清华新闻网信息检索系统

李泽龙

计54

2015011321

1. 基本功能要求的实现
2. 爬取清华新闻网中的新闻信息：通过crawler.py程序实现，程序中主要通过MyHTMLParser、ListName两个分别继承于HTMLParser、SGMLParser的类，来爬取清华新闻网上，news子域名下的网页，并将新闻页面中的url、标题、正文、时间通过正则表达式分离，导出至news\_201609101625.txt，导出时，出于对之后处理数据便捷性的考虑，在导出数据之间加入了一些分隔符；
3. 处理爬取页面信息：主要通过mycopy.py和list\_index.py两个程序实现。
   1. 为每个新闻页面赋予一个ID:爬取网页部分采用的是广度优先搜索（BFS）的思想，赋予ID的采用的是访问爬取网页的时间戳，即最先爬取到的新闻页面的ID为1，最后大约爬取了近50000个页面；
   2. 抽取网页关键主体内容：用正则表达式剥离有效信息，同时用replace函数，将从网页上爬取下来的内容的多余的换行符删去，以便后用；
   3. 对正文和标题进行分词及对每个分词建立倒排列表：分词采用jieba中的lcut\_for\_search函数可以得到分词list，得到分词之后通过将list转为set再转为list，从而获得无重复的分词表，对于一段文本，即正文和标题，立刻进行分词，将分到的词和ID立刻存入word\_dict中，word\_dict以词为key，以ID构成的list为value，在此过程中，由于先操作ID较小的网页，所以倒排列表的要求 也随即完成；
   4. 导出分词结果：通过json转码解码，并通过存取二进制文件，从而将数据量非常大的word\_dict导出到index.pkl中，为之后编写网页搜索操作提供静态数据库。
4. 实现网页：在/form/templates/form下编写index.html和result.html，实现检索系统前端，建立Django模型。在/form/view.py中，在执行处理输入搜索框的文字，进行分词，对数据库的查询，并显示搜索结果页面。
5. 学习并使用CSS框架，美化页面：编写了简单的/form/static/form/style.css文件对index.html和result.html页面进行外观美化，并在.html文件中进行配置