得都让你觉得恐怖，是社会蚕食了他们，还是他们原本就是养成这个社会的蛀虫？

分享 | 邀请回答 | 我要回答

4个回答

食用菌商务 2018-03-24 15:40:25
借用小说里的一句话，受尽世间冷眼，我自坦然面对。

流云如水sakura: 淡然处事 2018-03-24 16:17:55 评论

还可以输入140字 发表

林清名斋 2018-03-24 19:29:07
社会没有病，因为每个人的看法都不一样，所以就会有你所认为的，，，，， 稍安勿躁，，，，，

流云如水sakura: 这段时间又出了空姐遇害的事件，为她惋惜，可是事件的后续，我觉得无关于道德什么的了，说道德都觉得侮辱这两个字。 2018-03-24 09:49:35 评论

爬取及解析结果：

Wen:社会病了吗，还是我们根本就没有底线？ da:借用小说里的一句话：受尽世间冷眼，我自坦然面对。

Wen:借用小说里的一句话：受尽世间冷眼，我自坦然面对。 da:淡然处事

Wen:社会病了吗，还是我们根本就没有底线？ da:社会没有病，因为每个人的看法都不一样，所以就会有你所认为的.....

Wen:社会没有病，因为每个人的看法都不一样... da:这段时间又出了空姐遇害的事件...

Wen:社会病了吗，还是我们根本就没有底线？ da:改革开放是成功的，资源分配出头了，千万富翁多了，十万以下人也多了.....水污鱼少。

2、可可英语高质量问答对：

数据来自于 <http://talk.kekenet.com/> 所有情景对话问答对

见文件： 可可英语高质量问答对/jiaoyu_dialogue.txt

3、微博及评论

爬取单个用户的所有微博及其以下所有评论

Quick start:

1、下载 phantomJS, 并记住其安装路径，下载 selenium（主要一定要 2.x 版本，3.x 版本不兼容 phantomJS）

2、cd weibo 及所有评论

3、设置 conf.py 文件

4、Python weibo.py 生成文件 weibo_detail

4、tieba_欢乐斗问题反馈专区：

数据来源：<https://tieba.baidu.com/f?kw=%BB%B6%C0%D6%B6%B7%B5%D8%D6%F7&fr=ala0&tpl=5>

所有帖子标签及详细内容

Quick start:

cd tieba_欢乐斗问题反馈专区

python main.py