Search Engine 实验报告

学号-20142112280-谢非

2016 年 12 月 28 日

目录

1 实验目的

1. 应用信息提取的基本原理来实现一个简单的搜索引擎

2 实验要求

1. 实现一个信息提取系统（搜索引擎）
2. 实现引擎的模糊搜索和全文检索功能

3 实验环境

实验代码使用 python3.5 实现. 测试数据是 `http://www.cs.cmu.edu/afs/cs/project/theo-11/www/naive-bayes.html`这个网站上的 20 个新闻组资料, 即之前文本分类的 20 个新闻组资源. 不过实际试运行时只选取了测试数据中的 100 个文本数据.

4 实验原理

4.1 创建索引

首先将所有文本数据去除标点符号以及停用词[1]剔除，之后将每个文本转化为一个单词列表. 然后，记录每个单词出现的索引位置. 实现时采用字典存储，单词是键，值是其出现的索引位置的列表. 在处理完所有的单个文件之后，再次扫描所有的索引[2]，建立全局的倒排索引表，键是单词，值是一个字典（该字典内，键是文件名，值是该单词在文件中的索引列表）.

4.2 实现查找

具体查找时需要对输入的字符串进行分词，对所有的单词，在全局索引表中获取该单词对应的文件和文件内位置. 之后根据文件内总体的关键字出现次数，来计算出该文件的权重，由权重对文件进行重排. 重排后根据第一个索引的位置构造出该单词在文件中出现的大体位置（相邻的字符串）返回结果. 运行样例如下，

```
indexData.getStopWords("stopWords.txt")
allIndexs = indexData.createIndex()
allIndexs.loadAllFiles(["./data/"])
queryTool=query.QueryStr(allIndexs.allIndex,allIndexs.allDocs)
queryTool.query("i am good")
```

返回结果如下（包括该文件以及关键词出现的位置）:

```
文件及文件内关键字符串提取如下：
./data/39078  ...a truly lossless compression algorithm i e one that guarantees......avoided or limited but I am not optimistic about it.....compress images 5 Wh
./data/38853  ...model2 406 vertices 296 useful i e referred to in......hal physics wayne edu I am writing my own visualization......24 bit scanning Get a good 24
./data/38375  ...some stuff for scientific graphs i e log axes free.....tool and it is very good for this purpose Dimple...
./data/37930  ...familiar with graphics and I am not expecting for any......know if there are any good 2d graphics packages available...
./data/38856  ...lines just a hair backwards i e smaller VRC Z...
./data/38921  ...lines just a hair backwards i e smaller VRC Z...
./data/38606  ...on 32768 colors when I am actually in 1024x768x65000 color...
./data/38571  ...Lines 15 Hi there Well i have a 386 40...
./data/38753  ...free information for the common good Pat Cadigan _Mindplayers_ an...
./data/39664  ...vu nl NNTP Posting Host am ucsc edu hed cats.....an alt group Wouter Very good point Is there someone...
./data/38700  ...sun ac za writes I am also looking for a.....chessboard Right here is as good a place as any...
./data/38370  ...Rourke Subject Re Need a good concave convex polygon algorithm...
./data/38470  ...got a line on a good book to help answer...
./data/38904  ...acct writes Hi there I am interested in facial animation...
./data/37942  ...edu Lines 9 Hi I am in immediate need for...
./data/37947  ...comp graphics animation Lines 3 i am sorry but this.....graphics animation Lines 3 i am sorry but this genoa...
./data/39049  ...render into the same image i e Bezier curves to.....do this In fact I am interested in sources for...
./data/38459  ...to manipulate postscript files I am specifically interested in drawing...
./data/38980  ...programs never seen such in good view and are not...
./data/38942  ...of our results so any good data that we will...
./data/38421  ...32 GMT Lines 14 I am looking for some graphic...
./data/38907  ...a flame or advertisement Where am I Who turned off...
./data/38998  ...37 GMT Lines 17 I am revamping some computer aided...
./data/38699  ...xxxx xxxx xxxx xxxx The good thing about this perspective...
./data/38750  ...48 EST Lines 8 I am scanning in a color...
./data/39000  ...plugged in at the time i e not using the...
./data/38683  ...be appreciated Thanks A very good modeling package I found...
./data/38473  ...animation unless you had a good Amiga animation program Otherwise...
./data/38603  ...files It took me a good 20 minutes to start...
./data/38577  ...your disk back This is good for people who don...
./data/38251  ...B Subject Re Fractals what good are they Message ID...
./data/38466  ...proposed newsgroup split I personally am not in favor of...
./data/38402  ...about this chip Yes I am very interested in this...
```

[1]停用词采用的是网上查找的停用词

[2]写实验报告的时候发现此处实际上只需要扫描统计一次就够了

5 实验总结

本次实验过程中我对信息提取技术和搜索引擎的具体实现有了更深刻的理解和认识，深切体会到为什么搜索领域的整体思路很简单，就是建立索引然后查找．但是具体实现的时候有无数的实现细节和需要考虑的地方需要我们注意和学习．因为实现时时间比较紧张，许多细节没有考虑（也同时感受到需要考虑的细节很多很多），希望之后还可以改进这个很原始的搜搜索引擎.