美食搜索引擎——设计文档

计45 邵韵秋

1. 基本介绍

本美食搜索引擎支持搜索维基上的所有关于蛋糕和沙拉的网页，在搜索框中输入关键词，打开相应链接，便可搜到维基百科上相应的网页。

1. 设计思路及实现方法

该作业的实现主要分为从爬取网页、提取关键词、建立到排列表、web搜索、页面美化这几个部分。

主要使用了python和Django来实现，爬取网页时使用HTMLParser解析根网页代码，爬取相关网页。（由tol.py,cake.py,salad.py实现，网页分别放在了cake 和salad文件夹中）

提取关键词时，使用了Beautifulsoup和字典以及集合的数据结构，来提取关键词并且去重，将关键词的倒排列表放在keywords.py中几个字典中。

在web部分，使用Django来编辑搜索框，根据用户输入，使用正则表达式在关键词的倒排列表中搜索，并且显示相应页面，同时还实现了分页功能。

美化界面时，使用了css，美化搜索框，以及搜索信息的表格。

此外，还实现了父类搜索，在刚进入页面时有两个选项，search for cake 和 search for salad,如果点选进去，得到的信息会在相应的分类下，而不会互相串起来。

在Django的templates文件夹中，分别放置了模板，搜索功能主要通过foodsearch文件夹中的search.py实现。

1. 总结与思考

Python在爬取网页方面的确功能强大，在此次大作业的过程中，熟悉了python的基本语法以及一些用来解析网页的方法，同时学习了基本的django和css样式表的使用方法，感觉收获很大。

然而此次大作业中仍存在编码问题尚未解决，通过查阅相关资料和咨询同学，已经将编码转成utf-8，但在网页显示概要内容时还是会出现一些乱码，这个问题尚未解决。

备注：keywords中的读取文件使用的是绝对路径，在其他电脑上运行时，需要改一下路径。

2015年9月13日星期日