使用说明

一些重要的东西

由于涉及写入文件可能报毒。信任就好！！

仅限Windows系统！！手机退散！！Mac/Linux自行虚拟机或更新python安所有依赖库！！

本教程看似很长其实很短也很好懂！！阅读五分钟省您几小时！！如果您只关心基本用法，阅读完"一些重要的东西"一栏即可解决99%以上的问题。

用途

为了解决大量集中查询成语的需求，包括成语单OCR结果格式化，自动改错，批量查询成语或词语释义。可以进行几乎全自动的，速度飞快的批量操作。

可以查询（包括但不限于）汉字，词语，术语，成语。

用法

需要一个成语单和一个OCR（图片转文字）软件，把成语单的OCR结果存到word或记事本（这很重要！）。如果有成语单的文本也可以直接用。

然后运行exe文件，程序内附有相当完备的使用说明。

关于各种文件

首先查看文件后缀名。（参考这里）。

后缀为py和md的文件删了也没事，那是给开发者看的。

需要运行的是exe文件，用于储存数据的是自动生成的txt文件和data文件夹，这几个不能删，且必须在同一文件夹中。

exe文件根据操作系统留一个删一个

xp版是Windows通用的。但是查词逻辑相当落后，不支持自动改错，错误率较高。而且巨！慢！无！比！。Windows xp只能用这个版本。现已停止更新功能。

标准版只能用在Win7以上操作系统，速度飞快，支持自动改错。

关于两个txt文件

两个总集.txt收录了所有用IDerek查询过的成语和释义，可随意更改内容，但不可重命名，为了保存历史也尽量不要移动。

好用的OCR软件

白描（擅长打印体，自带扫描），讯飞输入法（打印手写体都擅长，建议先用其他软件（如白描或cs扫描王）扫描再识别）。

请尽量使用这两种OCR软件，这很重要。也请尽量拍清楚些。

权威性

你完全可以把它当成一个快捷版百度汉语。百度汉语权威性还是有保障的，至少比百度百科靠谱。对付高考还是绰绰有余。

没有采用某些现成的开源数据库，一是其中数据有很多错误，二是考虑到时效性。为了权威性和全面性还特意做了很多努力。

支持

如果这个项目帮到了你，去github给我个星标吧！

也可以扫这个赞助二维码请我喝杯咖啡哦~

不那么重要的东西

对OCR结果格式的要求

成语和成语之间有任意的非汉字字符（包括空格和换行）即可。

例如像下面这些都符合要求:

> 558.一无所获，364、二三其德
>
> 四体不勤/五谷不分 六神无主
>
> 七拼八凑djjdii妄九言十， ？'?啰嗦，

唯一不好弄的例子是下面这种:

> 荤荤大端耳濡目染声嘶力竭著书立说

或者这样:

> 眈眈逐逐，上行下效，迫在眉睫，定夺

（拿什么分隔都行就是别两个成语之间只用一个逗号！！那是标识八字成语用的！！）

这种可能需要你去程序界面按几十个enter才行……更好的选择是换一个能帮你排版的OCR软件。

原理

通过除逗号外的非汉字字符划分输入的OCR结果来格式化同时允许中间有逗号的成语的存在。

通过百度汉语的（伪）API接口抓取词条页面，百度汉语中能查到的视为正确。

通过在成语数据库中进行相似度比对实现自动改错。

爬取成语数据库的脚本在我的项目my-baiduhanyu中，还附带一个爬释义的。

自定义成语释义

自行改动user_data.json，此处不赘述json文本格式。

作者

反馈我都会回的，但是可能比较慢。

- Github：@This-username-is-available
- 邮箱：792405142@qq.com

新版会发布在这里。（打开慢属正常现象，且尽量不要用IE浏览器打开）

这里有一些可能的改进方向，如果确实需要可以附在反馈里。

- 支持带序号输出
- 支持成语之间空行

示例文本

> 435、亦步亦趋436、溢美之词437、蝇营狗苟438、越俎代厄439、振振有词
>
> 440、自命不凡?441、坐而论道442、敝帚自珍443、抛砖引玉444、无功受禄
>
> A45、激谢不地446、不列门墙447、信笔涂鸦448、德薄才疏449、趋之若條
>
> 450、一得之愚451、不足挂齿452、雕虫小技453、便短汲深454、东涂西

（这感人的识别率）

可能出现的问题

关于查询

包括但不限于:查到一半突然停止，输出的不是成语释义而是别的或者根本输不出来。

出现这种情况一般是百度汉语的（伪）API接口变动，或者界面源代码变动，或者网络请求超时，还有可能是某些神奇的词语造成的。请务必反馈！！

关于无法写入文件或者乱码

编码错误，把自带两个txt文件重新用ANSI或GBK编码保存即可。（参考这里）。