广东食品药品职业学院

毕业论文

|  |
| --- |
| **浅析Java程序设计在证券分析实务中的应用** |
| **的应用** |

|  |  |
| --- | --- |
| 毕业生姓名 | **李源** |
| 毕业生学号 | **117612101023** |
| 毕业生班级 | **17移动应用开发1班** |
| 专业名称 | **移动应用开发** |
| 所属学院 | **软件学院** |
| 指导教师 | **谭新** |
| 论文提交日期 |  |

**浅析Java程序设计在证券分析实务中的应用**

李源

【**摘 要**】本文在总结证券研究活动中的程序化工作，金融实务分析员无法快速进行工作时，利用Java语言编写程序，以达到快速自动化任务。同时将举例一些在证券分析实务中普遍用到的程序化应用，对其进行探讨与实现。

【**关键词**】Java；证券分析；金融实务；数据；程序

本文在总结作者从事证券分析实务工作以来，遇到过的程序化工作问题，开发程序并运用到实践中，提高证券分析工作的效率与专业性。本文将按流程和难度不同，分别举例两个Java应用，并对其工作原理、代码实现与实际应用进行剖析。

举例的第一个较为简单的程序为“上市公司及其子公司公司注册地分析”。该工具通过调用“Wind全球企业数据库“API接口，进行遍历查找、整合、存储。且日后根据证券分析师的需要，提供改写程序，就可以不断为证券分析师提供精简准确的数据分析源。

举例的第二个较为复杂的程序为“证券研报相似度分析”。该工具将通过算法，对同一证券的单篇深度研报与市面上其他券商深度研报的相似度、数据来源等方面计算，得出其相似指数，为证券分析师的深度研报研究价值提供借鉴。

**1上市公司及其子公司公司注册地分析**

**1.1概述与实现思路**

随着编程技术的不断发展，证券分析方法也愈来愈复杂多样。证券分析师撰写企业研究报告时，通常都会需要大量的数据，而有一些数据的提取并不方便，例如筛选所有归母净利润增长率为100%以上的上市公司、筛选子公司/分公司数量超过10个的上市公司等等。此类数据，往往需要对所有上市公司乃至所有分支公司的信息进行比对分析。如果进行人工的机械化查找比对，费时费力的同时，又不能保证数据的完整性和准确性，故人工分析不可取。因此，程序分析可以大行其道。

在一次证券研究的过程中，分析师需要获取所有公司注册地在无锡的上市公司及其上市公司所有子公司，以用于分析地域/行业对企业发展的影响。其中，若子公司还有子公司则全部获取，也就是数据结构中所涉及的树与森林结构遍历算法，程序获取到所有数据后需要再进行数据整合，以便为证券分析师研究提供参考数据源。要实现本程序，有以下要点：如何获取所有上市公司？如何获取某个公司的子公司？如何获取公司的公司注册地？如何按母公司-子公司-孙公司…的结构保存所有上市公司？

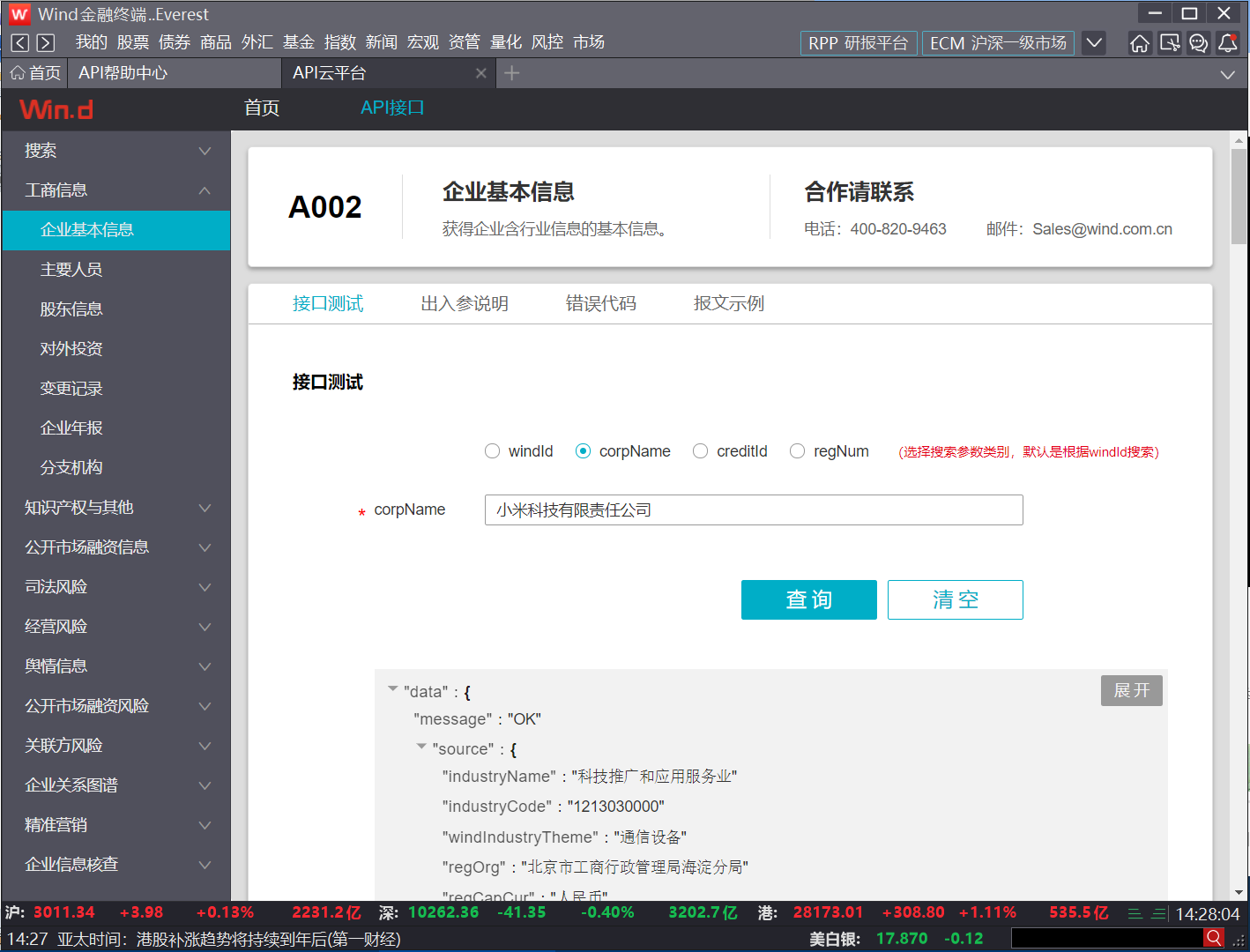
要实现此程序，首先，要通过检索，下载到所有上市公司的证券代码对应的公司全称，再通过该全称列表，即可遍历全部上市企业。然后，获取子公司的方法通过检索比对可知，获取企业信息的一个既高效又简单的方法是，使用“Wind金融终端”软件下的“量化”-“API帮助中心”-“企业库API”。通过授权购买后，获取本API，以下（图1-1）为本软件的API接口帮助信息界面。

图1-1：Wind金融终端 - API云平台

从帮助信息中可知，本API数据库返回的数据为JSON数据；该API有工商信息、企业关系图谱、企业分支机构等API子项；通过本API的出入参说明以及报文示例可以初步形成代码的业务逻辑。了解完API后，即可编写链接服务器的代码，API授权用户登录代码、获取token代码等（图1-2）。然后通过给定的参数以及本地的上市企业列表调用该企业API，编写并测试好API查询调用逻辑代码（图1-3），最后运行程序，导出数据，即可完成对目标数据的获取与筛选。

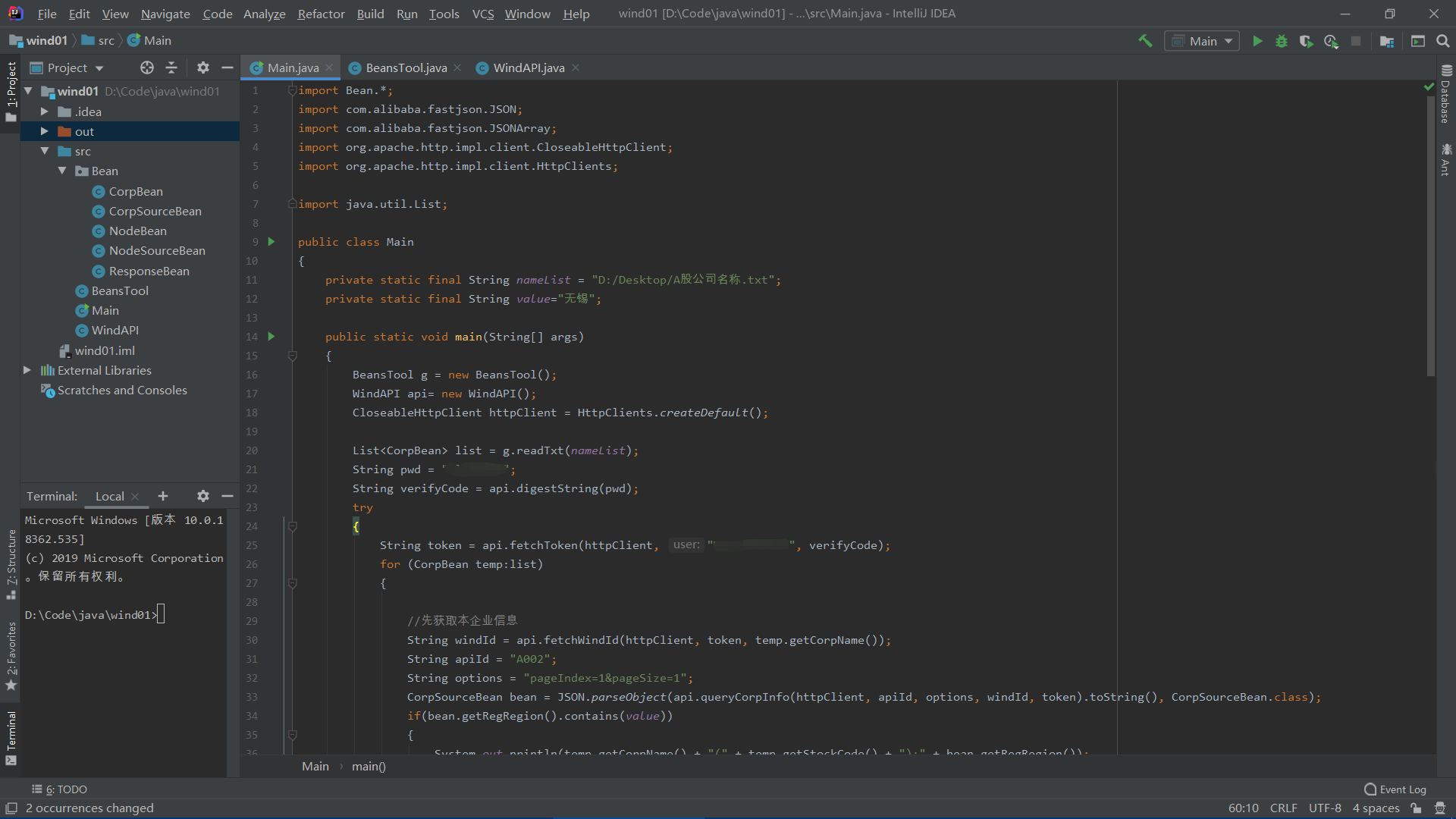


图1-2：部分代码实现

**1.2代码业务逻辑与优点**

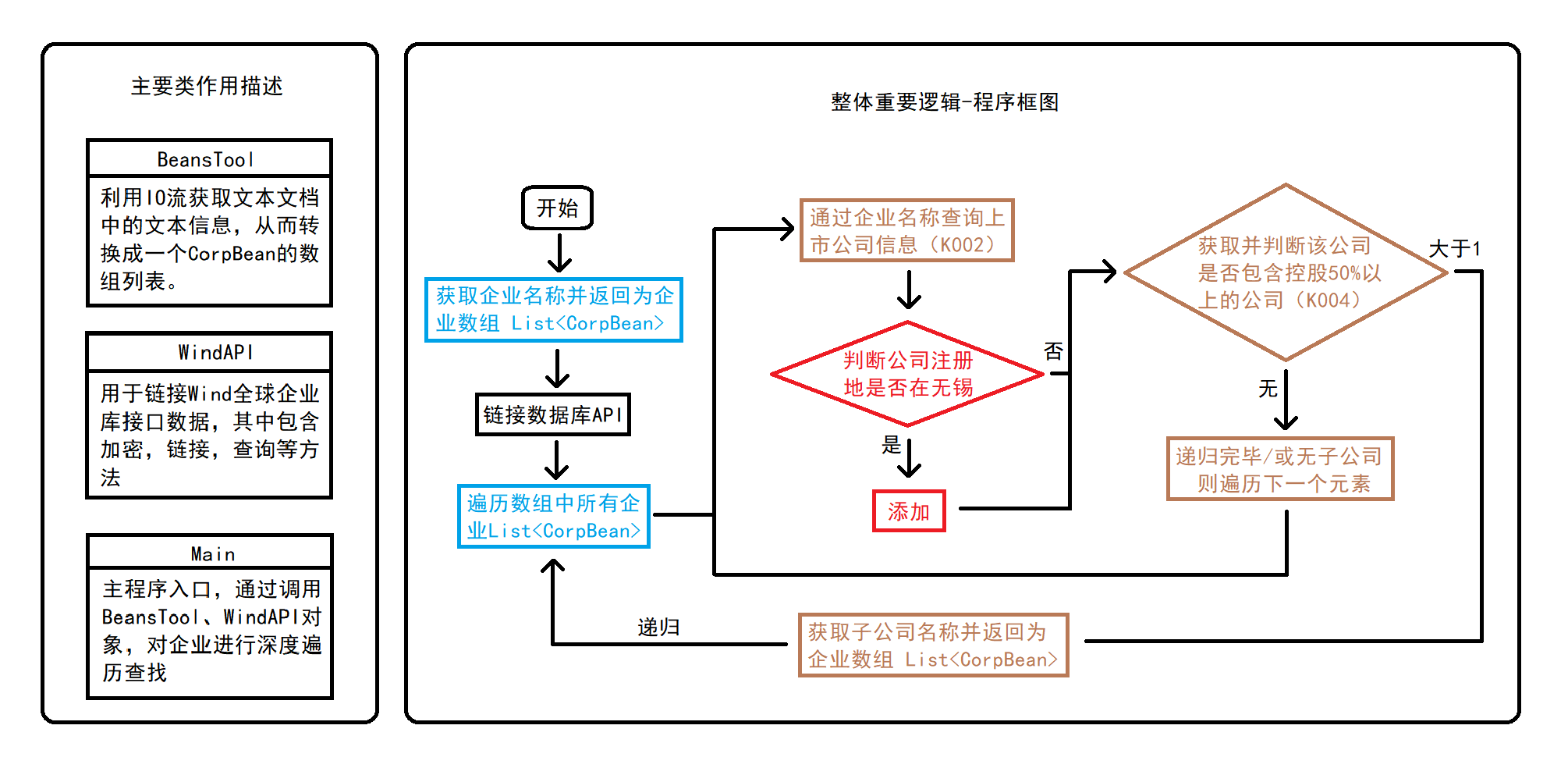


图1-3：主要类作用描述及业务逻辑

此程序主要有以下优点：

1. 程序可维护性高。后期可根据判断内容的不同，更改条件，以获取相应的符合要求的企业及所有对应子公司。例如获取所有涉及人工智能行业的公司、获取所有拥有N个子公司以上的公司等等；也可以更改数据源，也就是更改所导入的公司列表，实现小范围查询、指定行业查询、指定公司池查询等功能。
2. 程序较为轻量化。因使用API模块，故后期可加入相应接口规范API类，以达到只修改API类，就可以达到更改数据获取源而不影响程序逻辑，大大降低程序耦合性。当今金融行业流行的金融数据API，东方财富Choice、彭博社Bloomberg、汤森路透Eikon等，打包成API对象导入程序，即可实现原有功能。
3. 程序代码复用性高。遍历所有公司与遍历子公司的算法逻辑高度重合，故此程序将其提炼，实现了仅使用一个遍历算法就组成整个程序。

**2证券研报相似度分析**

**2.1概述与实现思路**

证券研究报告（以下称研报，图2-1）是证券分析师对证券及证券相关产品的价值或影响其市场价格的因素等进行分析后，总结并撰写含有对具体证券及证券相关产品的价值分析、投资评级意见等内容的报告文件。

具有发布证券研究报告业务证书的证券从业人员担任分析师并发布研报时，需要对其所发布的研报负责。现如今，研报抄袭现象屡见不鲜，轻者构成民事侵权，损害公司名誉，重者或构成[侵犯著作权罪](https://www.baidu.com/s?wd=%E4%BE%B5%E7%8A%AF%E8%91%97%E4%BD%9C%E6%9D%83%E7%BD%AA&tn=SE_PcZhidaonwhc_ngpagmjz&rsv_dl=gh_pc_zhidao)，从业人员甚至公司会被追究法律责任。

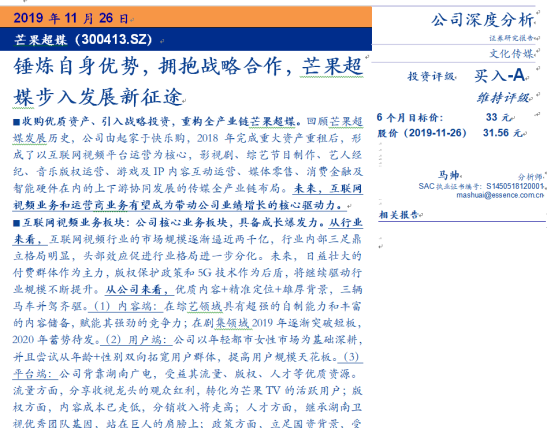
在撰写研报时，研报内容难免出现与其他研报内容数据来源相似、研究观点相似、统计内容相似、部分表达相似等可能引发法律纠纷的内容（图2-2）。这就需要从业人员在发布研报时，对研报内容进行查重，而目前尚未有针对研报的查重系统，多数为论文查重系统，查重样本并非研报，故需开发一款证券研报相似度分析的程序，从而为从业人员进行研报查重提供便利。

图2-1：证券研究报告 图2-2：研报抄袭新闻报道

要实现本程序需要解决以下问题：

1，如何将其他研报作为比对样本？研报本质是一篇PDF格式的文章，由于受PDF特性及字体编码不同等因素影响，导致无法使用Java程序直接解析字符，现有的诸如pdfbox、Spire.PDF、iTika等PDF解析框架至今仍存在着很多缺陷，例如对中文支持不友好、存在乱码、成本较高、系统庞大等。

2，如何快速对研报内容进行比对分析？研报相似度分析针对的是文字分析，也就是字符比对，但是一篇深度报告往往动辄上万字，整段文章比对或单文字比对显得不切实际。

3，如何将相似度量化？判断一段文字是否相似，这似乎很容易，但有时会遇到常用短语或术语，专有名词等，这种相似是可以忽略不计的。分析结果展示时，如何清晰明确地展现研报之间的相关性也是一个重要的问题。

4，如何准确定位相似内容？通过程序进行客观分析可以大大加快速度，但有时还需要对相似内容进行人工校验，故需要保存所获取相似的文段的对应位置。

针对以上问题，反复探寻后得出以下解决思路：

1. 使用PDF转换软件，将被分析文档与比对样本集所有PDF转成TXT文本。
2. 对被分析文档提取关键词后，逐个关键词与比对样本各文段进行比对。
3. 可对各关键词进行加权，权重越高影响越大，权重与出现频率成反比。
4. 对每个关键词进行分割，每个关键词即对应文中某处，可通过检索定位。

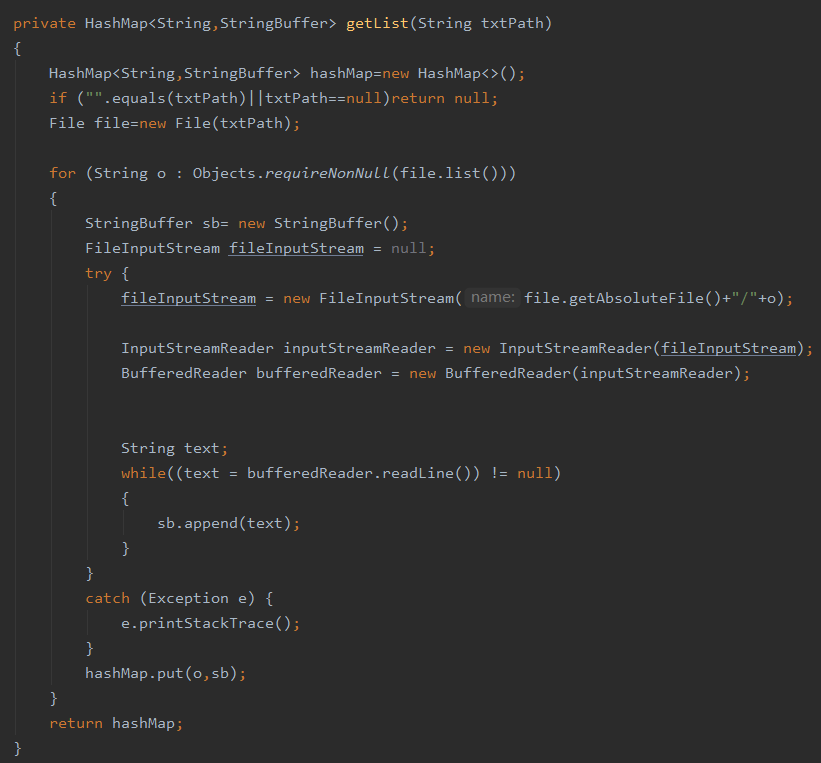
**2.2程序业务逻辑**

本程序主要由文件操作（File Operation）类、相似度分析（findEquals）、比对结果并输出（printRes）等方法构成，下面将逐一解析其代码逻辑。

**2.2.1文件操作（File Operation）**

此类的主要由读取包含所有对比样本文件夹里的所有文件、读取被分析文档并转化成关键词数组等方法组成，读取文件列表较为基础，主要为IO流操作，下面主要针对第二个方法进行解析。

第二个功能，也就是读取被分析文档并转化成关键词数组的方法，实现逻辑是将所有文段进行分割，分割条件为遇到例如句号逗号等标点符号、制表符回车符等特殊标识符、“的，则，和，在”等字眼、谓语动词、人称代词等时，替换成特定分隔符，并在之后进行空白行与空白格删除，去除重复项等操作，达到（图2-4）的效果，并逐个关键词保存，组成关键词数组列表。



锤炼自身优势|拥抱战略合作|步入发展|新征途收购优质资产|引入战略投资|重构全产业链|发展历史|公司由起家于|快乐购|完成重大资产重组|互联网视频平台|运营为核心|综艺节目制作|音乐版权运营|IP内容互动运营|消费金融|智能硬件|上下游协同发展|传媒全产业链布局|互联网视频业务|运营商业务有望成|带动公司业绩增长|核心驱动力|互联网视频业务板块|公司核心业务板块|具备成长爆发力|从行业来看|互联网视频行业|市场规模逐渐|逼近两千亿|行业内部三足鼎立格局明显|头部效应促进行业格局进一步分化|三辆马车并驾齐驱|内容端|综艺领域具有超强|自制力

图2-3：部分文件操作类代码 图2-4：关键词转化完成后效果

**2.2.2相似度分析（findEquals）**

此方法嵌套了两重循环，实现了遍历每一个关键词时，每一个关键词都对每一个文件的文件内容进行查找。查找的实现方式是调用search方法后，整合查找结果，汇总后返回一个HashMap集合（图2-6）。search方法将对关键词，和对应的文字段落进行循环地逐字拆解文段操作，每次拆解完成都使用indexOf方法比对关键词并累计次数。

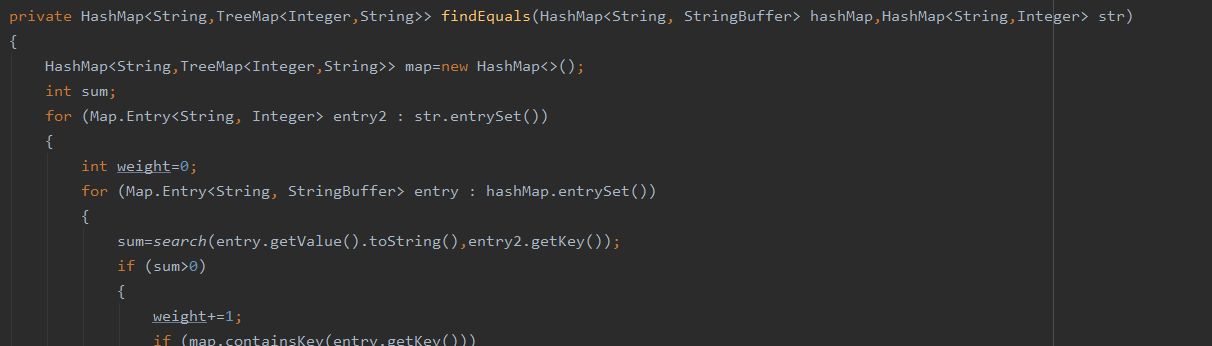
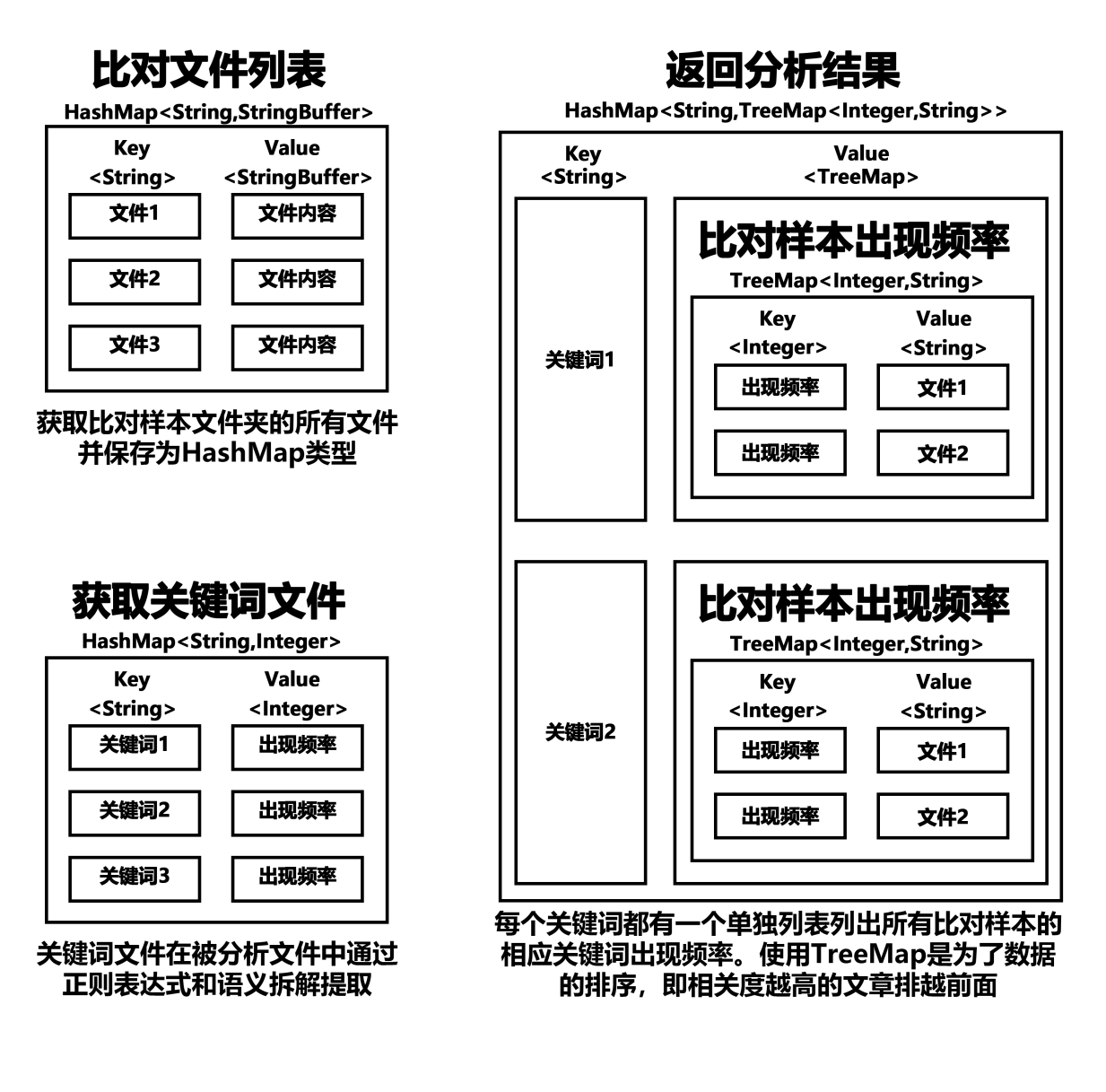
为了实现预期效果，使用了多重集合框架类型（图2-5）。从抽象上看，此方法逻辑为：传入“比对文件列表集合”及“获取关键词文件集合”，通过比对并汇总整合，最后生成“返回分析结果集合”，到此方法执行完毕。

图2-5：部分相似度分析方法代码

图2-6：数据结构图

**2.2.3** **比对结果输出（printRes）**

程序输出结果分为两部分，一部分为控制台文本输出（图2-7），另一部分为文本文件输出（图2-8）。控制台输出的文本为详细数据，可供于优化算法或调试，文本文件输出为最终分析结果。

这里我们举例一个在实际中运用的例子：把1份研报作为被分析文档，24份文档作为比对分析样本，导入文件后运行，得出（图2-7）运行结果，我们可以看到有5列运行结果，从左到右分别代表相似值、文件名、权重、频率、关键词。每一列，对应着每一行的5个参数，每一行又代表着每一份比对分析样本对应每个关键词的分析数据。

图2-7：控制台文本输出（部分）

下面为控制台及文本文件中，部分重要输出数据列释义：

权重：对应关键词在其他研报出现次数，权重为1时最大。

频率：对应关键词在该研报中出现的次数，为零则不显示。

相似值：根据权重、频率、关键词长度等数值计算生成。

相似度：根据每份研报中所有关键词的相似值相加得出。

最后，在进行了多轮不同类型的研报组分析，定制了文本文件输出结果（图2-8）的“区间参考定义”列，通过代入相关度的值后，供比对做参考

|  |  |  |
| --- | --- | --- |
| **研报编号** | **相关度** | **区间参考定义** |
| 20190826-国信证券 | 3969 | **2000以上** 有大量独立重复内容,疑抄袭 |
| 20190729-广证恒生 | 1339 | **500-2000** 可能引用了同一数据源 且该数据源较少被使用 |
| 20190830-华创证券 | 1091 |
| 20181026-国盛证券 | 1055 |
| 20190509-长城证券 | 1030 |
| 20180915-东吴证券 | 739 |
| 20191008-国金证券 | 684 |
| 20191021-招商证券 | 410 | **0-500** 为基本无重复  （大部分研报都存在的内容， 则该内容无权重,将不被计入） |
| 20180828-西南证券 | 394 |
| 20190724-国盛证券 | 334 |
| 20191106-国盛证券 | 311 |
| 20190711-天风证券 | 310 |
| 20180906-华泰证券 | 210 |
| 20190819-民生证券 | 171 |
| 20180831-国海证券 | 152 |
| 20190109-东北证券 | 148 |
| 20180902-招商证券 | 70 |
| 20181022-国金证券 | 59 |
| 20180808-广证恒生 | 50 |
| 20190307-长城证券 | 31 |
| 20190529-联讯证券 | 26 |
| 20180827-联讯证券 | 13 |
| 20190715-长城证券 | 5 |
| 20190524-招商证券 | 4 |

图2-8：文本文件输出

**2.3在实际运用中的效果**

在上节的例子中，运行程序并获取到分析结果的文本文件后，根据相似度于区间可得0-500区间的17份研报相似度较低，初步判定这些比对样本与被分析样本无过高相关性，而被分析文档与“20190826-国信证券”呈高度相关，需重点审阅分析。经过人工比对验证后，确定被分析文档与该文档或存在抄袭现象，需让相关从业人员进行改正。下表为部分相似区域截图（图2-9）。

|  |  |  |
| --- | --- | --- |
|  | 被分析文档 | 20190826-国信证券 |
| 相似处1 |  |  |
| 相似处2 |  |  |
| 相似处3 |  |  |
| 相似处4 |  |  |
| 相似处5 |  |  |
| 相似处6 |  |  |
| … | … | … |

图2-9：部分相似区域截图

**参考文献：（宋体小四号、加粗，首行不缩进）**

[1] 陈婕 . Python 技术在商业银行信贷风险数据方面的运用 .[I] . 软件和信息技术服务业 ,2019(12)

[2] CHEN Huimin XU Yunbao LI Yuanhui . Teaching Reform of Mathematical Finance and R Language Application . [G] . 文化科学 . 2019(04)

[3]Wind . Wind量化接口FAQ . 2014(6)

[4]陈红玲 周立 . C#与Matlab混合编程技术在金融行业中的应用 . 银行业务