NEWSMAP

关闭报告

小组成员：杨天健、刘华秋、张振宇

2017.6

一、功能实现的难度与工作量

|  |  |  |  |
| --- | --- | --- | --- |
| 功能 | 完成人 | 代码行数 | 难度【注】 |
| 爬虫部分（新浪、搜狐）  csv网站元数据解析 | 杨天健 | 330 | 0 |
| 索引及分词部分的java实现 | 张振宇 | 200 | 1 |
| Redis数据库的探索 | 张振宇 | 100 | 1 |
| 索引评分公式建模 | 张振宇 | - | 3 |
| 环境的配置，相关参数的配置 | 杨天健 | - | 2 |
| Web后台 | 刘华秋 | 430 | 0 |
| Web前端（基于e-chart） | 刘华秋 | 400 | 2 |
| 索引部分Java->Python的移植 | 杨天健 | 110 | 2 |
| Lxml + cssselector的探究 | 杨天健 | - | 1 |
| 合计 |  | 1530 |  |

【注】难度级别：

0 现成库函数调用

1 涉及到工程思考

2 涉及到代码内层思考

3 涉及到数学、调度知识

二、项目完成情况

|  |  |
| --- | --- |
| 功能 | 完成情况 |
| 爬虫 | 100% |
| 索引 | 100% |
| Web前后端 | 100% |

三、经验及教训

本次项目让我们最值得欣喜的地方是web界面的展示和接口清晰的思路。前者使软件更加直观、易用、跨平台，后者保证小组三人可以并行工作，一人对代码的修改不影响另一个人（只要接口不变），使得我们对于程序的稳定性、安全性可以各司其职地优化。

本次项目也有一些点值得我们深入学习。对于二维表格型的数据库，mysql无法发挥优势。前两星期我们才开始对mangodb等nosql数据库做详细了解，没有来得及配置，因为那时数据已经很多了，换数据库会带来较大的时间成本。

由于时间原因，基于lxml的更专业版本的爬虫仍在实验中。作为外国人开发的软件，lxml的cssselector出现了编码问题，并且不支持自定义标签，这使得抓取关键部分不得不以正则表达式做匹配（beautifulsoup由于效率问题一般不使用）。