#### **指导**

MapReduce是利用多进程并行处理文件数据的典型场景。作为一种编程模型，其甚至被称为Google的”三驾马车“之一(尽管目前由于内存计算等的普及已经被逐渐淘汰)。在编程模型中，Map进行任务处理，Reduce进行结果归约。本周作业要求利用Python多进程实现MapReduce模型下的文档库（**搜狐新闻数据(SogouCS)，下载地址：https://www.sogou.com/labs/resource/cs.php**），注意仅使用页面内容，即新闻正文）词频统计功能。具体地：

1. Map进程读取文档并进行词频统计，返回该文本的词频统计结果。

2. Reduce进程收集所有Map进程提供的文档词频统计，更新总的文档库词频，并在所有map完成后保存总的词频到文件。

3. 主进程可提前读入所有的文档的路径列表，供多个Map进程竞争获取文档路径；或由主进程根据Map进程的数目进行分发；或者单独实现一个分发进程，与多个MAP进程通信。

4. 记录程序运行时间，比较不同Map进程数量对运行时间的影响，可以做出运行时间-进程数目的曲线并进行简要分析。

注：搜狐新闻数据的下载需要将下载按钮对应的链接  [http://www.sogou.com/labs/sogoudownload/SogouCS/news\_sohusite\_xmlfullzip](http://www.sogou.com/labs/sogoudownload/SogouCS/news_sohusite_xmlfullzip" \t "/Users/xihe/Documents\\x/_blank) 后面部分改为xml.full.zip