# 过滤公众号知识性文章

### 步骤

##### 1. 过滤

1.1 获得文章列表，共N1篇文章。根据收藏量，过滤出收藏量最高的N2篇文章。

1.2 对剩下的N2篇文章，根据文章内容长度，过滤出长度多于200字的M篇文章，作为初始语料库。

##### 2. 文章关键字

2.1 利用TF-IDF算法，求出每篇文章每个单词的TF-IDF值，将最高的100个词作为文章的关键字。

##### 3. 特征提取

3.1 文章内容序列特征（token level）：包括内容单词个数、内容不同的单词个数、内容句子个数、内容平均每句单词个数、标点个数。

3.2 文章表现特征（performance level）：包括文章转发数、点赞数、点击量、收藏量、平均阅读时间、平均阅读完成率。

3.3 文章内容单词特征（word level）：包括知识性单词个数、第一人称单词个数、第二人称单词个数、第三人称单词个数。

3.4 文章词性标注特征（pos level）：包括名词、动词、副词、形容词、介词、数量词在整篇文章中出现的平均数和方差（使用之前提到的histogram特征统计方法）。

以上特征共27个。

##### 4. 分类

4.1 共M篇文章，27个特征。使用支持向量机做分类器，进行二分类（知识性和非知识性）。得到每篇文章被分到知识性类别的概率，按照这个概率进行排序，取前K篇文章作为该语料库中知识性的文章。

##### 5. 标题简化

5.1 由于标题中存在一些冗余部分（如“一定要收藏”、“必须转载”），所以需要去除这些对文章内容理解没有意义的表示。按照标题中的标点符号进行分句。如“干货| 说说什么是对冲资金（一定要转载）”这个标题，被分句为[“干货”, “说说什么是对冲资金”，“一定要转载”]，共三个部分，可以看出来第二部分是真正有意义的表示。

5.2 每个子句分句之后，计算其与文章前L个关键字的Jarccard相似度作为第一个指标，每个子句的长度乘以0.5作为第二个指标，将两个指标相乘作为每个子句的得分，按照得分排序，得分最高的子句作为简化后的文章标题。

综上所述，得到了K篇标题简化后的知识性的文章。

### 结果

取1000篇文章作为语料库，标注400篇文章，未标注文章中top100作为知识性文章。由于计算单词TF-IDF值的程序运行较慢（之后如果要大量运行可以把计算TF-IDF值得程序写成mapred版本），所以现在处理的文章数量较少，但是结果比较理想的。

##### 1. 知识性词语

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 讲堂 | 干货 | 学堂 | 分析 | 揭秘 | 技巧 | 收藏 | 答疑 | 解密 | 解读 |
| 策略 | 如何 | 问题 | 方法 | 怎么 | 指南 | 误区 | 模式 | 研究 | 专题 |
| 技术 | 有哪些 | 机制 | 问答 | 课堂 | 区别 | 知识 | 原创 | 哪些 | 手法 |
| 什么 | 办法 | 措施 | 解答 | 理论 | 学习 | 做法 | 经验 | 系列 | 大全 |
| 常识 | 方式 | 各种 | 怎样 | 指导 | 方案 | 为什么 | 设计 | 图解 | 介绍 |
| 最全 | 实用 | 必备 | 深度 | 汇总 | 搭配 | 制度 | 秘诀 | 教程 | 攻略 |

##### 2. 知识性文章标题

|  |
| --- |
| 干货 | 新三板尽职调查指引详细解读（建议收藏） |
| 证券从业资格考试测试制度改革常见问题解答 |
| 【考试动态】证券业从业人员资格考试测试制度改革常见问题解答 |
| 史上最全：新三板审计问题总汇 |
| 【逐条解读.干货】互联网金融基本大法20条逐条解读 |
| 商业银行不良资产处置模式汇总 |
| 沪指尾盘为何现小幅跳水 11大权威机构本周策略大全(股) |
| 【收藏干货】新三板上市指引及财务要点 |
| 【财税学习】 IPO公司同一控制下业务重组相关问题小结 |
| 【互联网金融基本法来了】央行等十部委发布《关于促进互联网金融健康发展的指导意见》 |
| 与建设工程档案管理有关几个法律实务问题分析总结(2015)|法客帝国 |
| 尽职调查白皮书（适用于：IPO、新三板、PE、风险投资） |
| 推荐收藏：中国宏观经济数据分析入门 |
| 桥水内部宏观研究框架，让你彻底搞懂经济机器，力荐收藏！ |
| 干货回顾！余浩：另类资管产品常见交易结构及风险简析 |
| 【重磅首发】国家能源局关于推进新能源微电网示范项目建设的指导意见，国能新能[2015]265号 |
| [投资干货]下一个风口！284 家央企上市公司大全（附最全图表） |
| 创业者最应关心的十七个法律问题 |
| 一般纳税人增值税账务处理全攻略 |
| 【重磅干货】炒股赚钱三个方法 |
| 新婚姻法中关于房产问题的解读 |
| “营改增”对流转税管理的冲击 金融企业怎么办？ |
| 上海高院：盗窃罪若干新疑难问题解析（二） |
| 银行都怎么了？为什么这样说保险... |
| 【股权众筹】股权众筹融资操作方法 |
| 【谨记在心】大盘反弹 按照这十条秘诀炒股的都赚钱了 |
| 学堂｜短线博差价的3个基本技巧 |
| 【干货】7家顶级机构投资策略分析（红杉、IDG、DCM、经纬、纪源、晨兴、创新工场） |
| [交易技巧]主动解套的常用方法！ |
| 银行核心业务系统采用互联网分布式架构替代主机集中式架构的研究 |

##### 3. 非知识性文章标题

|  |
| --- |
| 踩踏暴跌落幕 这些超跌绩优股再不把握你就没机会了！ |
| 二季度基金重仓股全部出炉，你的股票上榜了吗（名单） |
| 【前瞻】7月20日淘金早参：操盘策略+热点追踪（附股） |
| 创业板成长个股逐一点评 |
| 中国银行正式发布回收硬币价格表，一不小心变土豪！ |
| 公安部出手“空头”颤抖 三类股望率先反攻 |
| 【重点关注】机构抢筹的38只股，是否有您持有的？ |
| 国家队护盘机构跟进，这些个股被密集买入！ |
| 7月14日晚间两市利好公告一览 |
| 独家深度揭秘：中科汇通举牌，翻倍个股集中营（续三） |
| 国家队持仓大揭秘：狂买30股 列入9股十大股东(附表) |
| 周一（13日）复牌名单一览 快看有您的股票吗 |
| 【财经早餐】2015.7.20星期一 |
| 【P2P信心大调查】2015年倒闭潮过后，谁能挺进P2P前5强？ |
| 【企业】12家公司豪推“10转20” 中报高送转行情引爆 |
| 公司半年报出炉：27家业绩增幅超10倍（四大板块突出） |
| 谁在掌控中国金融系统 现任政治局常委们与金融圈关系大起底 |
| 下周操盘策略（附股） |
| 【重磅】军工股连涨三天 “真假”军工股要分清 |
| 供应链从业人员必须理解的专业术语 |
| 内保外贷及信用证套利操作步骤和案例 |
| 锁定第二批央企改革概念股（附股） |
| 【推荐】现任政治局常委们与金融圈关系分析 |
| 谁的国开行？财政部终究败给了央行 |
| 基金二季报抢先看84股获增持3行业受青睐，看有投资机会？ |
| TMT板块18个股迎高送转，业绩+题材双轮驱动供你选择（附股） |
| 国家队钦点操盘手策略曝光！3路径把握投资机会 |
| 税务总局：建筑业跨省异地工程人员个人所得税征收管理问题的公告 |
| 价值投资者必读的23本书 |
| 股市面临4大挑战（7月17日很重要吗？） |

##### 4. 简化的知识性文章标题

|  |
| --- |
| 新三板尽职调查指引详细解读 |
| 证券从业资格考试测试制度改革常见问题解答 |
| 证券业从业人员资格考试测试制度改革常见问题解答 |
| 新三板审计问题总汇 |
| 互联网金融基本大法20条逐条解读 |
| 商业银行不良资产处置模式汇总 |
| 11大权威机构本周策略大全 |
| 新三板上市指引及财务要点 |
| IPO公司同一控制下业务重组相关问题小结 |
| 央行等十部委发布《关于促进互联网金融健康发展的指导意见》 |
| 与建设工程档案管理有关几个法律实务问题分析总结 |
| IPO、新三板、PE、风险投资 |
| 中国宏观经济数据分析入门 |
| 桥水内部宏观研究框架 |
| 另类资管产品常见交易结构及风险简析 |
| 国家能源局关于推进新能源微电网示范项目建设的指导意见 |
| 家央企上市公司大全 |
| 创业者最应关心的十七个法律问题 |
| 一般纳税人增值税账务处理全攻略 |
| 炒股赚钱三个方法 |
| 新婚姻法中关于房产问题的解读 |
| “营改增”对流转税管理的冲击 |
| 盗窃罪若干新疑难问题解析 |
| 为什么这样说保险 |
| 股权众筹融资操作方法 |
| 按照这十条秘诀炒股的都赚钱了 |
| 短线博差价的3个基本技巧 |
| 红杉、IDG、DCM、经纬、纪源、晨兴、创新工场 |
| 主动解套的常用方法 |
| 银行核心业务系统采用互联网分布式架构替代主机集中式架构的研究 |