文晴曼-毕设代码文档说明

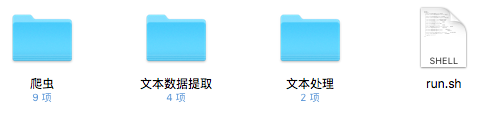
PS：Mac系统 Python 3.6版本，MongoDB数据库。

Mac系统对于文件夹里的内容处理与Windows有所不同，在遍历文件夹时有时先过滤了.Dsstore文件。在代码里有遍历文件夹的地方需要注意。

Python3.6与Python2有很大不同，本毕设全部用Python3.6完成。

由于使用到的Python Module有点多，只需根据提示，缺少什么就安装什么即可。其中有几项必须安装的有Scrapy , Jieba, lda,以及下文中提到的百度AI提供的SDK文件。

1. 在毕设代码目录下打开终端（命令行）运行命令： sh run.sh XX （XX为需要运行的项目名称，具体用记事本打开run.sh文件可以看到对应的项目）

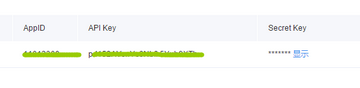


1. 爬虫文件夹下为爬取不同网站的爬虫代码，具体的可以参照每个文件夹下的readme.txt，如果没有readme.txt可以直接查看代码中的注释。爬虫用到的数据库为MongDB，需要先安装。有的爬虫用到了Scrapy框架和Selenium，也需要自行先安装。（运行代码的时候如果差什么配置文件就安装什么）
2. 文本数据提取中，首先需要下载百度AI平台的SDK文件到本地，参照百度AI平台网页中的说明

<http://ai.baidu.com/docs#/ImageClassify-Python-SDK/top>

另外，需要创建百度AI平台应用。具体步骤参照以下网址。

<https://blog.csdn.net/shuihupo/article/details/79866439>.



在代码中填写自己的这三个参数。

4. 文本处理的文件夹里面没有readme.txt的可以直接参照代码注释。在运行过程中提示缺什么module就自行安装。