# 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 pvips。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

……



 统计股票名称相关的各种统计数据,即输出中的1、2，某个股票名称的总搜索量和

含有以上某个股票名称的top100搜索词及搜索量。

1的输出格式为：股票名称\001pv\001ips

2的输出格式为：搜索词\001pv\001ips

 抓取关键词的百度搜索显示的链接。使用类cn.clickwise.web.bd.WordLinks完成此功能。

使用方法是：

java –cp user\_click.jar cn.clickwise.web.bd. WordLinks <word> <totpage> >> result.txt

word :待分析的关键词

totpage:抓取前totpage页的搜索结果。每页抓取10个链接。

 dmp平台也有一些搜索词的点击链接数据，和上一步的结果合并。共同形成关键词的点击链接数据。下一步是将这些链接对应的网页内容抓取下来。该功能是通过类

cn.clickwise.web, FieldCrawl完成。该类的使用方法是：

Usage:<field\_num> <crawl\_index> <separator> <is\_debug>

field\_num : 输入的字段个数

crawl\_index: 抓取的字段索引，从0开始

separator:字段间的分隔符，001 表示 字符001，blank 表示\s+ 即连续空格 ,tab 表 示

is\_debug:是否进行调试

脚本执行命令举例：

cat word\_links\_result | java –cp user\_click.jar cn.clickwise.web.FieldCrawl 2 1 001 false >> result2.txt

抓取的网页内容放在最后一个字段，并用separator与原有字段分隔开来。

 对上一步抓取下来的网页内容进行分词。该功能通过类

cn.clickwise.clickad.seg. Segmenter完成。使用方法是：

Usage:[dict] <field\_num> <seg\_field\_index> <separator>

dict : 外加词典路径

field\_num : 输入的字段个数

seg\_field\_index: 要分词的字段编号，从0开始，即0表示第一个字段

separator:字段间的分隔符，001 表示 字符001，blank 表示\s+ 即连续空格,tab表示

脚本执行命令举例：

cat result2.txt | java –cp user\_click.jar cn.clickwise.clickad.seg.Segmenter so\_dict.txt 3 2 001 >> result2\_seg.txt

如果加词典，第一个参数为词典的路径。如果不加，忽略掉此参数即可。

统计某个股票名称点击文章中关键词的得分，和各个文章的得分。即输出中3、4、5

该功能通过类cn.clickwise.liqi.str.app. TFIDFStatisticN完成。该类的使用方法是：

Usage: <field\_num> <url\_index> <text\_index> <input> <output> <docscore>

field\_num:字段的个数,默认用\_分隔

url\_index:url所在的字段索引，从0开始

text\_index:网页内容(分词后)所在的字段索引，从0开始

input:输入文件，即每一行包含链接、网页内容(分词后)等字段

output:关键词得分的输出文件，每一行是一个关键词，跟着该词的得分，词和得分之间用\_分开

docscore:每篇文章(即输入的每行)的得分，输出到路径为docscore的文件中

脚本执行命令举例：

java –cp user\_click.jar cn.clickwise.liqi.str.app. TFIDFStatisticN 3 1 2 result2\_seg.txt word\_score.txt doc\_score.txt