代码说明

汪楚文 2018202114 05/05/2020

目录结构

```
MovieRP/
|-- MovieRP/
|    |-- __init__.py
|    |-- items.py
|    |-- middlewares.py
|    |-- pipelines.py
|    |-- settings.py
|    |-- __pycache__
|    |    |-- __init__.cpython-37.pyc
|    |    |-- items.cpython-37.pyc
|    |    |-- pipelines.cpython-37.pyc
|    |    |-- settings.cpython-37.pyc
|    |-- spiders
|    |    |-- __init__.py
|    |    |-- __pycache__
|    |    |    |-- __init__.cpython-37.pyc
|    |    |    |-- movie.cpython-37.pyc
|    |    |-- movie.py
|
|-- dict/
|    |-- comment.txt
|    |-- ex.txt
|    |-- cut_str.txt
|    |-- Typeface.ttf
|    |-- back.jpg
|    |-- cloud.png
|
|-- scrapy.cfg
|-- 代码说明.docx
|-- 代码说明.pdf
```

scrapy工程详情

一.spider文件movie.py

name：movie，通过scrapy crawl movie 执行

代码解释：通过xpath把符合正则表达式的信息分别存入item['name']和item['content']中，并返回item。设置爬虫爬取页数spider_end = 3， 爬虫页数控制及末页控制if self.count < self.spider_end: self.count = self.count + 1，当未爬取到目标页数时，爬取下一页 yield scrapy.Request(nextPage, callback=self.parse);爬取结束时爬虫退出。

二.items.py

```
class MovierpItem(scrapy.Item):
    content = scrapy.Field()
    name = scrapy.Field()
```

name用来存取爬取到的评论者id，content用来存取具体评论

三.settings.py

```
BOT_NAME = 'MovieRP'

SPIDER_MODULES = ['MovieRP.spiders']
NEWSPIDER_MODULE = 'MovieRP.spiders'
HTTPERROR_ALLOWED_CODES = [403]
DOWNLOAD_DELAY = 1.5
ROBOTSTXT_OBEY = False
```

上述settings定义spider爬取数据的方式，防止数据爬取受阻

四.__init.py__和middlewares.py

保持默认

五.pipelines.py

[功能]

将iteam存取的内容进行处理并存入相应文件

[代码实现方法]

(1)爬到的内容存取到comment.txt

```
content = jieba.lcut(item['content'])
name = item['name']
self.file.write('['+ name +']' + ':' + " ".join(content)+'\r\n')
```

(2)ex.txt用来设置被排除的个别词语

```
jieba.analyse.set_stop_words("./dict/ex.txt")
```

(3)从content中统计出的词频序列，存入cut_str.txt，用于构造词云图

```
tags = jieba.analyse.extract_tags(content, topK=1000, withWeight=True)
for v, n in tags:
   # 将词频记录下来
    data_str = v + '\t' + str(int(n * 10000)) + '\n'
    file_object.write(data_str）
```

(4)构造词云图

```
wc = WordCloud(font_path="./dict/Typeface.ttf",background_color="white", max_words=500, mask=alice_coloring,stopwords=stopwords, max_font_size=40, random_state=42)
```

六./dict 目录

里面有爬取到的comments.txt，排除的词ex.txt，统计出的词频序列cut_str.txt
生成词云图的背景图back.jpg,字体Typeface.ttf,以及生成的词云图cloud.png

七.scrapy.cfg

保持默认

附

爬取到的评论文件comment.txt放在/dict下

超链接：点击此处可在github中查看该项目