EverythinNKU

主要实现功能

- 网页爬取
- 倒排索引构建
- 利用PageRank将结果排序进一步优化
- 自己编写的VSM排序查询结果

web crawler

查看源码

1. 将南开大学官网作为种子url
2. 根据种子url，自己编写程序进行网页爬取，将该url加入used url列表， 将网页中的链接加入到unused url列表，防止重复爬取。
3. 将网页中重要的信息（一般包含在a标签下），比如标题、锚文本、相对链接等保存为数据库文件供建立索引。 将文本文件根据相应的映射存储为txt文件

做的优化

1. 多线程
2. 锁机制
3. 爬取日志

index

查看源码

以BSBI为基础，在parse block的同时进行自定义目标数据结构的维护，并用pickle、json等包的dump函数将其保存到磁盘中。

下次使用时，再利用load函数将其加载到内存。

类内主要多维护了这几个属性：

```
self.term_tf_docid = defaultdict(lambda: [0, set()])
self.term_idf = defaultdict(float)
self.docid_terms_tfs = defaultdict(lambda: defaultdict(int))
```

分别用来存储词向的倒排索引及其词频、词向的逆文档频率、文档内词向的词频

PageRank

查看源码

1. 根据爬取的链接间的关系，构建有向图
2. 利用networkx进行PageRank计算

3. 将结果保存到磁盘中

VSM

查看源码

1. 继承自BSBIIndex，主要将之前计算出来的tf、idf进行最后的相乘求和并归一化， 并扩展了对外的查询接口，能够提供文档查询个性化查询等服务。
2. 此外，会对每位用户的查询进行日志记录，便于提供更好地个性化查询。现阶段的个性化查询比较简单粗暴， 日后会对其算法进行进一步的改进。

遇到的问题

网页爬取内容一直会出现各种各样的问题

这是一个不断发现问题并解决问题的过程，现在我已经爬取的数据中还是会有很多没有考虑到的问题。日后会进一步改进。

对Python语言的掌握度不够，发生错误但是一直没有发现

```
uniq_urls = [i for i in set(new_urls) if i not in unused_url and used_url]
```

该语句并不能将i排除在两个列表之外，而是只能将i排除在第一个列表之外。

此bug可能是导致我数据库内容大量重复的原因之一。

修改为

```
uniq_urls = [i for i in set(new_urls) if i not in unused_url and i not in
used_url]
```

避免死锁

编写acquire函数，给锁进行相应的编号，使得两个锁的获得必定是有序的

相对链接地址

```
<div class="text-muted description">
    <p class="text"><a href='/2020/1114/c19665a317736/page.htm' target='_blank'
title='金融学院召开课程思政建设工作推动会'>（通讯员：徐静）为进一步贯彻落实《南开大学课程思政建设实施方案》和南开大学课...</a></p>
</div>
```