可复现说明

2017211106 班 2017212116 杨诺诚

所有文件已同步到 https://github.com/YangNuoCheng/NLP_Keywords

Python 环境和库版本

实验使用的是 python 3.7.6

使用到额外的库有

库名称	版本号
Jiagu	0.2.3
Pandas	1.0.1
Jieba	0.42.1
Selenium	3.141.0
BeautifulSoup	4.8.2
Gensim	3.8.0
Sklearn	0.22.1

此外爬虫使用的是 selenium 在 GoogleChrome 下的无头浏览器，需要下载对应版本的 chromedriver.exe 放置在脚本同一目录下。

chromedrive 可以在 http://npm.taobao.org/mirrors/chromedriver 找到。

项目是在 spyder 中创建的，可以在 spyder 中打开这个工程文件。

主要脚本文件使用方法

文件中的脚本按照运行次序排序可得：

①getData.py

②jiagu.py

②LSA_LSI_LDA.py

②TextRank.py

②Tfldf.py

③Analysis_Draw.py

步骤①中的 getData.py 用于爬取网站的信息，并将结果保存在/data 文件夹下的 data_sample.csv 和 data_sample.txt，这两个文件中包含文章名称、摘要正文以及作者给出的标准关键词，并生成一个 keywords&url.csv 用于保存中间步骤的结果，这个文件中有文章名称、标准关键词和文章所在的 URL 地址。

步骤②中的各个算法文件单独运行，都会在/result 中生成 keys_算法名.csv 的结果文件，如 keys_Jiagu.csv。但这一步需要在/data 目录下有 data_sample.csv 文件。

步骤③中的 Analysis_Draw.py 用于分析/result 中的结果文件，先做预处理再计算各种参数，参数结果会保存在/result 中的 analysis.csv 中，之后程序会绘制一些图像，保存在/pictures

中。

主要文件夹

项目包括 data、result 以及 pictures 文件夹，/data 用于存储原始数据、各类停顿词(stopword.txt)，训练语料(corpus.txt)以及爬虫中间保存的数据；/result 用于存储各个算法运行结果和分析的结果；/pictures 用于存储各类分析生成的图片。

文件夹树结构

```
C:\USERS\18801\DESKTOP\NLP_KEYWORDS\NLP_KEYWORDS
|   Analysis_Draw.py
|   chromedriver.exe
|   debug.log
|   getData.py
|   jiagu.py
|   LSA_LSI_LDA.py
|   TextRank.py
|   TfIdf.py
|
├─.spyproject
|   └─config
|           |   codestyle.ini
|           |   encoding.ini
|           |   vcs.ini
|           |   workspace.ini
|           |
|           ├─backups
|           |       codestyle.ini.bak
|           |       encoding.ini.bak
|           |       vcs.ini.bak
|           |       workspace.ini.bak
|           |
|           └─defaults
|                   defaults-codestyle-0.2.0.ini
|                   defaults-encoding-0.2.0.ini
|                   defaults-vcs-0.2.0.ini
|                   defaults-workspace-0.2.0.ini
|
├─data
|       corpus.txt
|       data_sample.csv
```

```
|       data_sample.txt
|       keywords&url.csv
|       stopword.txt
|
├─pictures
|       Acc_Dim.png
|       Acc_Pre_Rate.png
|       BestAcc.png
|       Pre_Dim.png
|       句子个数的分布.png
|       数据集词数分布.png
|       运行时长.png
|       运行时长 2.png
|
├─result
|       analysis.csv
|       analysis.txt
|       keys_Jiagu.csv
|       keys_LDA.csv
|       keys_LSI.csv
|       keys_TextRank.csv
|       keys_TFIDF.csv
|       keys_TFIDF.txt
|
└─__pycache__
        jiagu.cpython-37.pyc
```