# 期末大作业 搜狗日志查询分析

|  |
| --- |
| 【大作业名称】搜狗日志查询分析 |
| 【大作业目的】 |
| 使用Spark大数据平台对搜狗日志数据进行查询分析。 |
| 【大作业原理】 |
| 使用HDFS存放日志数据；使用Spark RDD转换API对数据进行查询分析。 |
| 【大作业环境】 |
| 开发环境:   1. PCharm, Vi编辑器 2. 搭建PCharm开发环境   运行环境:  （1）OS: Ubuntu16.04  （2）已经部署好的Hadoop+Spark环境 |
| 【需求描述】 |
| 从搜狗实验室下载搜索数据，且有中文搜索关键字。  数据格式说明：（包含6个字段）  日志格式：访问时间 用户ID [查询词] 该URL在返回结果中的排名 用户点击的顺序号 用户点击的URL  (分隔符是个数不等的空格，有些行有6个字段，有些达不到6)  数据之间的分隔符是1个或多个空格。  某些行的数据可能达不到6个字段。  要求：  （1）统计文件总行数。  （2）统计符合要求的行数(判断每一行是否为空，列数是否为6)。  （3）统计搜索结果排名第2，点击顺序排在第1的数据。  （4）统计用户查询次数排行榜（降序）。  （5）统计top10敏感关键词，并使用词云图展示。 |
| 【实验步骤】 |
| （1）解压数据源，并上传到hdfs，保存的目录以个人学号区分，如001为学号。  tar -zxvf SogouQ.tar.gz  hdfs dfs -put SogouQ /001/  （2）编写Python文件，使用转换API处理日志数据。  （3）保存日志数据至HDFS，目录为：/001/output，其中：001为个人学号。 |