实验报告——文档过滤与信息提取

161220179 周科

本次实验我需要对于从vscode网站中爬取到的816个issue文件进行文档过滤。可以看到的是，文档中各类url以及id信息属于上传者的个人信息部分，对本地实验是无用信息。本次实验的数据支持主要是由issue的“title”栏和“body”栏的具体内容构成。

由于文档中的title和body的信息存储格式均为"title": "…"和"body": "…"类型，故而选择正则表达式匹配来处理数据。构造出的正则表达式为“title/body”:\s”.\*”,利用regex\_search函数逐个处理这816个文档，将匹配到的title和body的具体内容输出到title.txt和body.txt这两个文件中。

主要部分代码如下：

smatch mat1;

regex pattern1("\"title\":\\s\"(.\*)\"");

string::const\_iterator start1 = allTitle.begin();

string::const\_iterator end1 = allTitle.end();

while (regex\_search(start1, end1, mat1, pattern1))

{

//cout << "1" << endl;

string msg(mat1[1].first, mat1[1].second);

fout1 << msg << endl;

start1 = mat1[0].second;

}