# 简介

该需求是从以股票名称为目标对象，以dmp平台获取的用户搜索为基础，统计含有股票名称的搜索词的搜索量分布情况。

并以用户搜索某个词时点击的文章为依据，统计这些文章中得分高的搜索词，及得分高的文章。

# 输入输出

* 输入：股票名称，例如老板电器、罗莱家纺、九阳股份、百润股份
* 输出：1. 含有以上某个股票名称的总搜索量

2. 含有以上某个股票名称的top100搜索词及搜索量

3. 按未分类词与某个股票名称相关的文章中top100个得分最高的关键词

4. 按已分类词与某个股票名称相关的文章中top100个得分最高的关键词

5. 与某个股票名称相关的得分最高的十篇文章

# 处理流程



从各地dmp平台搜集搜索的数据，直接使用数据加工程序处理后的结果，存放在hdfs的/user/clickwise/nstat/20150412/se\_\*/keywords文件夹下

(注:\*表示 baidu、360、google等)

hdfs里只存放最近一周的数据，历史数据存放在本地文件系统中。存放机器一般为dmp平台存储节点的第二台机器，例如

浙江 39 /data7/pp\_se/output/

海南 103  
山西 39  
辽宁 214  
江苏 106

本步骤输出结果的格式为：keyword\001pv\001ips



用各支股票的名称过滤出包含该股票名称的搜索词。不同股票名称过滤出的结果存储在不同的文件中。每一行的格式也是keyword pv ips。对文件按照pv数进行排序。

例如，九阳股份的输出结果是：

九阳股份 39 37

九阳股份有限公司 5 5

九阳股份股票 3 3

九阳股份股吧 3 3

002242九阳股份 3 3

……













# 程序