实验一 面向百度百科的深度优先爬虫

一、实验目的

通过设计、编写一个面向百度百科的深度优先爬虫程序，实现指定网站（百度百科）的特定词条下页面的爬取，进一步掌握常用的网络爬虫的基本架构、数据爬取策略和抽取方法的掌握，并加深对网络爬虫工作过程的理解。

二、实验内容

设计并实现对百度百科的词条（词条可任意指定）下页面的爬取，并将爬取结果页面的部分数据内容进行存储。

三、实验要求

  （1）采用深度优先策略实现爬虫程序；

  （2）对百度百科的词条（词条可任意指定）下三层节点页面进行爬取；

  （3）对爬取的页面抽取“词条名称、URL、描述、关键字信息”的内容保存到本地的txt类型文件中。

四、实验环境

  （1）pc机

  （2）Python程序开发环境

五、实验报告要求

  实验报告应包含以下内容：

  （1）实验目的

  （2）实验内容

  （3）实验要求

  （4）实验环境

  （5）实验步骤（含设计思想、算法结构说明、程序代码）

  （6）运行结果分析（含测试数据及运行截图）

  （7）实验总结