实验四 爬取天气预报数据

4.1 实验介绍

本实验通过对中国天气网（www.weather.com.cn）分析，采集某个城市天气及和该城市所有关联网页，并存储。

4.2 实验目标

了解 HTML 文档结构

了解深度优先算法和广度优先算法

掌握网站遍历和数据采集方法

掌握 BeautifulSoup 用法

4.3 实验原理与方法

HTML 是超文本标记语言（Hyper Text Markup Language），是一种用于创建网页的标记语言。HTML 语言编写的文件后缀名为.html 或者.htm，文件运行在浏览器上，由浏览器解释代码。用 HTML 编写网页，就是用标签来标记文本信息。常用的标签有 html、title、body、script、style、p、a、li、table、tr、td、input、button 等等，不同的标签对应不同的文本组织形式。每种标签有着各种属性，常见的属性有 class、id、name、value、href、style 等等，用于区别同种标签，使得同种标签也有不同的表达。

BeautifulSoup 库介绍

首先使用 pip 工具安装 BeautifulSoup 库：pip install beautifulsoup4

导入 BeautifulSoup 库：from bs4 import BeautifulSoup

BeautifulSoup 库用于管理标签树，因此并不限于 html 代码，xml 代码一样可以用该库来处理，这取决于使用的解释器。给定网页源代码和解释器后，就可以得到一个管理 html 标签树的对象，解释器一般使用 html.parse。通过该对象可以访问任意标签（Tag）、标签里的内容（NavigableString）和多种属性（Attributes，Name，Comment）。也可以通过访问标签的子标签（children，contents）、子孙标签（descendants）、父标签（parent，parents）、兄弟标签（next_sibling，previous_sibling，next_siblings，previous_siblings）来实现标签树的上行遍

历、下行遍历和平行遍历，注意，标签树的平行遍历是限于同一个父标签下的子标签，和二叉树的层次遍历有一定区别。

也可以使用 find_all()方法获取指定的标签内容。其中参数包括 name（标签名称检索）、attrs（标签属性值）检索、recursive（是否对子孙标签全部检索）、string（标签中的字符串内容检索），这个函数会以列表形式返回符合要求的内容。另外，find 方法和 find_all 方法使用方法相同，但是只返回第一个符合要求的内容。

详细学习 BeautifulSoup 库的使用方法请参考官方文档。

4.4 实验步骤

1. 浏览中国天气网，查看网页源代码，分析网页结构。
2. 定位目标数据，分析目标数据在网页中的特征。
3. 学习 BeautifulSoup 库，并编写 Python 脚本，提取目标数据
4. 编写 Python 脚本，将数据存入数据库中。

4.5 实验要求

1. 简单分析网页源代码。
2. 编写 Python 脚本，实现中国天气网的数据爬取。
3. 撰写实验报告。