# 近义词过滤实验结果

### 算法介绍

###### 算法思路

算法思路是首先选择一些能表明该文章是知识性的单词（种子词语），如《【投资课堂】技术分析讲堂：主要方法类别》这个标题中“课堂”、“讲堂”、“类别”、“方法”都表明这篇文章很可能具有知识性。

然后根据这些种子单词，寻找这些单词的近义词，更新种子单词集合，再通过更新后的种子单词集合寻找更多的表明文章知识性的单词，经过多轮迭代和人工筛选，获得最终的知识性单词集合。

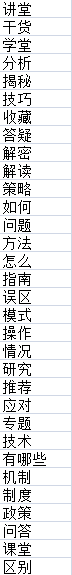
最终，统计每个标题中出现知识性单词的次数，并且排序，取前k个标题所对应的文章作为有知识性的文章。

###### 近义词搜索算法

1. 对每一篇文章的标题进行分词，构造每个单词的邻居单词矩阵（单词A的邻居单词指的是与单词A出现在同一标题中的那些单词）。
2. 根据单词的每一个邻居单词出现次数，构造出每个单词的邻居向量。
3. 对于一个查询单词，根据邻居向量，计算其他单词与查询单词之间的相似度，这里相似度使用余弦相似度。
4. 输出排序最高的k个单词。

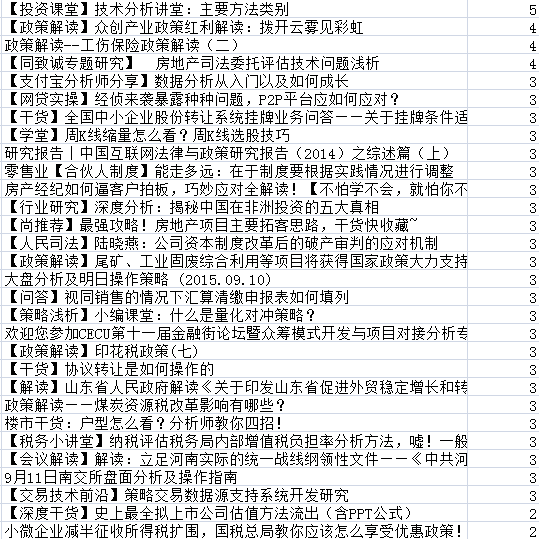
### 结果

在财经类别中，取6000篇文章进行知识性词语搜索，首先选择的知识性词语种子集合为——“讲堂”、“干货”、“学堂”、“分析”、“揭秘”。然后使用近义词搜索算法，更新后的种子集合增加了“技巧”、“收藏”、“答疑”、“揭秘”、“解读”。经过多轮筛选，最终选择了32个知识性的词语。



图一、32个知识性词语

根据这些知识性词语，统计每个标题出现这些词语的次数，降序排列，过滤出前50篇文章标题如下图。



图二、前50篇文章的标题