**武汉大学计算机学院**

**本科生课程设计报告**

**信息系统导论实验二**

专 业 名 称 ：软件工程

课 程 名 称 ：信息系统导论

指 导 教 师 ：王峰 副教授

学 生 学 号 ：2016301040050

学 生 姓 名 ：吕若凡

二○一八年十一月

**郑 重 声 明**

本人呈交的设计报告，是在指导老师的指导下，独立进行实验工作所取得的成果，所有数据、图片资料真实可靠。尽我所知，除文中已经注明引用的内容外，本设计报告不包含他人享有著作权的内容。对本设计报告做出贡献的其他个人和集体，均已在文中以明确的方式标明。本设计报告的知识产权归属于培养单位。

本人签名： 日期：

摘 要

信息系统导论实验二的实验目的是通过设计一个网络爬虫爬取知乎页面并用结巴分词库对内容进行分词整理统计词性和词频来熟悉python的语法和网络编程的基本内容。

实验设计主要遵循是python的基本语法和网络编程的基本原理。

实验内容主要包括：设计一个网络爬虫爬取知乎页面并用结巴分词库对内容进行分词整理统计词性和词频。

实验结论为在实现基本要求的基础之上，使用两种方法（使用selenium库模拟浏览器和使用re库发送http请求）爬取网页源代码，然后得到文本内容，根据词性分词，得到形容词，名词，动词和字母词的词频。

**关键词：** python；网络编程；爬虫；jieba

目录

[1.实验目的和意义](#_Toc9499_WPSOffice_Level1) [5](#_Toc9499_WPSOffice_Level1)

[2．实验方案](#_Toc6641_WPSOffice_Level1) [5](#_Toc6641_WPSOffice_Level1)

[3．结论与结果](#_Toc6649_WPSOffice_Level1) [10](#_Toc6649_WPSOffice_Level1)

[4．小结](#_Toc20764_WPSOffice_Level1) [10](#_Toc20764_WPSOffice_Level1)

[参考文献](#_Toc4783_WPSOffice_Level1) [10](#_Toc4783_WPSOffice_Level1)

**1.实验目的和意义**

**1.1实验目的**

本实验通过设计一个网络爬虫爬取知乎页面并用结巴分词库对内容进行分词整理统计词性和词频来熟悉python的语法和网络编程的基本内容。

**1.2实验意义**

对python运用更加熟练，了解了网络的工作原理，增加了对数据统计和数据分析的兴趣。深入了解form表单的提交，网页请求头的格式，网页数据传递方式，对网络编程更加得心应手。

**2．实验方案**

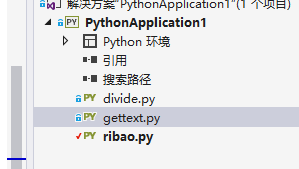
**2.1实验环境**

**2.1.1所用软件**

**Vs2017；chrome浏览器；chrome driver。**

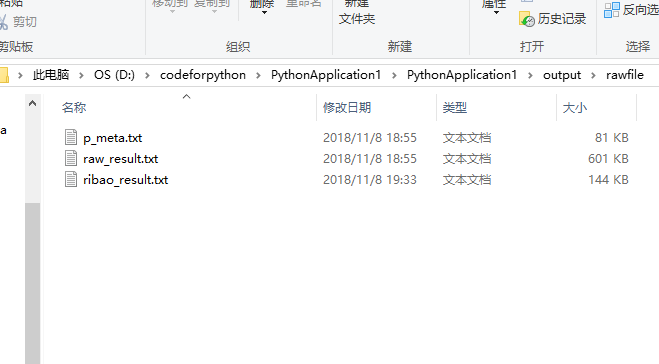
**2.1.2所用系统**

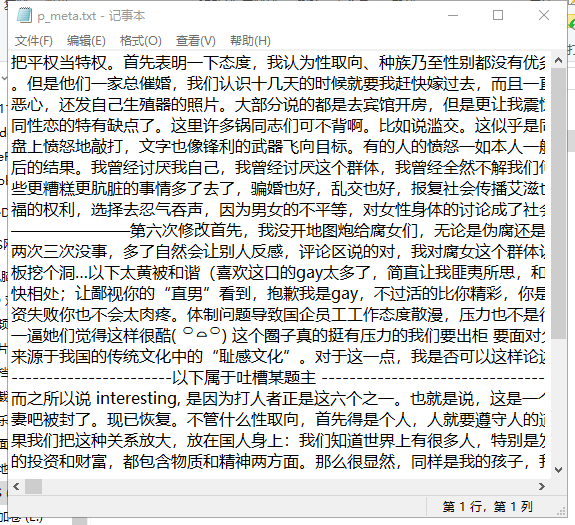
**Windows 10**

**2.2方案设计   
 2.2.1文件结构** 

可以看到整个项目有三个文件：divide.py，gettext.py和ribao.py

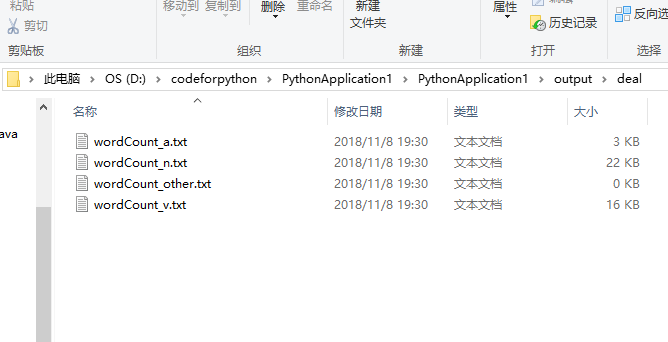
Gettext.py和ribao.py是分别从网页中获取源代码并保存为本地的文 本文件，再将源代码文本文件中的文本内容提取得到文本文件。而divide.py 是将得到的文本文件按照名词，动词，形容词和字母词分类统计词频。 **2.2.2从网站得到源代码** 使用两种方法得到网页源代码：

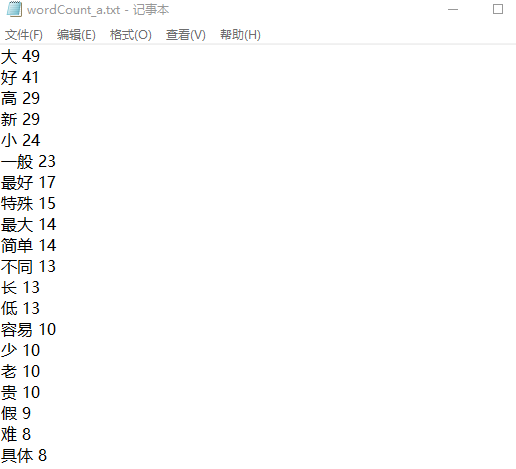
1.使用selenium库，模拟真实的网页运行来得到完整的网页源代码和文本信息。由于知乎页面采取异步加载的方法，普通的re请求得到的网页源代码只有一小部分，故使用selenium包，并下载chrome driver驱动，打开知乎网页，自动下拉使网页加载。并将得到的源代码文本保存在./output/rawfile/raw\_result.txt中。然后使用buautifulsoup4库直接通过result\_soup.find\_all()函数得到p标签之间的内容即纯文本内容保存到./output/rawfile/p\_meta.txt中。至此知乎回答内容已经得到。  
 2.使用re库,用python发送http请求,得到知乎日报的源代码，然后对知乎日报的源代码进行解析，得到每一篇文章的链接，再向各个链接发送请求得到其源代码，这几份源代码是我们需要的。然后再使用正则表达式得到其中的文本内容，并保存在./output/rawfile/ribao\_result.txt中。至此，所有的知乎日报文本内容已经得到。  



**2.2.3jieba分词**

使用结巴库里的函数可以很轻松的将一个完整的句子切割成很多词语并得到他们的词性。.flag是其词性，.word是其词本身。利用词典和列表两种数据形式可以循环遍历得到相同词的词频。而字母词在结巴库中并没有定义，所以需要自己定义，设计在jieba词库中添加所需单词，然后当遇到这个词之后把他加入到字母词文件中。按照词性的不同将词语分到四个文件：wordCount\_a.txt wordCount\_v.txt wordCount\_n.txt wordCount\_other.txt





**3．结论与结果**

项目中的三个程序可以完成老师的任务，并在其基础上做更多的事情。但各有优缺点，其中使用selenium库进行爬虫的代码文件，必须在装有chrome和chrome driver的电脑上运行，并且加载方式笨拙，所用时间较长。但优点是可以加载得到所有的（大部分）的答案。使用re库进行爬取的代码，优点是对环境要求低，可以在python环境下直接运行，请求方式优雅，请求速度快。缺点是无法异步加载。所以采取了发散式爬取，对于知乎日报页面来说，所有的内容都是规范的布置，所以可以对其链接进行再次发送访问请求，得到文章的真正内容。

**4．小结**

通过该实验对python运用更加熟练，了解了网络的工作原理，增加了对数据统计和数据分析的兴趣。深入了解form表单的提交，网页请求头的格式，网页数据传递方式，对网络编程更加得心应手。而且了解了get,post请求的区别，参数传递的位置，懂得了如何对网页进行调试。

**参考文献**

[1] 赫特兰(MagnusLieHetland,《Python基础教程》,人民邮电出版社,2010-6.