证券研究报告
专题研究 | 金融工程

CMS 招商证券

招金词酷

金融文本挖掘的分词工具

叶　涛
021-68407749
yetao@cmschina.com.cn
S1090514040002

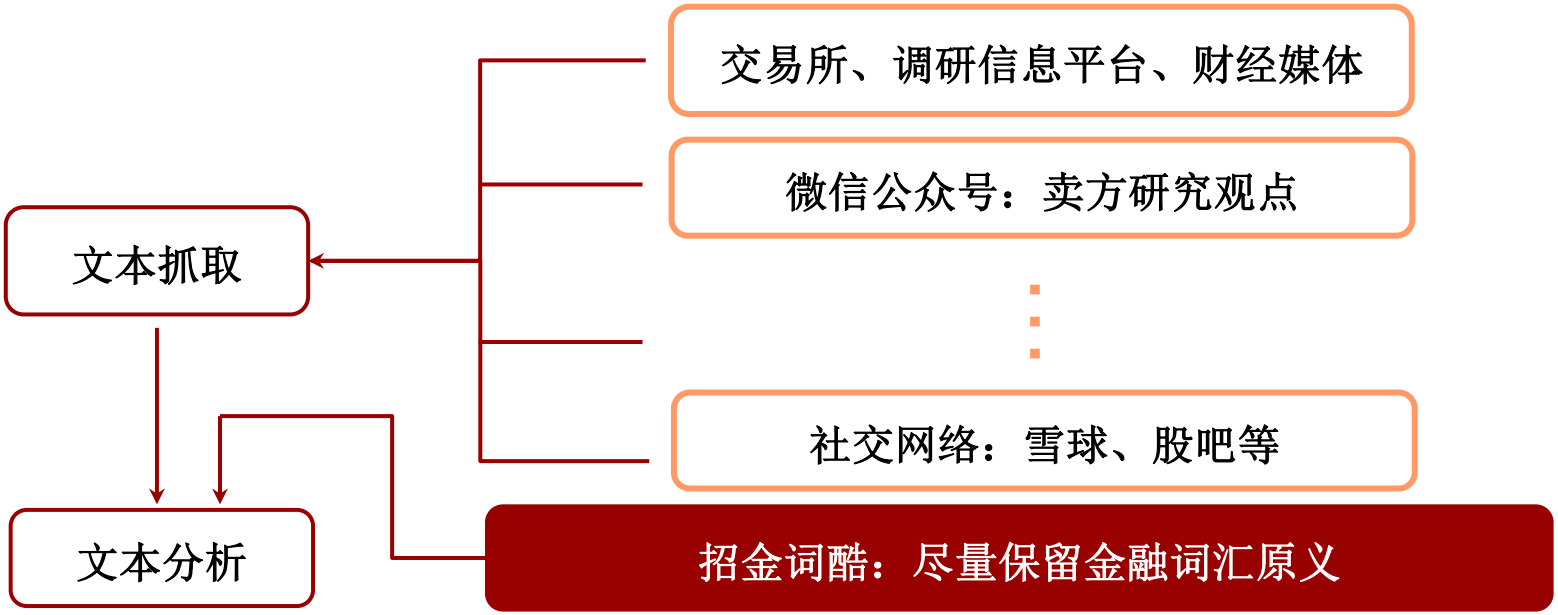
研究助理
赵月涓
021-33938893
zhaoyuejuan@cmschina.com.cn
S1090115060055

报告日期：2016 年 10 月

目录

- 一、搭建招金词酷
- 二、招金词酷赢在精度
- 三、手把手教您用招金词酷
- 四、PDF 批量转 txt、HTML 工具

搭建招金词酷

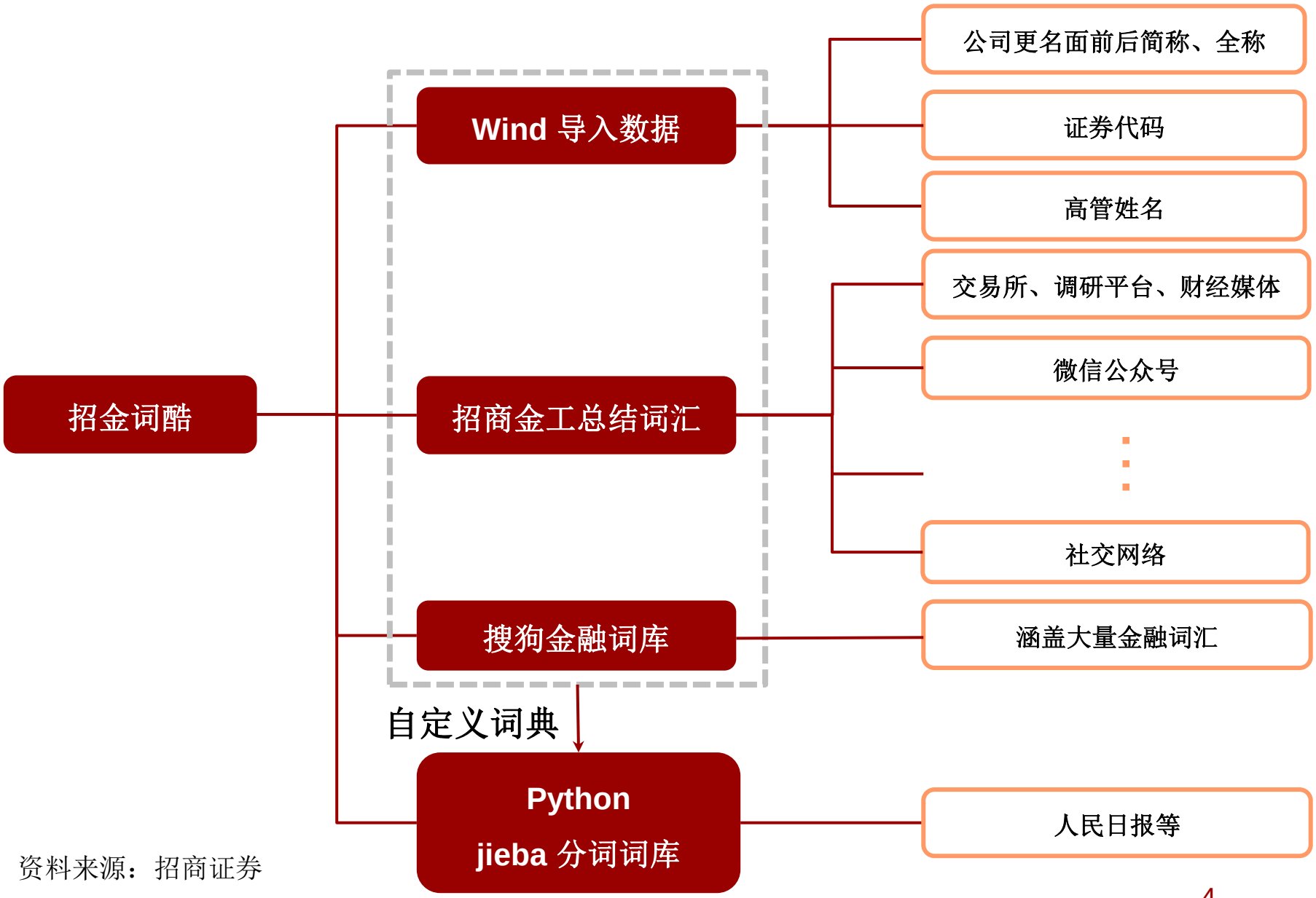


资料来源：招商证券

- 原文：“在《中国证券报》、《上海证券报》、《证券时报》刊登”
  - **传统方法**：在 / 中国 / 证券报 / 上海 / 证券报 / 证券 / 时报 / 刊登
  - **招金词酷**：在 / 中国证券报 / 上海证券报 / 证券时报 / 刊登

- 原文：“丽江玉龙旅游股份有限公司”
  - **传统方法**：丽江 / 玉龙 / 旅游 / 股份有限公司
  - **招金词酷**：丽江玉龙旅游股份有限公司

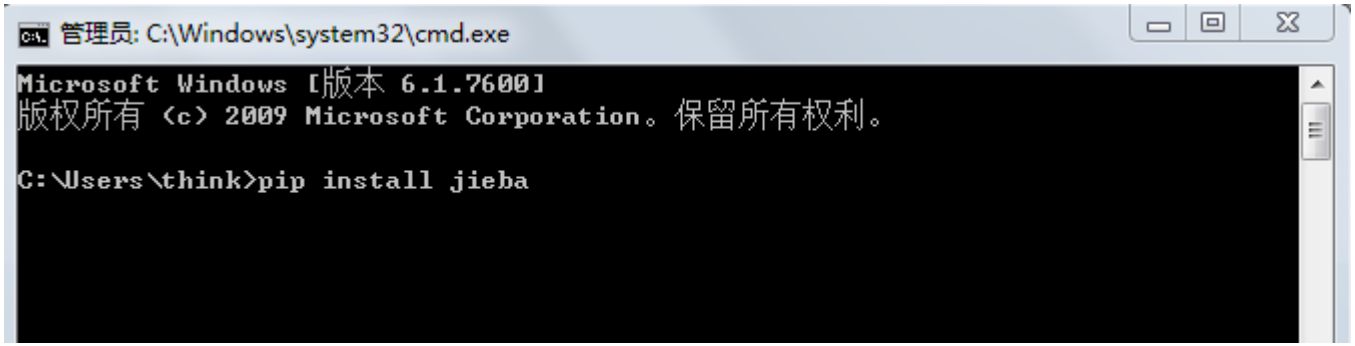
招金词酷组成部分



资料来源：招商证券

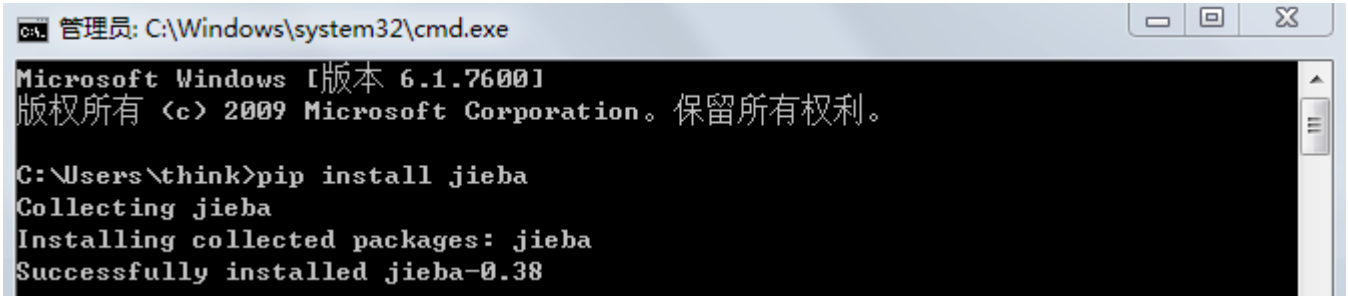
安装 jieba

- 安装 jieba 分词模块，在命令提示符（Windows）或者终端（macOS）输入 **pip install jieba** 进行自动安装，以 Windows 系统作为示例。



资料来源：招商证券

- 出现如下界面，表示 jieba 分词模块安装成功



资料来源：招商证券

基本命令

- jieba 的基本命令是 jieba.cut()，三个参数如下：
  - 需要分词的字符串
  - cut_all：控制是否采用全模式
  - HMM：控制是否使用 HMM 模型（隐马尔科夫模型）用于识别新词

- 代码示例一

```
seg_list = jieba.cut("我来到北京清华大学", cut_all=True) （全模式）
print("Full Mode: " + " / ".join(seg_list))
seg_list = jieba.cut("我来到北京清华大学", cut_all=False) （精确模式）
print("Default Mode: " + " / ".join(seg_list)) ）
```

  - 输出
    - 全模式： 我 / 来到 / 北京 / 清华 / 清华大学 / 华大 / 大学
    - 精确模式： 我 / 来到 / 北京 / 清华大学

- 代码示例二

```
seg_list = jieba.cut("他来到了网易杭研大厦")
（默认新词识别模式）
print(" / ".join(seg_list))
seg_list = jieba.cut("他来到了网易杭研大厦",HMM = False) （不使用新词识别模式）
print(" / ".join(seg_list))
```

  - 输出
    - 新词识别模式：他 / 来到 / 了 / 网易 / 杭研 / 大厦
    - 不使用新词模式：他 / 来到 / 了 / 网易 / 杭 / 研 / 大厦

词典加载方法

- jieba 中载入词典
  - 用户可以指定自己自定义的词典，以便包含 jieba 词库里没有的词。虽然 jieba 有新词识别能力，但是自行添加新词可以保证更高的正确率。
  - 用法

    jieba.load_userdict(file_name)
  - 词典格式
    - 与 jieba 自带的词典文件 dict.txt 格式相同，一个词占一行。每一行分三部分：词汇、词频（可省略）、词性（可省略），用空格隔开，顺序不可颠倒。

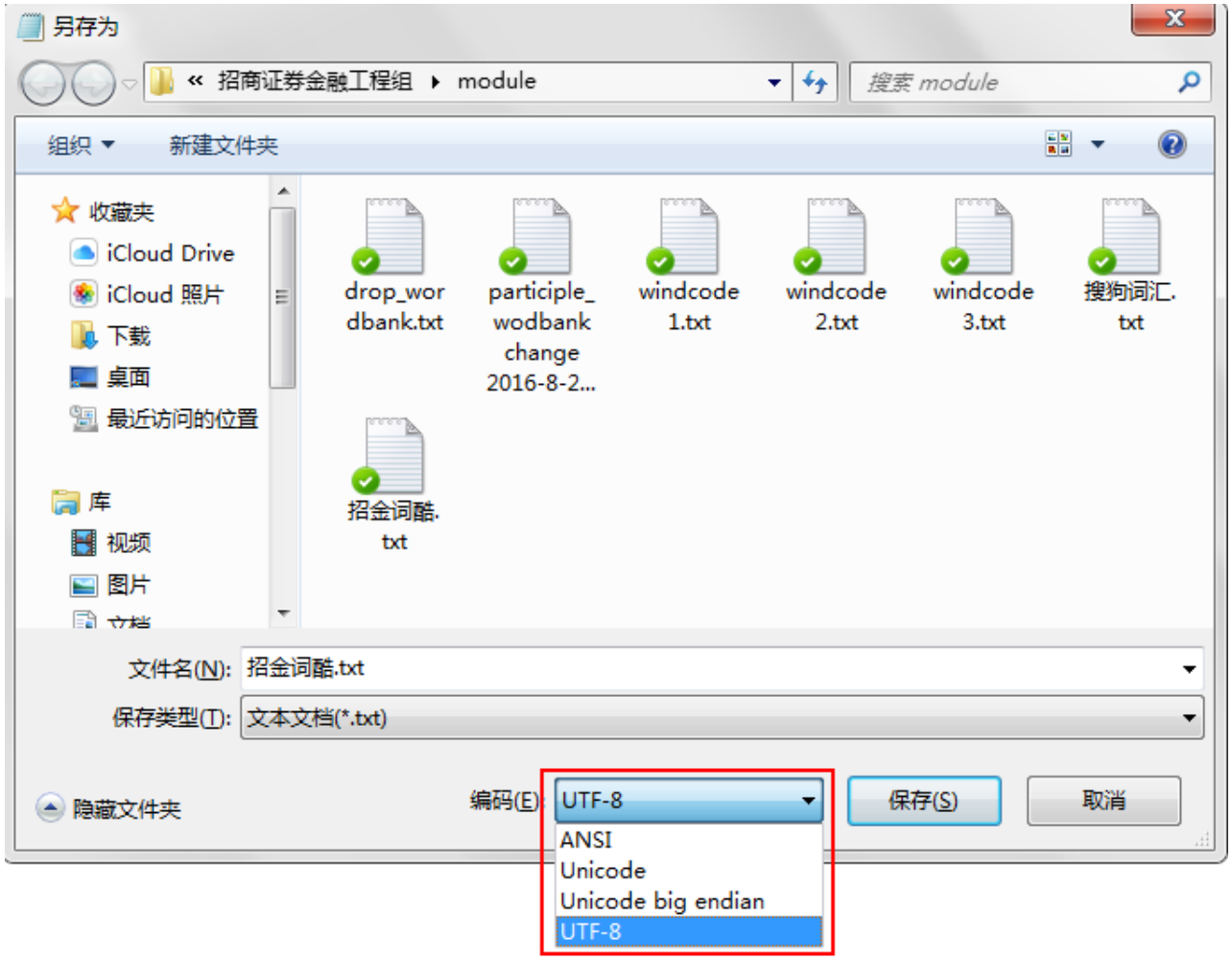
词汇	词频	词性
创新办	3	i
云计算	5	
凯特琳		nz
台中		

资料来源：jieba、招商证券

  - 路径设置注意事项
    - file_name 为自定义词典的路径，**建议使用英文**；若路径为中文或文件用二进制方式打开，则必须转为 **UTF-8** 编码。

词典保存方式

- jieba 中自行载入的自定义词典要以 **UTF-8** 格式保存
- 更多关于 jieba 分词模块的信息，可参见： https://github.com/fxsjy/jieba



资料来源：招商证券

jieba 词库示例

- jieba 分词工具通过对《人民日报》等海量通用类文章的分析已经收录了 **349046** 个常用词汇。

词汇	词频	词性
已达成	3	nrt
已过期	3	d
已近尾声	3	n
已远	3	d
已逝	3	v
已逸待劳	3	nz
已销	3	v
木牌	374	n
木牛流马	40	ns

资料来源：jieba、招商证券，截止 2016 年 9 月 20 日

搜狗金融词库

- 从搜狗官网下载金融词库 http://pinyin.sogou.com/dict/detail/index/334

资料来源：搜狗科技、招商证券

- 搜狗金融词库是特有的 **scel 格式**文件，需要**转换成 txt** 文件方能使用。

资料来源：搜狗科技、招商证券

搜狗金融词库

- “深蓝词库转换工具”可将 scel 格式文件转换成 txt 文件。

资料来源：招商证券

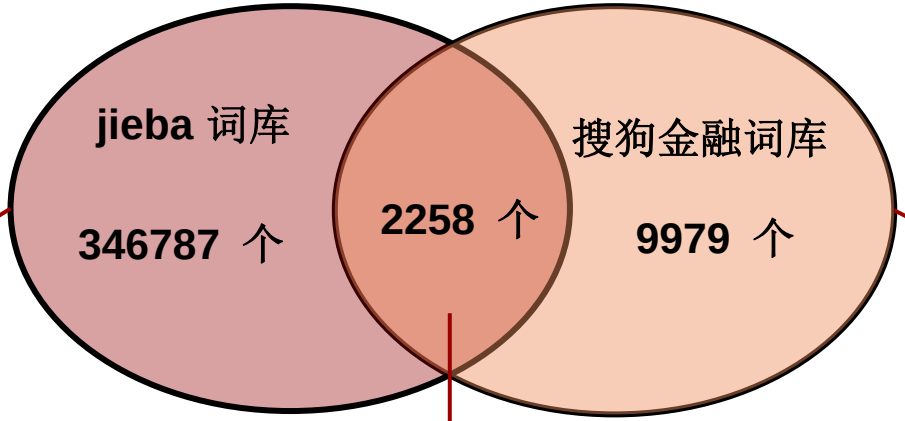
搜狗金融词库示例

- 通过整理，转为 jieba 可以使用的 txt 文件，并以 **UTF-8** 格式保存。经整合后，从搜狗金融词库中合计收录 **12337** 个词汇。

词汇	词频	词性
按计划	11	d
按季业务统计	11	v
按价值计算	11	v
按揭	11	ad
按揭财产	11	n
按揭贷款	11	n
按揭贷款安排	11	v
按揭贷款合同	11	n
按揭贷款契据	11	n

资料来源：搜狗科技、招商证券，截止 2016 年 9 月 20 日

jieba 词库与搜狗金融词库的关系



jieba 词库
已达成
已过期
已近尾声
已远
已逝
已逸待劳
已销
木牌
木牛流马

重合词汇
主营业务
买单
买卖双方
二季度
买超
买点
亏损
亏损面
买进

搜狗金融词库
按计划
按季业务统计
按价值计算
按揭
按揭财产
按揭贷款
按揭贷款安排
按揭贷款合同
按揭贷款契据

资料来源：jieba、搜狗科技、招商证券

Wind 词汇

- 通过 Wind 的 Python 插件，可以从 **Wind** 获取到包括**公司更名前后的简称**、**全称**、**证券编号**、**高管姓名**等经常会在金融类文本中出现的词汇，代码如下：

```
import WindPy
w.start
w.wsd("000001.SZ","sec_name,comp_name,boardchairmen,ceo,crtindpdirector,
frmindpdirector", "2016-09-13", "2016-09-13", "")
```

总经理

证券名称

公司全称

董事长

公司独立董事（现任）

公司独立董事（历任）

- **公司更名前后的简称、全称**可从 Wind 终端提取
  - 路径：“沪深市场概况”→“**股票更名**”（简称）、“**公司更名**”（全称）

Wind 词汇示例

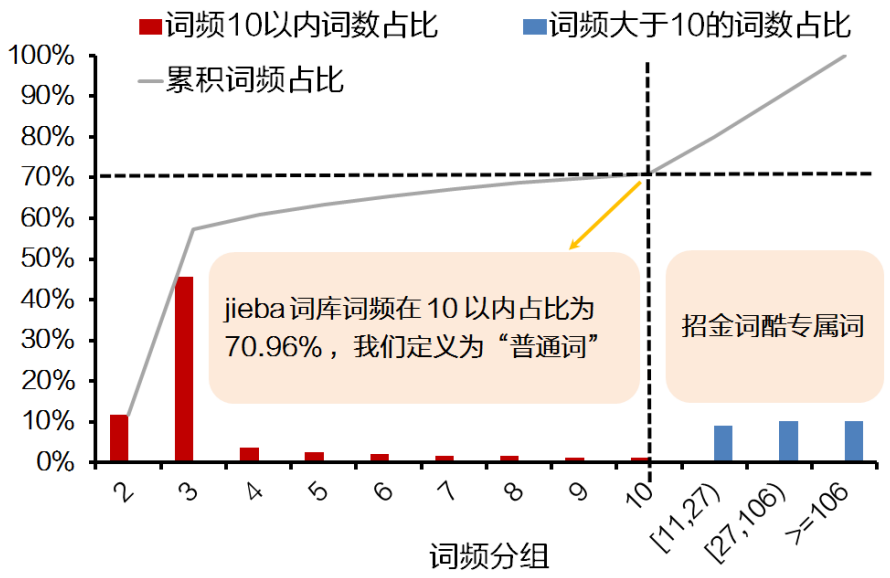
- 将 Wind 词汇整合、保存为 txt 文件，收录至招金词酷共 **10461** 个词汇。

词汇	词频	词性
天润乳业	20	nt
现代制药	20	nt
仰帆控股	20	nt
002455.SZ	20	nz
002456.SZ	20	nz
黄伟国	40	nr
金鑫	40	nr
300349	15	nz
300350	15	nz

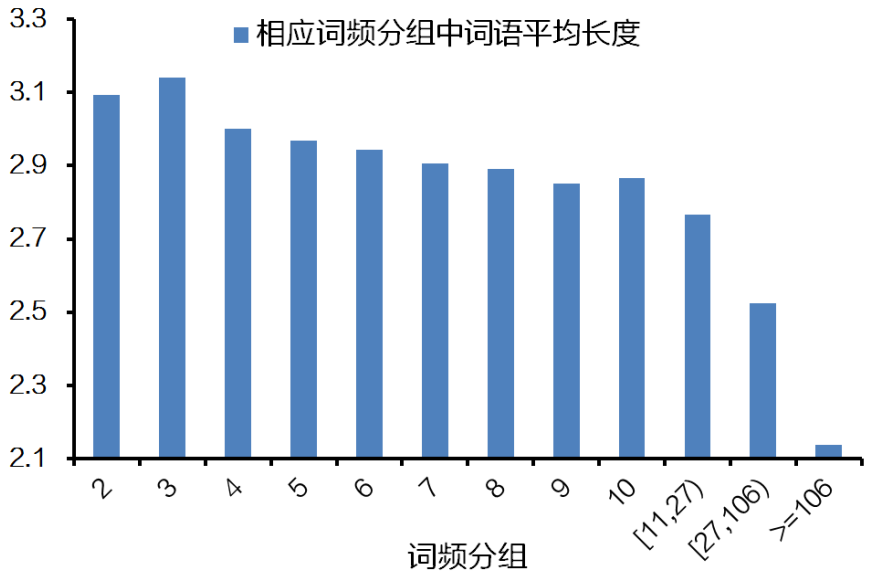
资料来源：Wind 资讯、招商证券，截止 2016 年 9 月 20 日

词频设置

- 依据实践中的经验，加入少量招商金工总结词汇，以便适应特殊形式的文本，最后将上述四部分整合搭建**招金词酷。**
- 依据 jieba 分词算法，原文词语会按照**概率连乘最大路径**来切割，**提高**（或**降低**）词库中词汇的词频，会使得这个词**能**（或**不能**）被识别出来。为了保持金融词语原义，适当**提高新收录词汇的词频**。
- 通过对 jieba 词典的词频统计，超过 70% 的词汇词频都集中在 10 以内。其中：词频大于 **10** 的词汇为“**招金词酷专属词**”；其余均为“**普通词**”。



资料来源：jieba、招商证券



资料来源：jieba、招商证券

词频设置

- 招金词酷的四部分按照重要性设置词频之间相对大小关系：

招商金工总结词汇 > Wind 词汇 > 搜狗金融词库 > jieba 词库

- 只要这四部分的词频相对大小遵循上述原则，分词效果就能达到预期目的。
- 在具体的词频数值设定上，jieba 词库的词频集中度很高，超过 70% 词汇词频都在 10 以内。因此，搜狗金融词库词频设定从 **11** 开始，其余两部分词频依次递增。

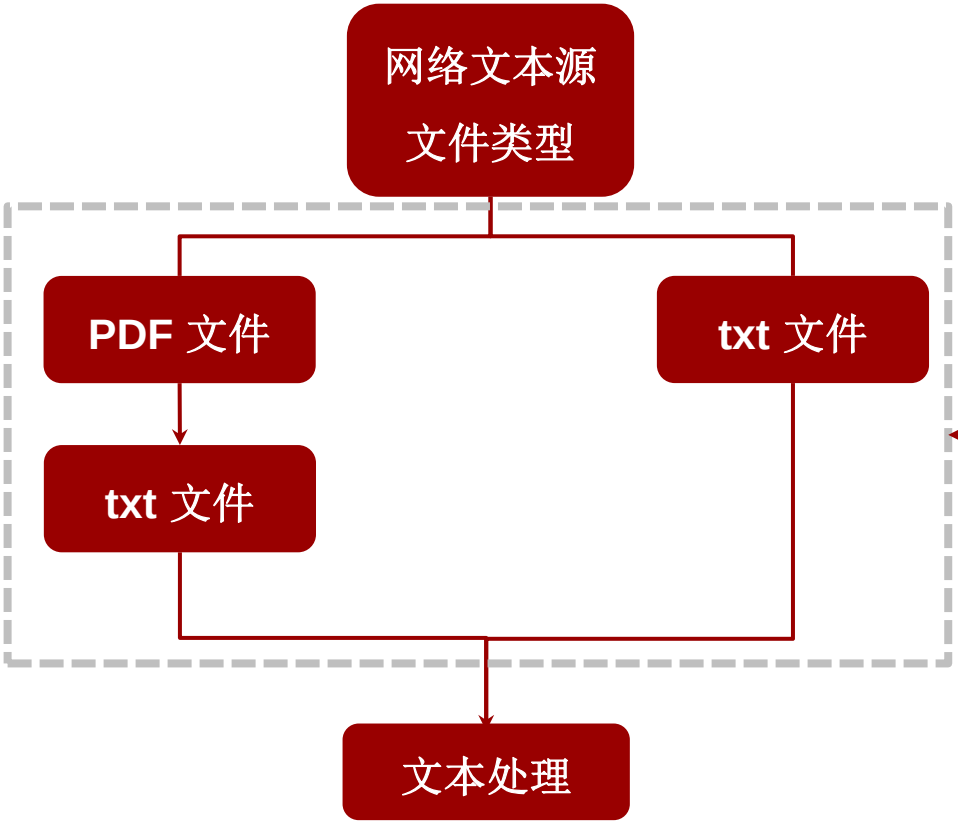
词库		词频
搜狗金融词库		11
Wind 词库词频	公司代码 “000000”	15
	公司代码 “000000.SH/SZ”	20
	公司简称	20
	公司全称	40
	高管姓名	40
招商金工总结词汇		200

资料来源：招商证券

目录

- 一、搭建招金词酷
- 二、招金词酷赢在精度
- 三、手把手教您用招金词酷
- 四、PDF 批量转 txt、HTML 工具

招金词酷赢在精度



资料来源：招商证券

- 将招金词酷和两款分词工具进行性能比较
  - txt 分词：与 IK Analyzer 进行**分词精度**的比较
  - PDF 分词：与某收费工具进行**得词率**与**分词精度**两方面的比较

txt 分词精度比较：IK Analyzer

- IK Analyzer 是一个开源的、基于 **Java** 语言开发的轻量级的中文分词工具包。
- 从 2007 年 1 月 4 日到 2016 年 7 月 25 日的业绩类公告中随机抽取 1 篇公告，分别用 IK Analyzer、招金词酷来分词，对比分词精度。

证券代码：002221 证券简称：东华能源 公告编号：2008-011

张家港东华能源股份有限公司

关于举办2007 年度业绩网上说明会的通知

本公司及董事会全体成员保证公告内容的真实、准确和完整，对公告的虚假记载、误导性陈述或者重大遗漏负连带责任。

本公司将于2008年4月15日（星期二）下午15:00-17:00在深圳证券信息有限公司提供的网上平台举行2007年度业绩网上说明会。本次业绩网上说明会采用网络远程的方式举行，投资者可登陆全景网 http://irm.p5w.net 参与年度报告说明会。

出席本次业绩说明会的人员有：公司董事长兼总经理方刚先生、副总经理华健镛先生、财务总监兼董事会秘书霍芝林先生、独立董事黄立峰先生、保荐代表人石丽女士。公司欢迎广大投资者积极参与！

特此通知。

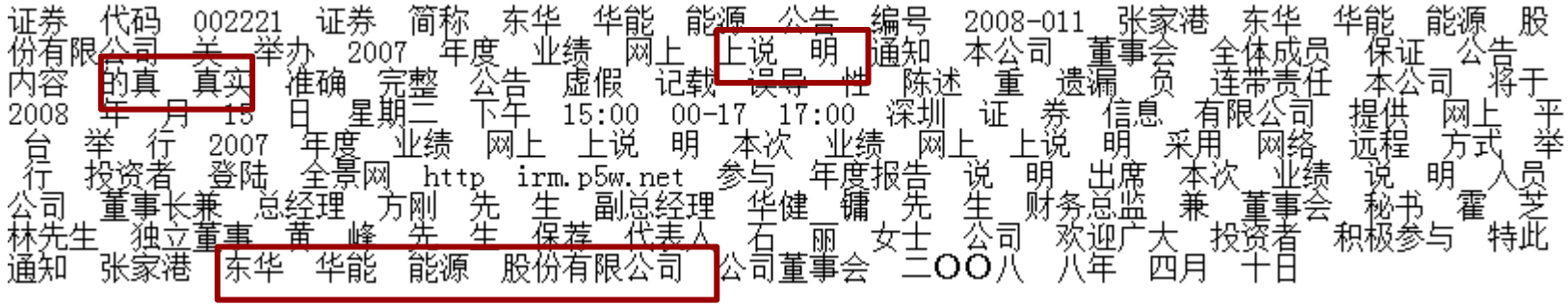
张家港东华能源股份有限公司董事会

二〇〇八年四月十日

资料来源：巨潮资讯、招商证券

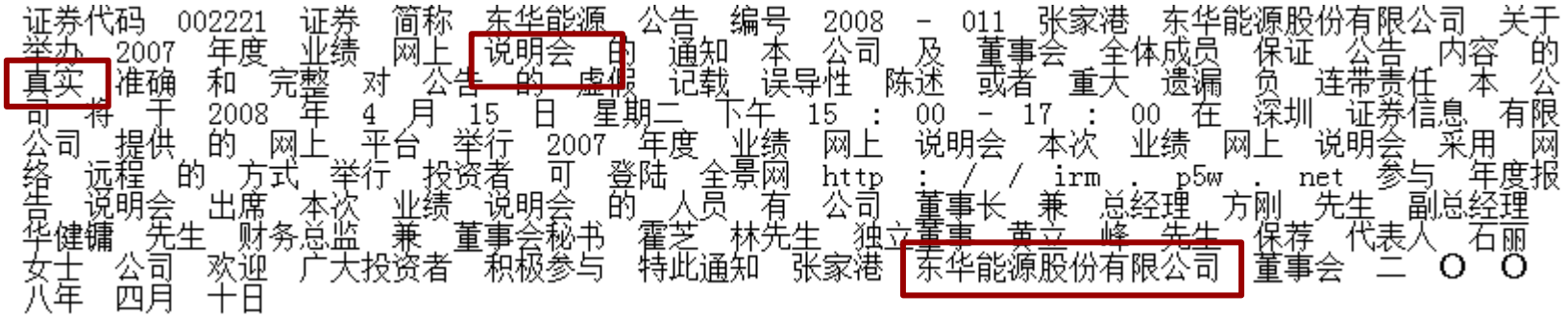
txt 分词精度比较：IK Analyzer

- IK Analyzer 不支持对 PDF 直接分词。
- 从 IK Analyzer 分词的结果来看，存在过度拆分和不当拆分的问题。


证券 代码 002221 证券 简称 东华 华能 能源 公告 编号 2008-011 张家港 东华 华能 能源 股
份有限公司 关 举办 2007 年度 业绩 网上 上说 明 通知 本公司 董事会 全体成员 保证 公告
内容 的真 真实 准确 完整 公告 虚假 记载 误导 性 陈述 重 遗漏 负 连带责任 本公司 将于
2008 年 月 15 日 星期二 下午 15:00 00-17 17:00 深圳 证 券 信息 有限公司 提供 网上 平
台 举 行 2007 年度 业绩 网上 上说 明 本次 业绩 网上 上说 明 采用 网络 远程 方式 举
行 投资者 登陆 全景网 http irm.p5w.net 参与 年度报告 说 明 出席 本次 业绩 说 明 人员
公司 董事长兼 总经理 方刚 先 生 副总经理 华健 镛 先 生 财务总监 兼 董事会 秘书 霍 芝
林先生 独立董事 黄 峰 先 生 保荐 代表人 石 丽 女士 公司 欢迎广大 投资者 积极参与 特此
通知 张家港 东华 华能 能源 股份有限公司 公司董事会 二OO八 八年 四月 十日

资料来源：招商证券

- 以下为运用招金词酷的分词结果，可以看出公司名称都被较好地分出，且不存在过度拆分现象。


证券代码 002221 证券 简称 东华能源 公告 编号 2008 - 011 张家港 东华能源股份有限公司 关于
举办 2007 年度 业绩 网上 说明会 的 通知 本 公司 及 董事会 全体成员 保证 公告 内容 的
真实 准确 和 完整 对 公告 的 虚假 记载 误导性 陈述 或者 重大 遗漏 负 连带责任 本 公
司 将 于 2008 年 4 月 15 日 星期二 下午 15 : 00 - 17 : 00 在 深圳 证券信息 有限
公司 提供 的 网上 平台 举行 2007 年度 业绩 网上 说明会 本次 业绩 网上 说明会 采用 网
络 远程 的 方式 举行 投资者 可 登陆 全景网 http : / / irm . p5w . net 参与 年度报
告 说明会 出席 本次 业绩 说明会 的 人员 有 公司 董事长 兼 总经理 方刚 先生 副总经理
华健镛 先生 财务总监 兼 董事会秘书 霍芝 林先生 独立董事 黄立 峰 先生 保荐 代表人 石丽
女士 公司 欢迎 广大投资者 积极参与 特此通知 张家港 东华能源股份有限公司 董事会 二 O O
八年 四月 十日

资料来源：招商证券

PDF 分词比较：某收费工具

- 从巨潮资讯（http://www.cninfo.com.cn/cninfo-new/index）抓取 2007 年 1 月 4 日到 2016 年 7 月 25 日的 45442 篇业绩类公告（PDF 格式），从中随机抽样一定数量的公告，与某收费工具进行比较。
- 抽样规则如下图，在本文之后的分析中，均采用相同的抽样规则。

资料来源：招商证券

- 由于该款收费工具**不支持批量导入分词功能**，因此按照上述抽样规则，从 45442 篇 PDF 业绩类公告中随机抽取 **50** 篇，从**得词率**和**分词精度**两方面来进行比较。

PDF 分词比较：某收费工具

CMS 招商证券

●随机抽取的 50 篇业绩类公告

公告编号	公告日期	公司代码	公告名称
1	20070413	000861	S*ST托普：业绩修正公告
2	20080129	600230	江苏阳光：2007年度业绩预增公告
3	20080411	600028	东华能源：关于举办2007年度业绩网上说明会的通知
4	20080627	600572	北海国发：2008年上半年业绩预亏公告
5	20080724	000423	潍柴动力：2008年上半年业绩预增公告
6	20100728	600689	宏达股份：2010年半年度业绩预亏公告
7	20101029	600854	华北制药：2010年度业绩预增公告
8	20110127	002265	威华股份：2010年度业绩预告的修正公告
9	20110412	600292	三一重工：2011年一季度业绩预增公告
10	20120110	600644	*ST盛工：2011年度业绩快报
11	20120222	002613	北玻股份：2011年度业绩快报更正公告
12	20120321	002033	博闻科技：2011年度业绩快报
13	20120329	300093	国民技术：关于举办2011年年度报告网上业绩说明会的公告
14	20120330	002528	江苏旷达：2012年度第一季度业绩预告
15	20120714	002416	远东传动：2012年半年度业绩预告修正公告
16	20121225	600751	开创国际：2012年度业绩预增公告
17	20130115	600726	海岛建设：2012年年度业绩预盈公告
18	20130131	000733	*ST能山：2012年年度业绩预告公告
19	20130131	002180	梅 花 伞：2012年度业绩预告修正公告
20	20130223	002314	濮耐股份：2012年度业绩快报
21	20130327	300120	华谊嘉信：长城证券有限责任公司关于公司收购上海东汐广告传播有限公司2012年度业绩承诺实现情况的核查意见
22	20130328	002449	九九久：2013年第一季度业绩预告
23	20130404	300015	立思辰：2013年第一季度业绩预告
24	20130411	300063	博实股份：关于举行2012年度网上业绩说明会的通知
25	20130413	002125	孚日股份：2013年一季度业绩快报

资料来源：招商证券

PDF 分词比较：某收费工具

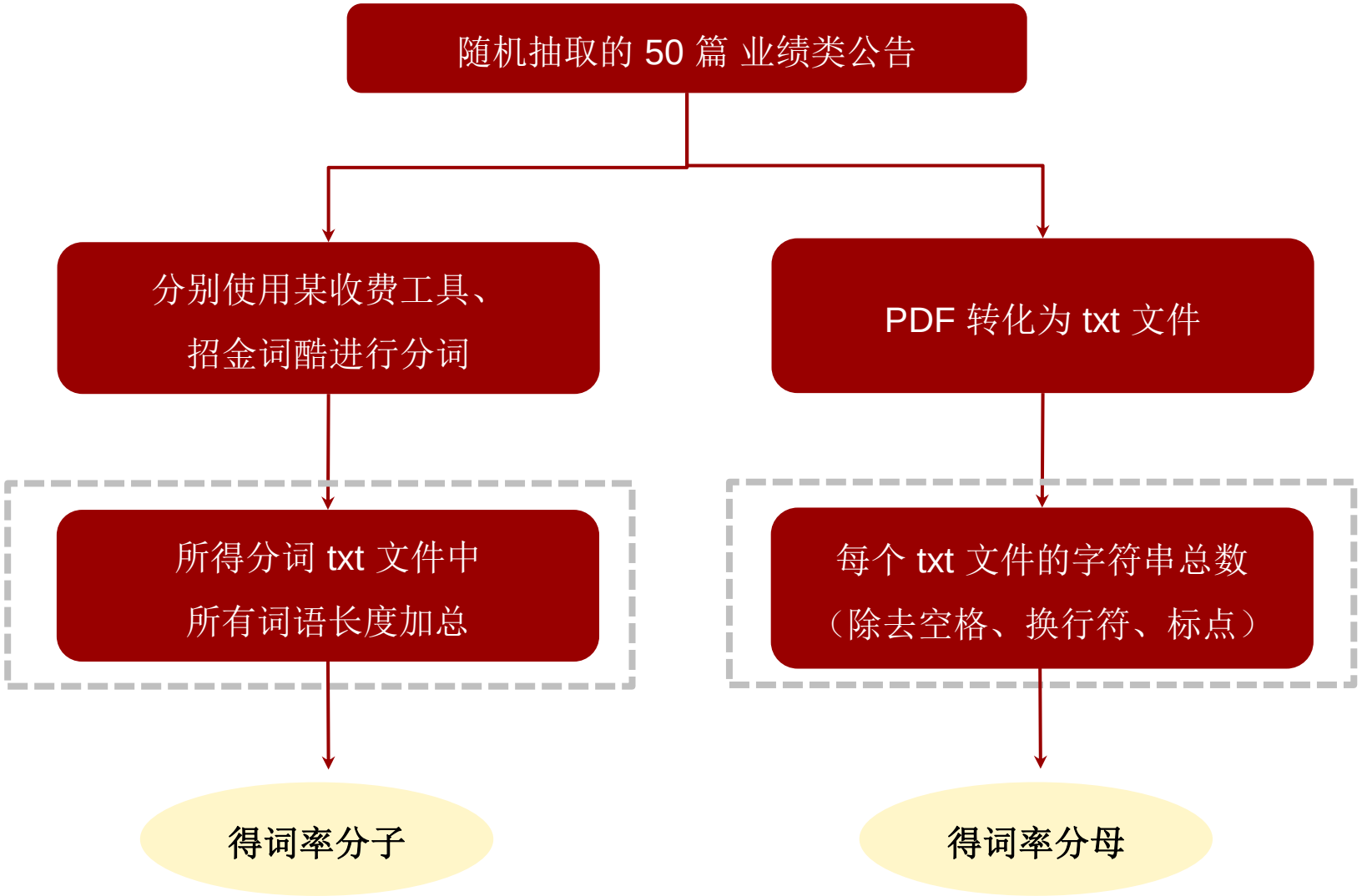
CMS 招商证券

●随机抽取的 50 篇业绩类公告

公告编号	公告日期	公司代码	公告名称
26	20130509	600730	华仪电气：关于2012年现场业绩说明会暨投资者接待日活动召开情况的公告
27	20130725	601179	渤海活塞：2013半年度业绩快报公告
28	20140116	600738	电子城：2013年度业绩快报
29	20140125	600058	中江地产：2013年度业绩预增公告
30	20140128	300264	光线传媒：2013年度业绩预告
31	20140129	600158	航天机电：2013年度业绩快报公告
32	20140328	000966	神火股份：2014年第一季度业绩预亏公告
33	20140418	601107	晋亿实业：关于召开2013年年度业绩说明会的预告公告
34	20140521	601168	广誉远：关于举行2013年度报告网上业绩说明会的公告
35	20140618	600720	亿利能源：关于召开2013年度业绩说明会的通知
36	20140702	300383	旋极信息：2014年半年度业绩预告
37	20140714	300122	智飞生物：2014年半年度业绩预告
38	20150117	000416	许继电气：2014年度业绩预告
39	20150129	600217	北方稀土：2014年度业绩预减公告
40	20150131	600533	中发科技：2014年年度业绩预亏公告
41	20150216	002219	科陆电子：2014年度业绩快报
42	20150228	300108	乐视网：2014年度业绩快报
43	20150306	300054	尤夫股份：关于举办网上业绩说明会和投资者接待日活动的通知
44	20150410	300010	立思辰：关于交易对手方对置入资产2014年度业绩承诺实现情况的说明审核报告（一）
45	20150429	002608	宝鼎重工：关于举行2014年度网上业绩说明会的通知
46	20150714	300285	汇冠股份：2015年半年度业绩预告
47	20150715	300466	高伟达：2015年半年度业绩预告修正公告
48	20151014	000557	佛山照明：2015年前三季度业绩预告
49	20160322	300017	网宿科技：2016年第一季度业绩预告
50	20160709	000626	吉林敖东：关于广发证券2016年半年度业绩预告的公告

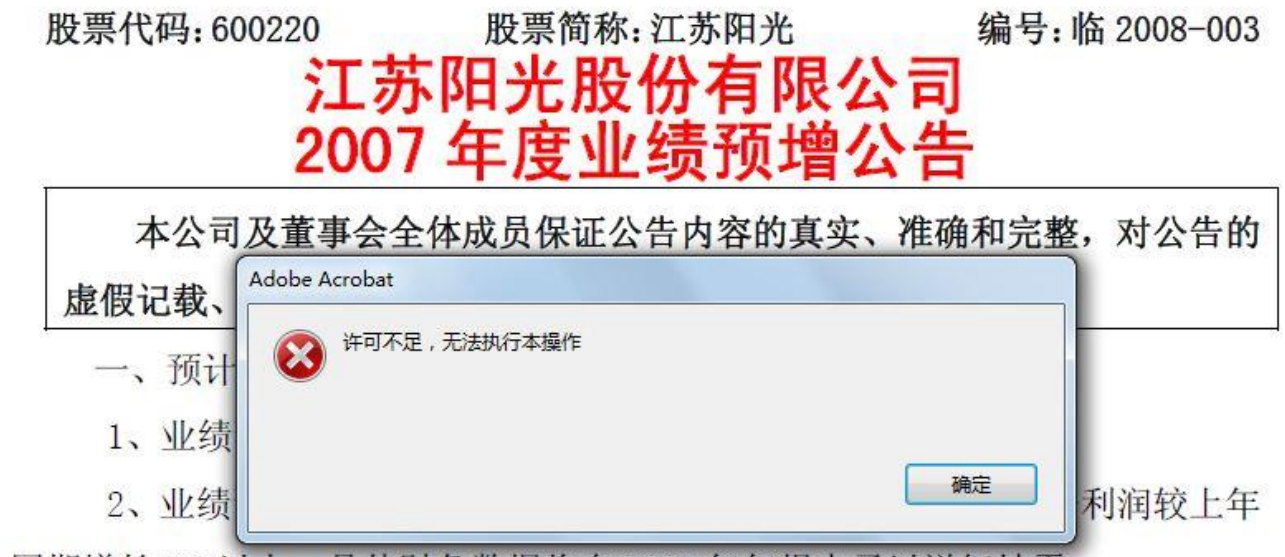
资料来源：招商证券

得词率


随机抽取的 50 篇 业绩类公告
分别使用某收费工具、
招金词酷进行分词
PDF 转化为 txt 文件
所得分词 txt 文件中
所有词语长度加总
每个 txt 文件的字符串总数
（除去空格、换行符、标点）
得词率分子
得词率分母

PDF 分词比较：得词率

- 在抽取的 50 篇 PDF 业绩类公告中，有 1 篇由于 PDF 文件有加密，转化成 txt 文件失败，因此可用于实际用于计算得词的公告为 49 篇。


股票代码：600220　　股票简称：江苏阳光　　编号：临 2008-003

江苏阳光股份有限公司
2007 年度业绩预增公告

本公司及董事会全体成员保证公告内容的真实、准确和完整，对公告的虚假记载、

Adobe Acrobat

许可不足，无法执行本操作

确定

一、预计

1、业绩

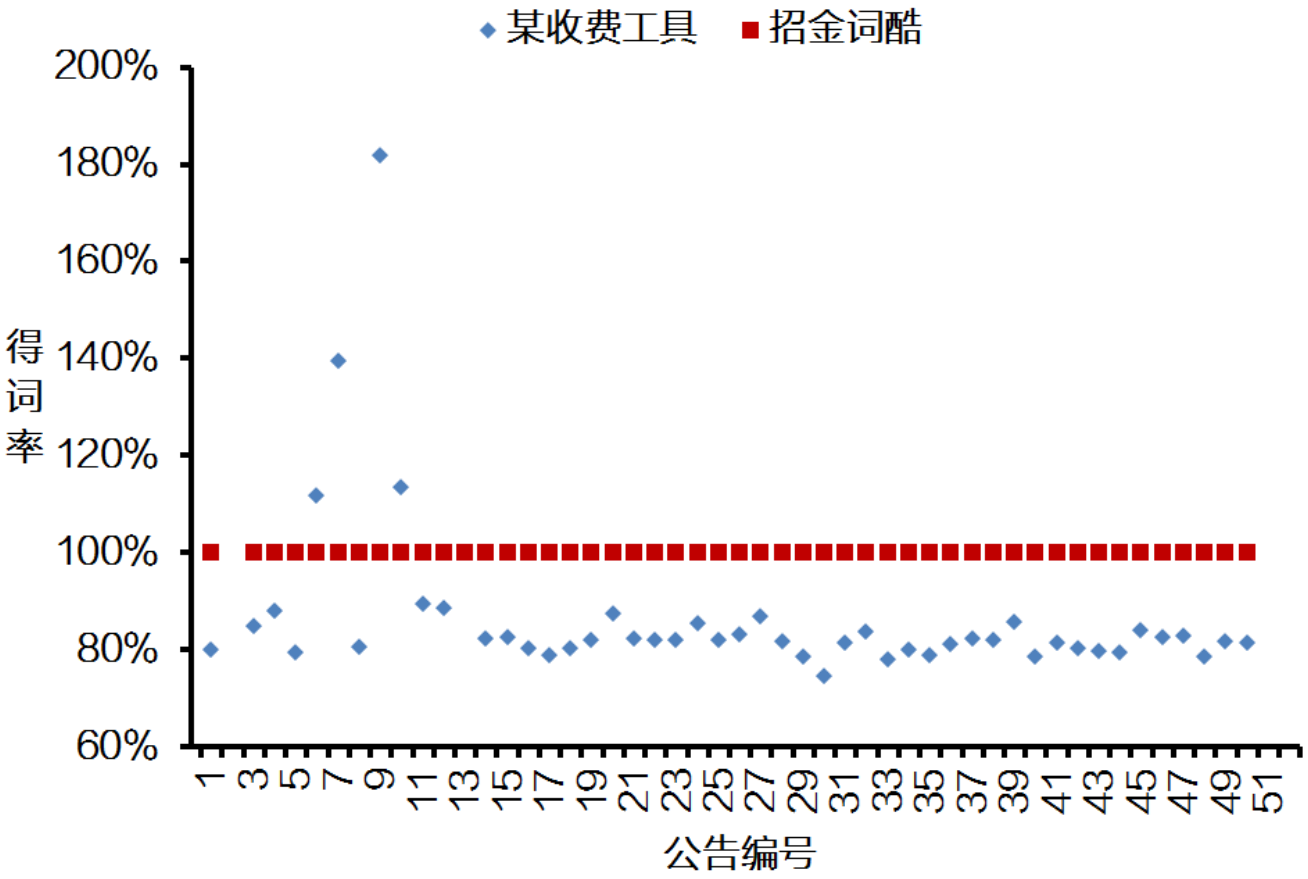
2、业绩　　利润较上年同期增长 50%以上，具体财务数据将在 2007 年年报中予以详细披露。

资料来源：招商证券

- 在 PDF 转化成 txt 文件成功的 **49** 篇 PDF 业绩类公告中，**招金词酷全部分词成功**，该收费工具有 **1** 篇分词失败。因此在之后**得词率**的计算比较中，该收费工具仅有 **48** 个点，招金词酷则有 **49** 个点。

PDF 分词比较：得词率

- 从得词率结果可以看出，该款收费工具相较招金词酷而言，存在两方面问题：
  - 得词率普遍小于 1，说明文件类型**转化**时词语**损失**量较大；
  - 得词率异常大于 1，说明存在**过度拆分**以及词语大量**重复出现**的现象。



资料来源：招商证券

某收费工具 PDF 分词精度展示

- 该收费工具虽能直接对 PDF 进行分词，但是对 PDF 中的表格部分处理效果不佳，且会出现丢词及过度分词的现象。

证券 代码 002427 证券 简称 尤夫 股份 公告 编号 2015 016 浙江 尤夫 高新 纤维 股份有限公司 举办 网上 业绩 说明 投资者 接待日 活动 通知 公司 董事会 全体 成员 保证 信息 披露 内容 真实 准确 完整 虚假 记载 误导 性 陈述 重大 遗漏 进一步 开展 浙江 尤夫 高新 纤维 股份有限公司 以下 简称 公司 投资者 关系 管理 活动 增进 公司 广大 投资者 沟通 交流 公司 举办 网上 业绩 说明 活动 投资者 接待日 具体 事项 公告 网上 业绩 说明 安排 公司 年度 业绩 说明 安排 活动 时间 2015 年 月 13 日 星期 15 00 17 00 召开 方式 利用 深圳 证券 信息 有限公司 提供 网上 平台 采用 网络 远程 方式 举行 投资 登陆 http irm p w net 参与 公司 2014 年度 业绩 说明 接待 人员 公司 董事长 兼 总经理 茅惠新 先生 财务 负责 人 兼 董事会 秘书 陈彦 先生 独立 董事 王华 平 先生 投资者 接待日 活

资料来源：招商证券

招金词酷 PDF 分词精度展示

CMS 招商证券

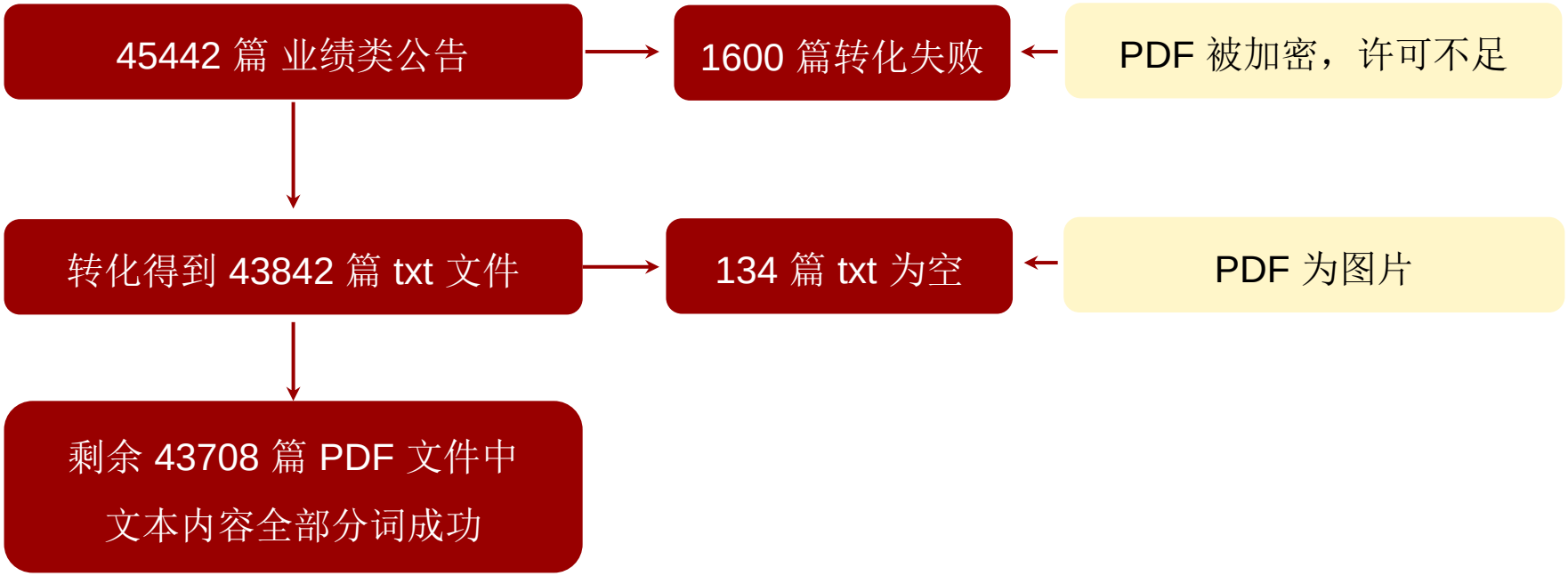
- 招金词酷分词优势凸显：
  - 支持批量导入分词功能
  - 信息保留度高
  - 处理金融词汇具有明显优势

证券代码 002427 证券 简称 尤夫股份 公告 编号 2015 - 016
浙江尤夫高新纤维股份有限公司 关于 举办 网上 业绩 说明会 和 投资者 接待日
活动 的 通知 本 公司 及 董事会 全体成员 保证 信息披露 的 内容 真实
准确 完整 没有 虚假 记载 误导性 陈述 或 重大 遗漏 为 进一步 开展
浙江尤夫高新纤维股份有限公司 以下简称 公司 投资者关系 管理 活动 增进 公司
与 广大投资者 的 沟通 与 交流 公司 将 举办 网上 业绩 说明会 活动 与
投资者 接待日 具体 事项 公告 如下 一 网上 业绩 说明会 安排 公司 年度
业绩 说明会 安排 如下 1 活动 时间 2015 年 3 月 13 日 星期五 15 :
0017 : 002 召开 方式 利用 深圳 证券信息 有限公司 提供 的 网上 平台
采用 网络 远程 的 方式 举行 投资者 可 登陆 http : / / irm . p5w .
net 参与公司 2014 年度 业绩 说明会 3 接待 人员 公司 董事长 兼 总经理
茅惠新 先生 财务 负责人 兼 董事会秘书 陈彦 先生 独立董事 王华平 先生

资料来源：招商证券

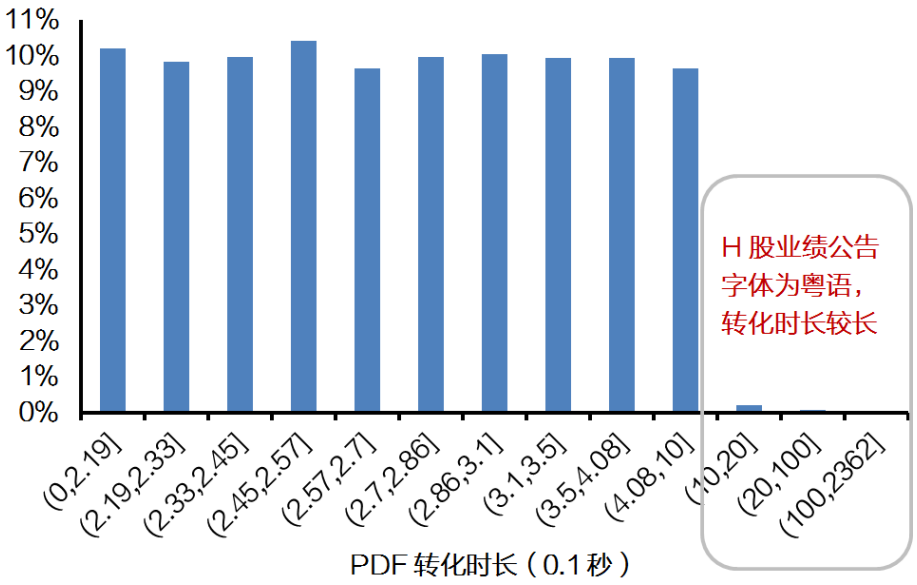
招金词酷批量分词结果

- 对 2007 年 1 月 4 日到 2016 年 7 月 25 日的 45442 篇业绩类公告进行分词，分词成功的共有 **43708** 篇，成功率高达 **96.18%**。



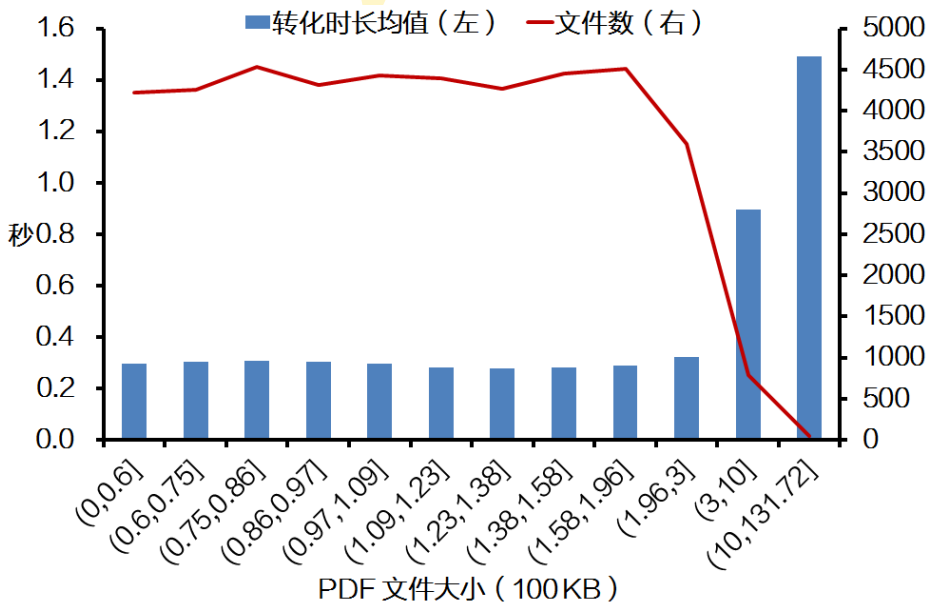
资料来源：招商证券

招金词酷 PDF 批量转化时长



资料来源：招商证券

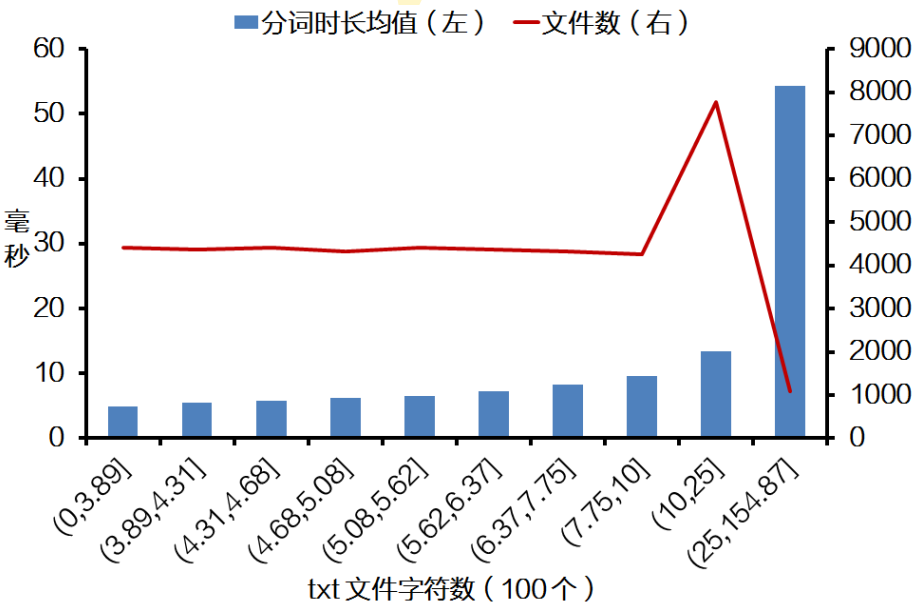
99% 的公告
转化时长不足 **1** 秒



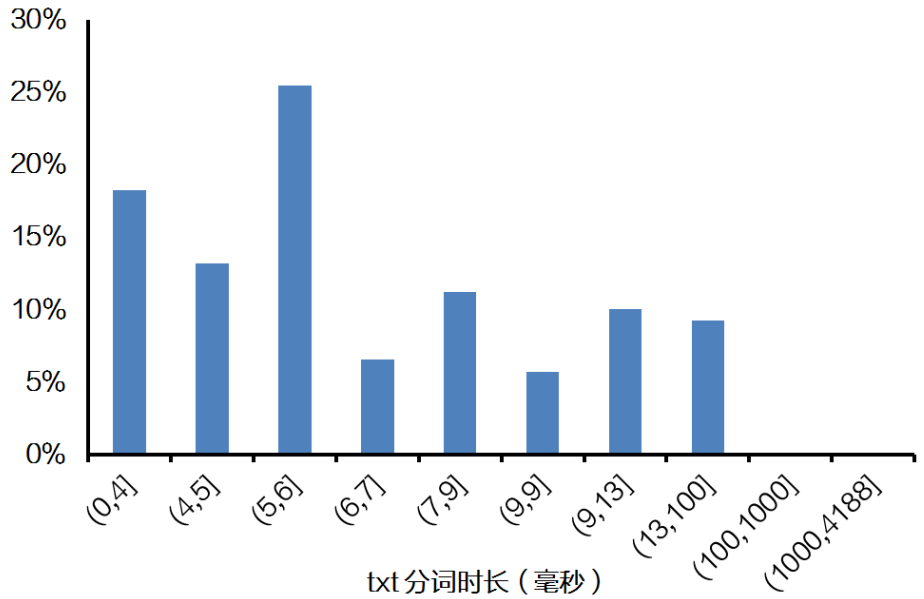
资料来源：招商证券

招金词酷批量分词时长

99% 的公告
分词时长不足 **0.1** 秒

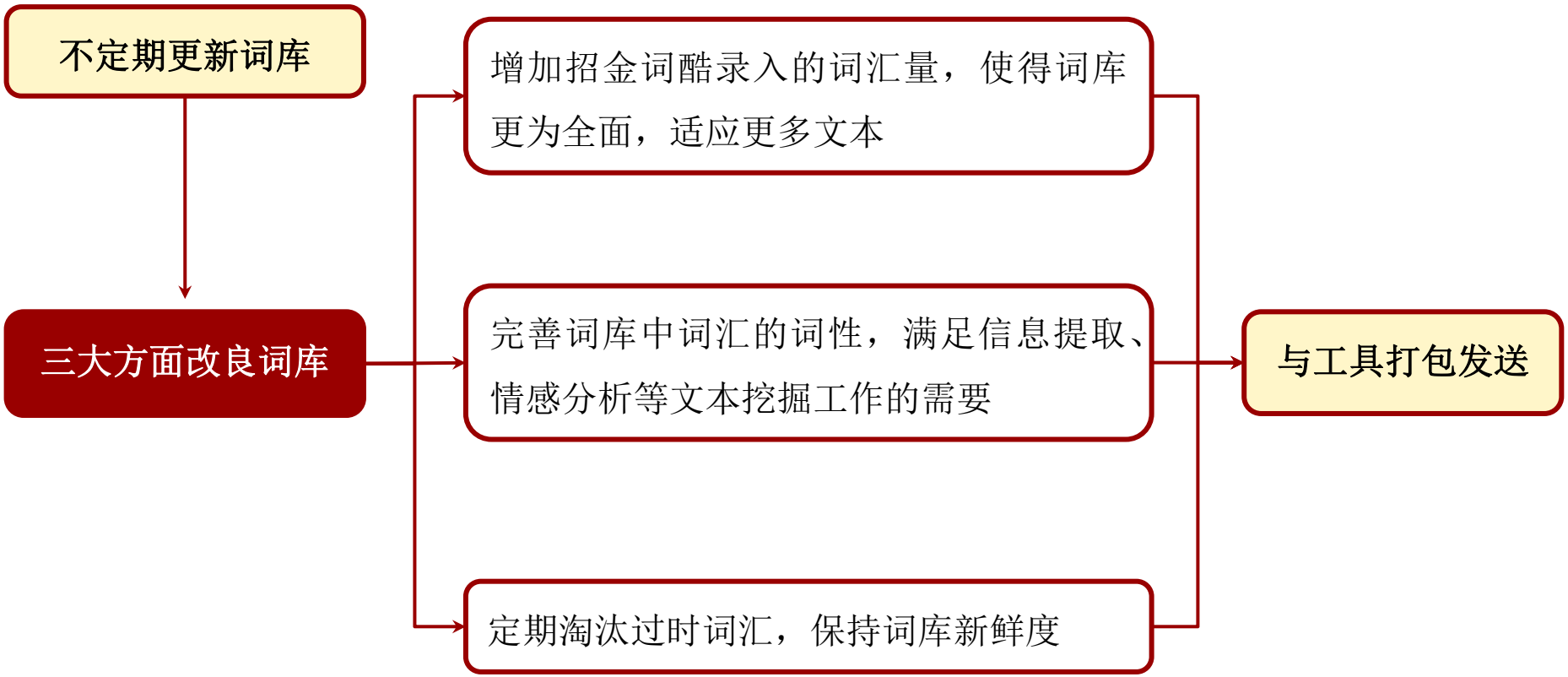


资料来源：招商证券



资料来源：招商证券

招金词酷维护方式



资料来源：招商证券

目录

- 一、搭建招金词酷
- 二、招金词酷赢在精度
- 三、手把手教您用招金词酷
- 四、PDF 批量转 txt、HTML 工具

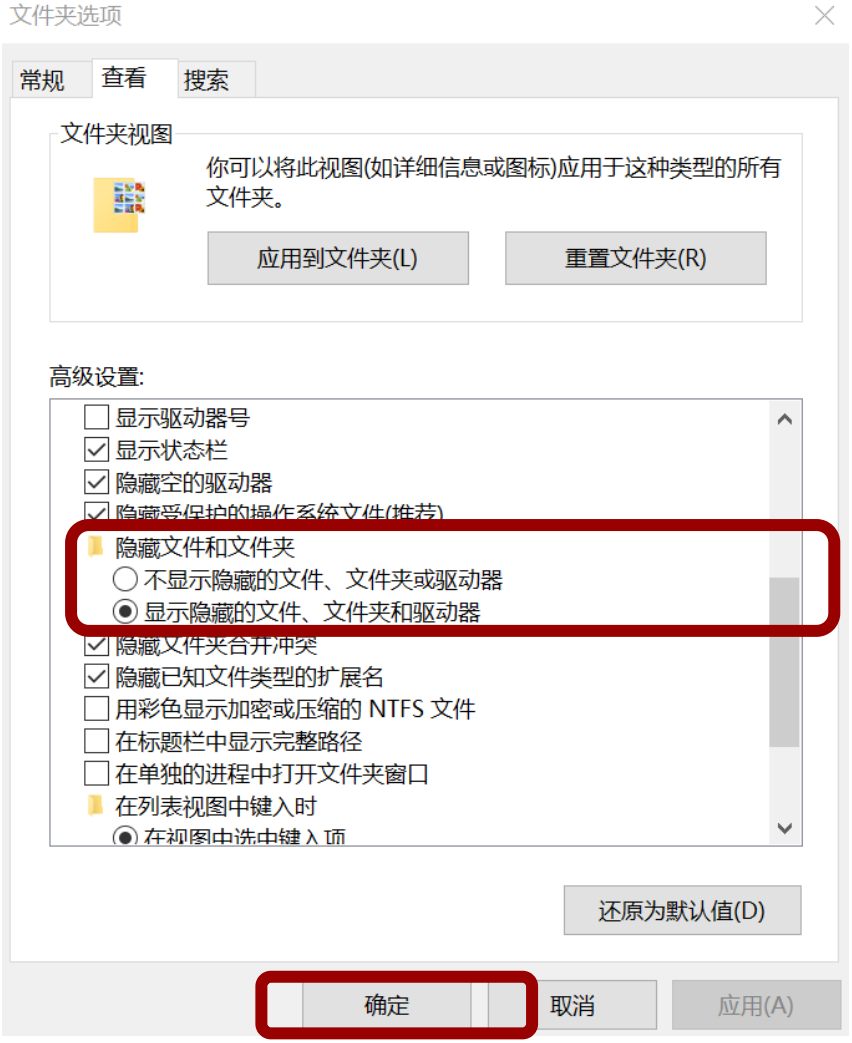
安装步骤

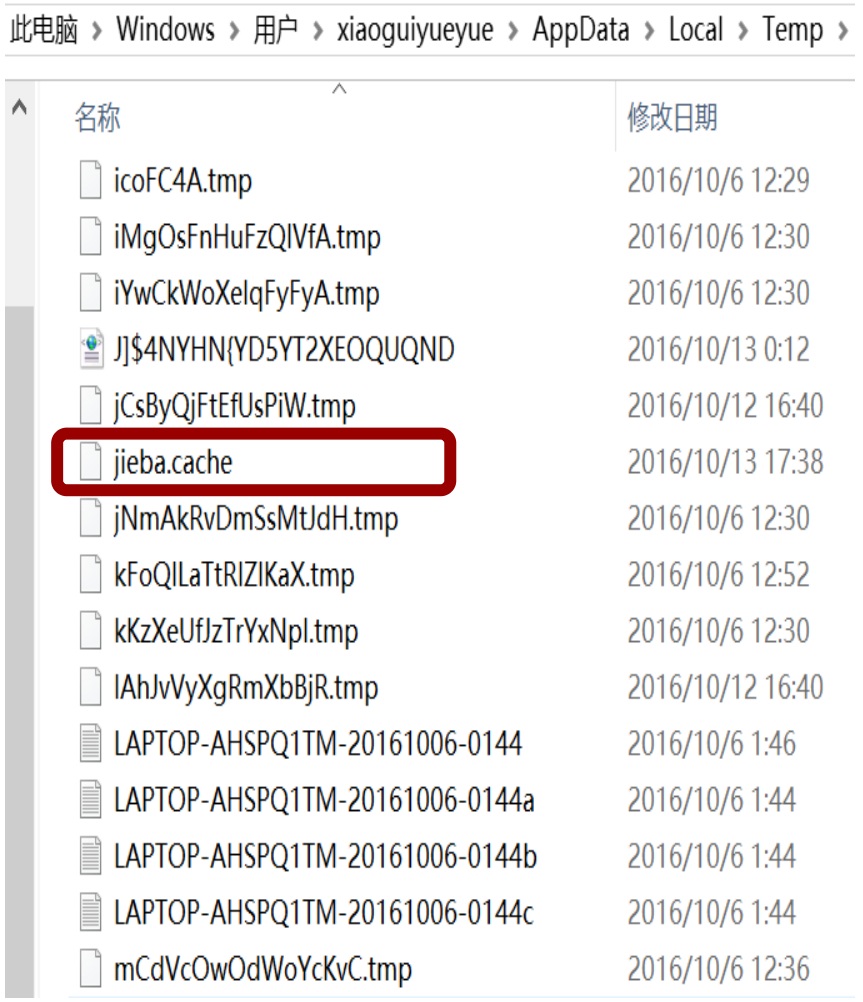
- 下载地址：http://pan.baidu.com/s/1boKzTdt　密码: 35vf
- 我们提供 32 位、64 位系统下分别对应的招金分词工具。
- 下载 Adobe Acrobat DC 软件：解压“**招金词酷 \ CFamily_Acrobat_XP85**”；在解压后的 ROOT 文件夹下点击 Setup.exe 即可进行安装。

资料来源：招商证券

安装步骤

- 打开“文件夹选项”，暂时将文件的**隐藏属性去掉**
- 将“**jieba.cache**”拷贝到“**C:\Users\~\AppData\Local\Temp**”，**恢复**文件夹选项**设置**

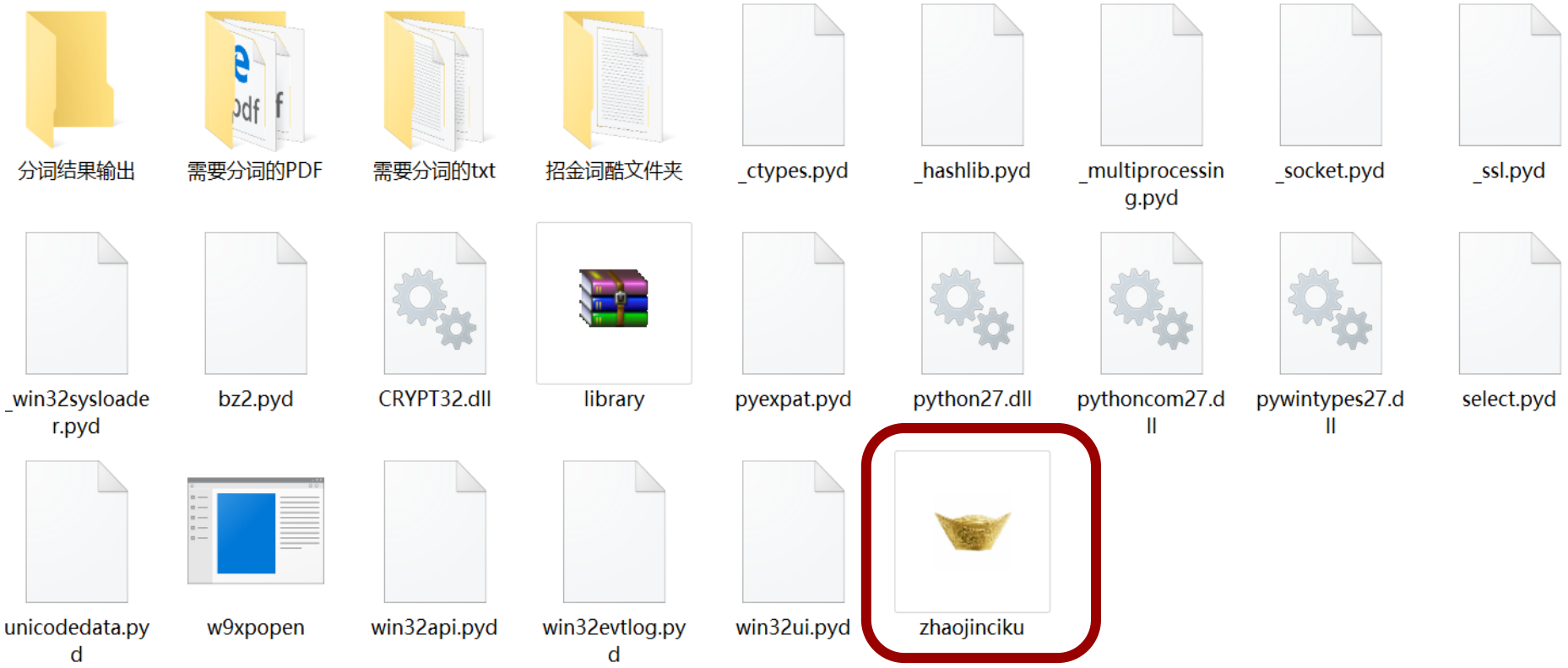




资料来源：招商证券

使用方法

- “【32位】招金分词工具”与“【64位】招金分词工具”使用方法类似，介绍前者为例。
- 解压“【32位】招金分词工具.rar”，并打开文件夹，见如下界面。

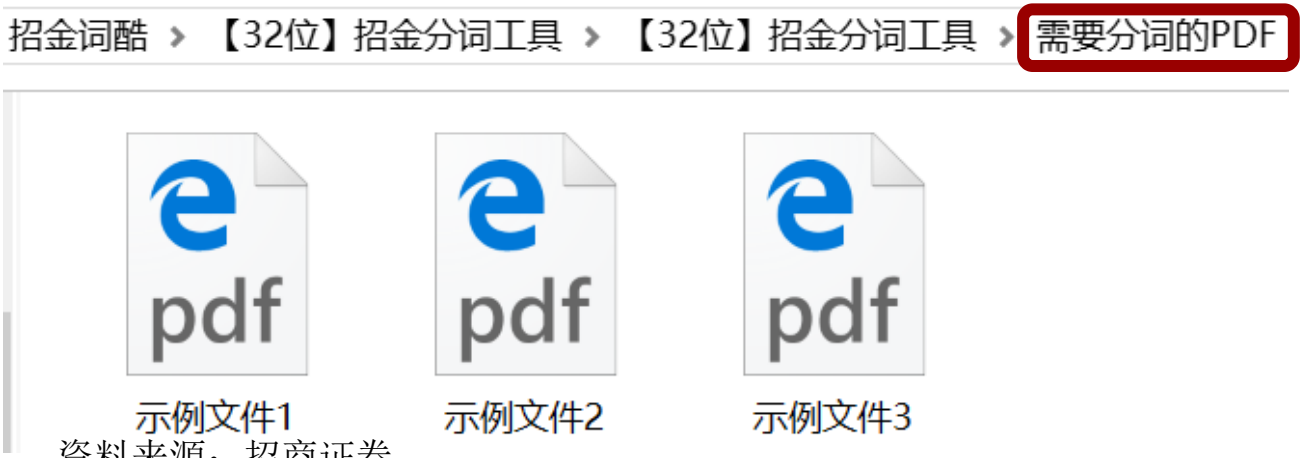


资料来源：招商证券

- **Tips！**
  - 请不要随意修改或移动上图所示文件夹及文件
  - 目前仅支持 Windows 系统

使用方法

- 若对 PDF 文件进行分词，将所有需要分词的 PDF 文件放入到“**需要分词的 PDF**”



资料来源：招商证券

- 若要对 txt 文件进行分词，将所有需要分词的 txt 文件放入到“**需要分词的 txt**”

资料来源：招商证券

- **Tips！**
  - 放入需要分词文件前，请保证这两个文件夹下无其他类型文件

使用方法

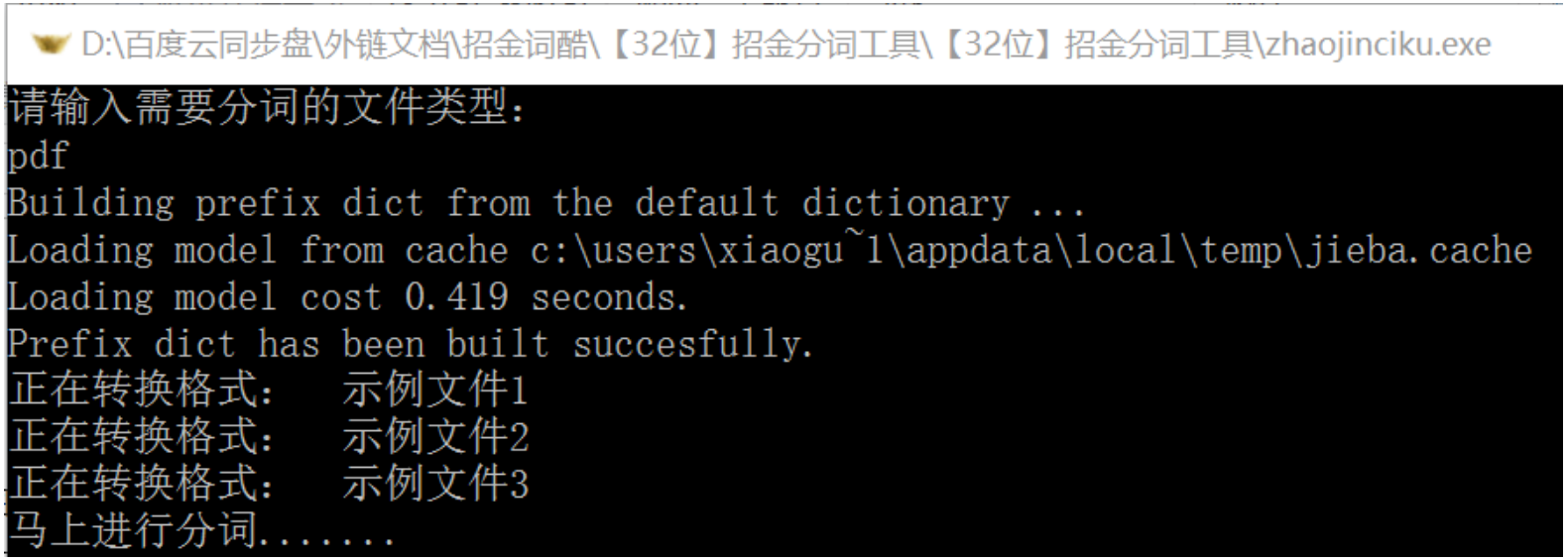
- 点击 “zhaojinciku.exe”，出现如下界面：

PDF 分词，请输入“**pdf**”，并回车

txt 分词，请输入“**txt**”，并回车

资料来源：招商证券

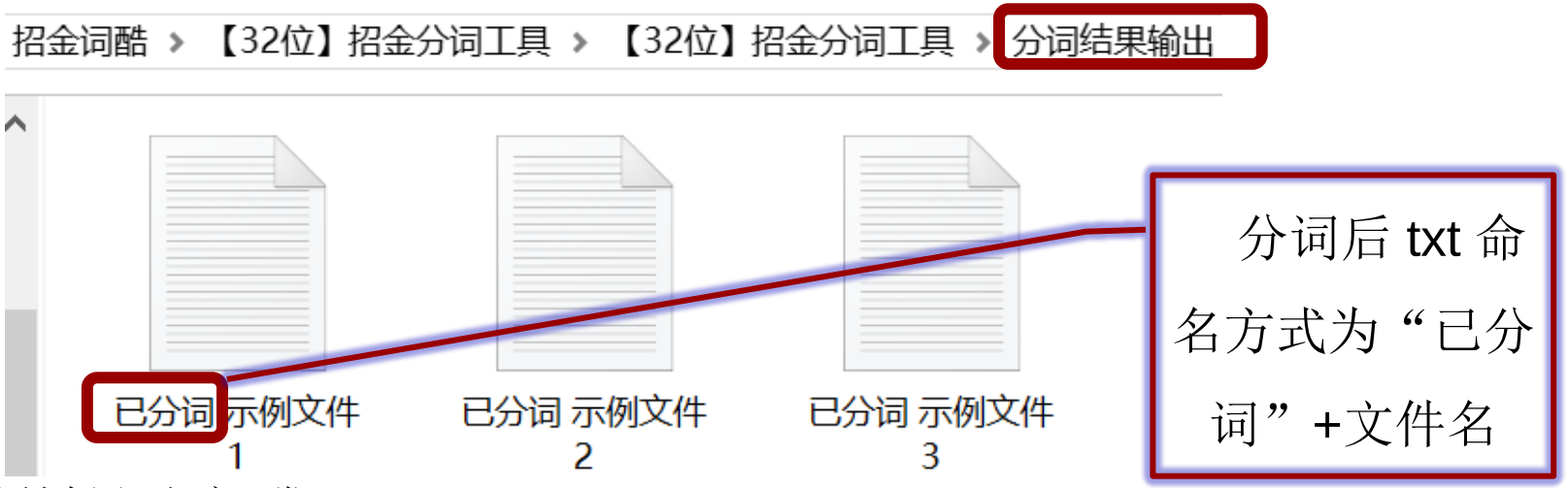
- 程序运行界面：



资料来源：招商证券

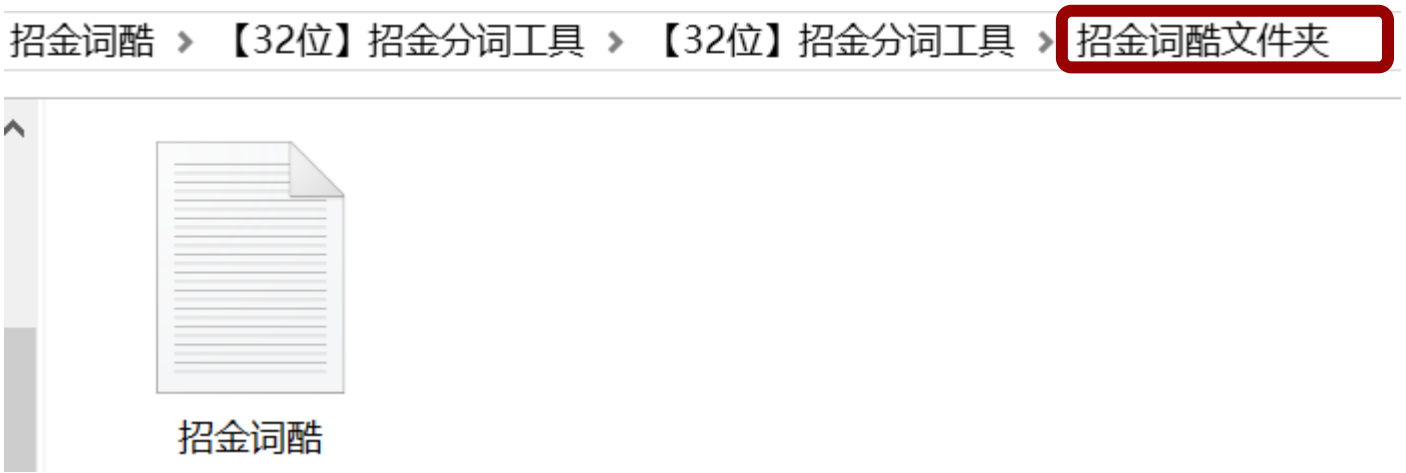
使用方法

- 程序运行完毕后，可在“**分词结果输出**”文件夹下找到已经分好词的 txt 文件



资料来源：招商证券

- 本团队提供的招金词酷放置在“**招金词酷文件夹**”中，分词时自动载入



资料来源：招商证券

目录

- 一、搭建招金词酷
- 二、招金词酷赢在精度
- 三、手把手教您用招金词酷
- **四、PDF 批量转 txt、HTML 工具**

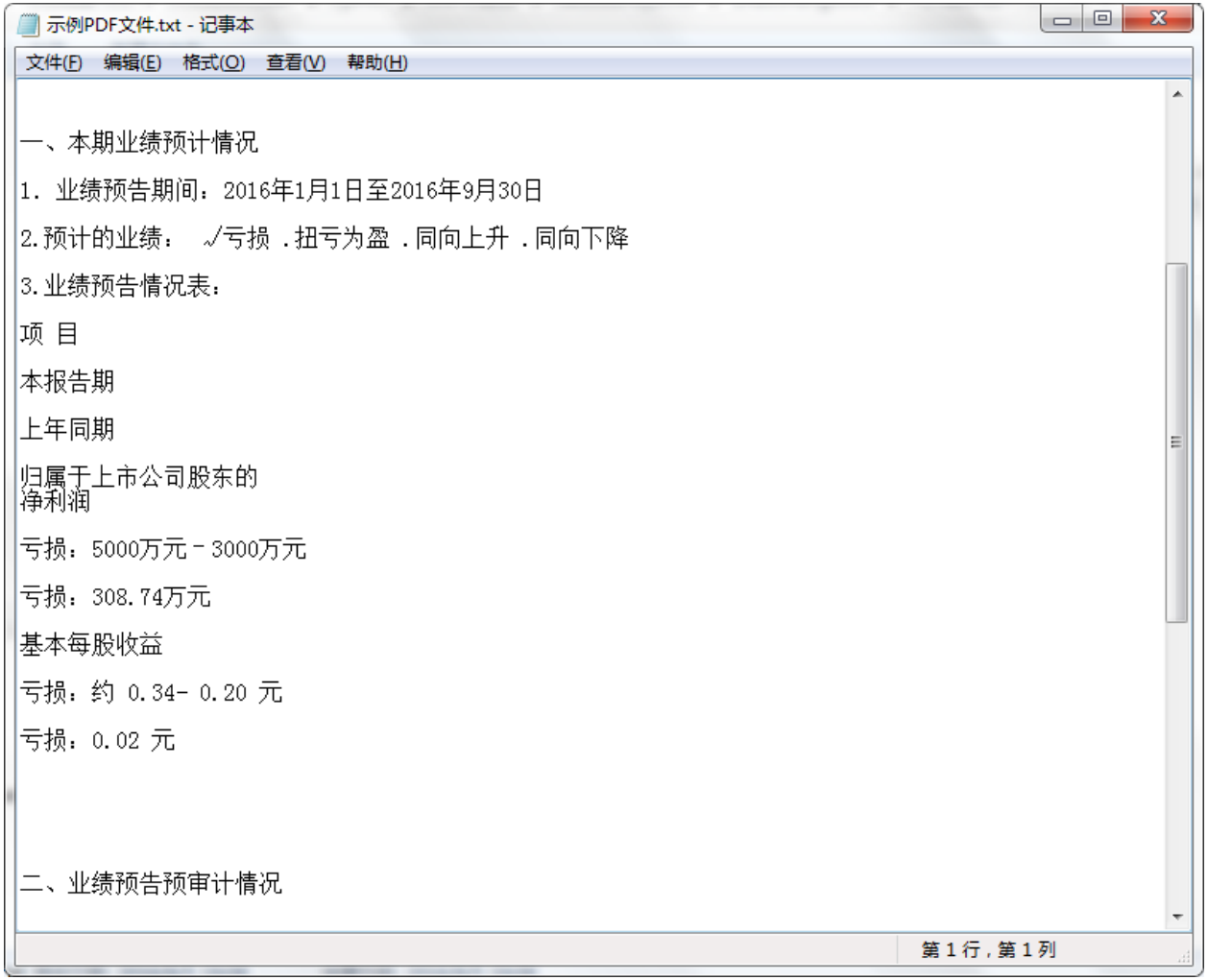
PDF 批量转 txt、HTML 工具

- 很多金融类文本为 PDF 格式，因此招商金工团队基于 python 提供一个**可批量**将 **PDF** 转为 **txt** 或 **HTML** 的小工具
- 优势：**免费**、**信息保留度高**
- 接下来我们选取一篇业绩类 PDF 公告来做示例：

资料来源：巨潮资讯、招商证券

格式转换效果展示

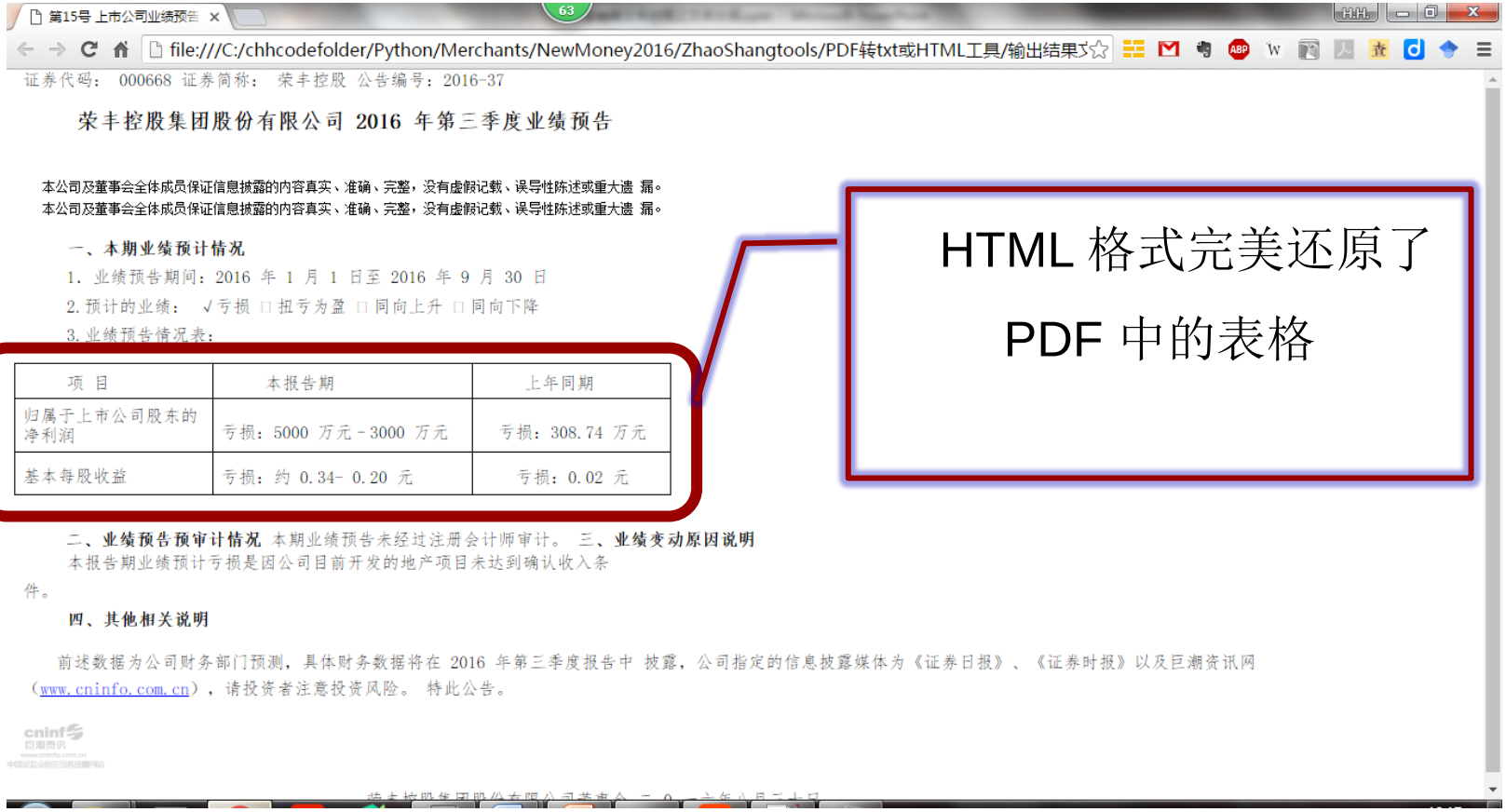
- 转换后的 txt 文件



资料来源：招商证券

格式转换效果展示

- 转换后的 html 文件



资料来源：招商证券

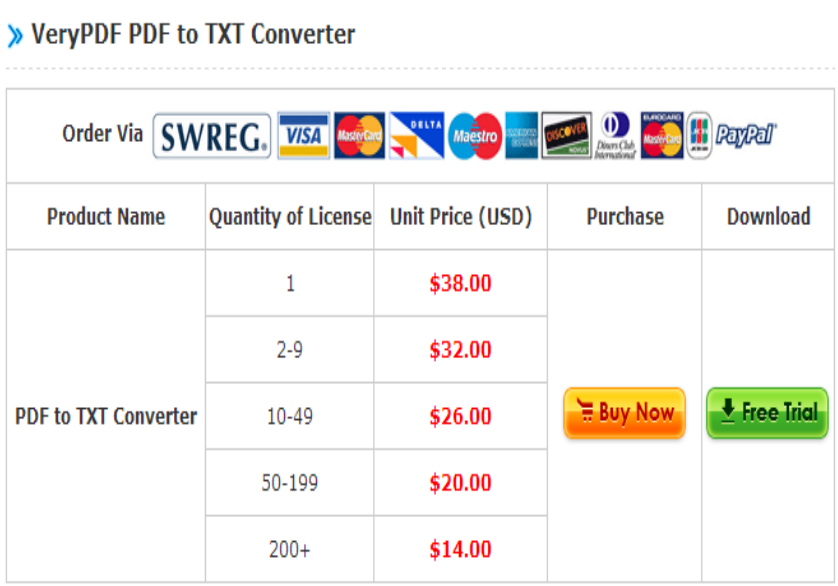
与 PDF2TXT 比较

- 下面将招商金工的 PDF 转换工具与 PDF2TXT 进行比较。

	PDF2TXT	招商金工 PDF 转换工具
费用	收费	免费
可转格式	仅为 txt	HTML、txt

资料来源：招商证券

- 由于 HTML 文件运用了大量标签，因此这种格式对于**抓取表格数据**有很大优势，可以**大大减少搜索范围**，实现**准确定位**。

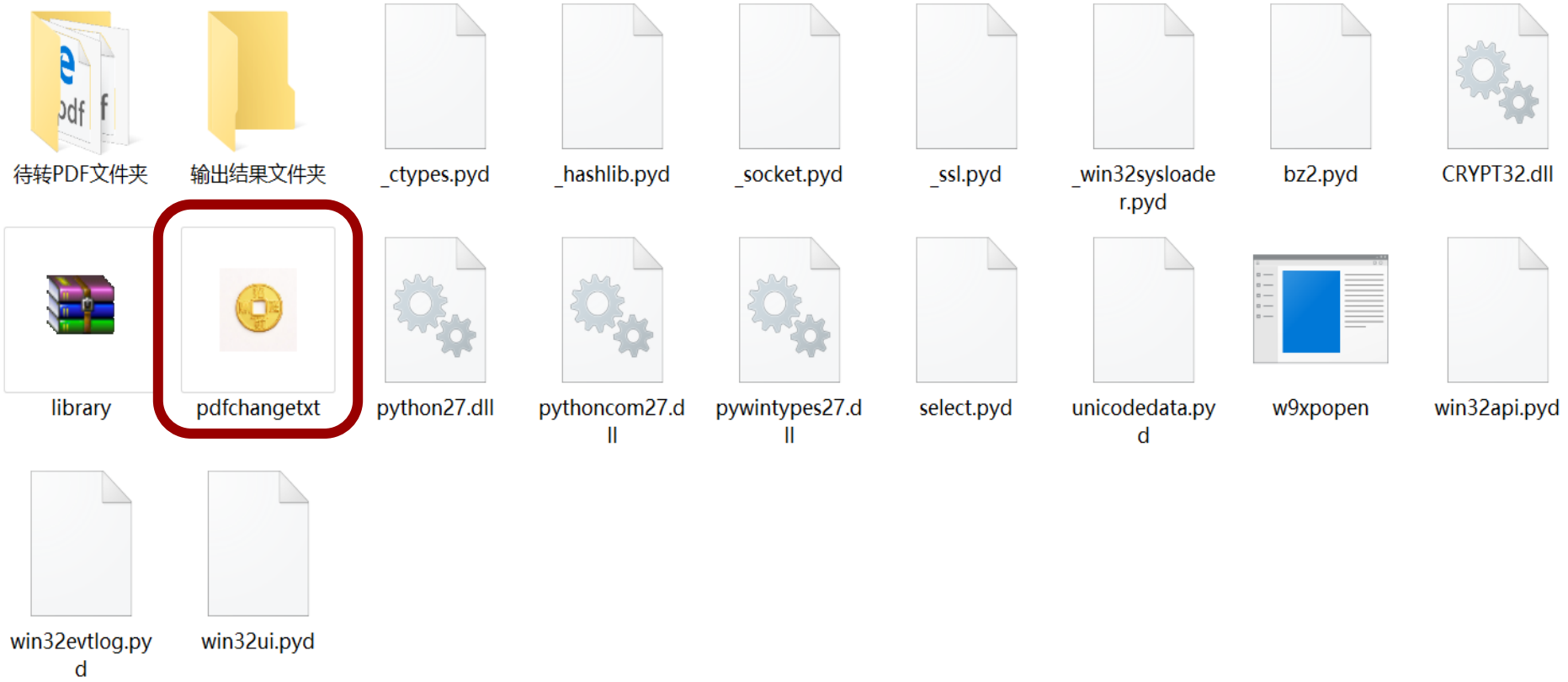


资料来源：互联网

资料来源：招商证券

安装步骤

- 安装 Adobe Acrobat DC，具体方法请见“手把手教您用招金词酷”
- 解压“【32位】招金PDF转txt或HTML工具”

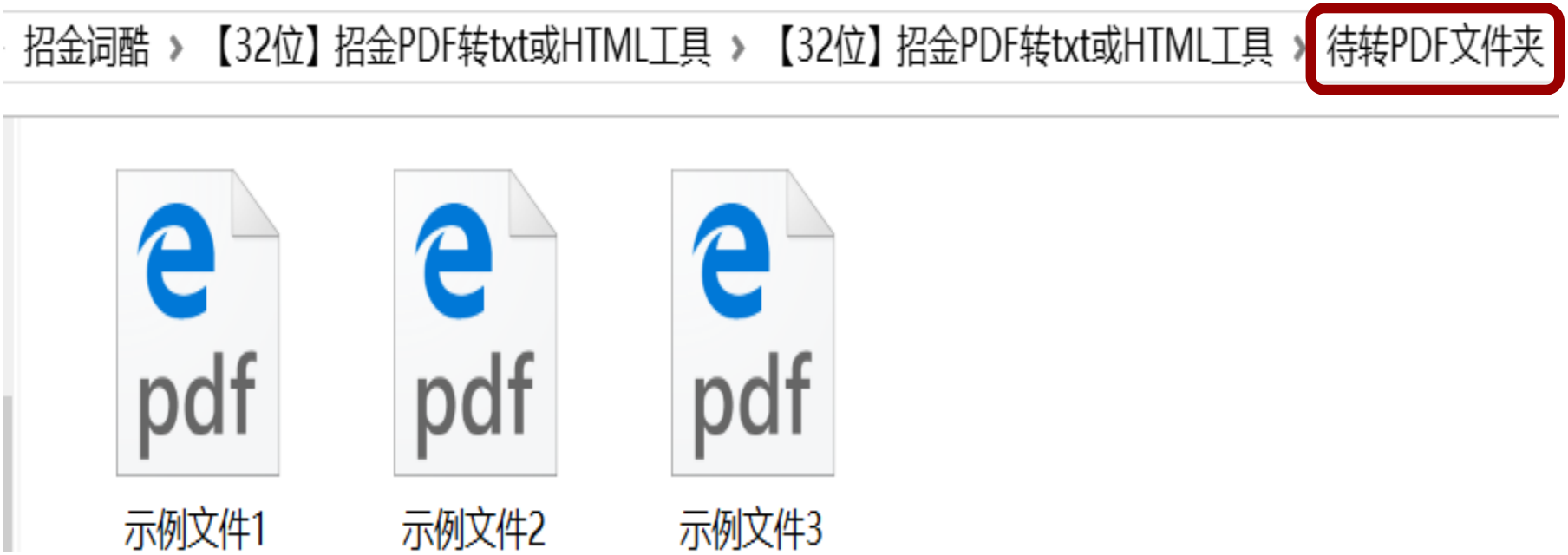


资料来源：招商证券

- **Tips！**
  - 请不要随意修改或移动上图所示文件夹及文件
  - 在转换格式前，请将“输出结果文件夹”清空

使用方法

- 将待转的 PDF 全部放入“待转 PDF 文件夹”，**放入前保证该文件夹为空：**



资料来源：招商证券

使用方法

- 点击“pdfchangetxt.exe”，出现如下界面：

资料来源：招商证券

转换为 txt，请输入“**txt**”，并回车

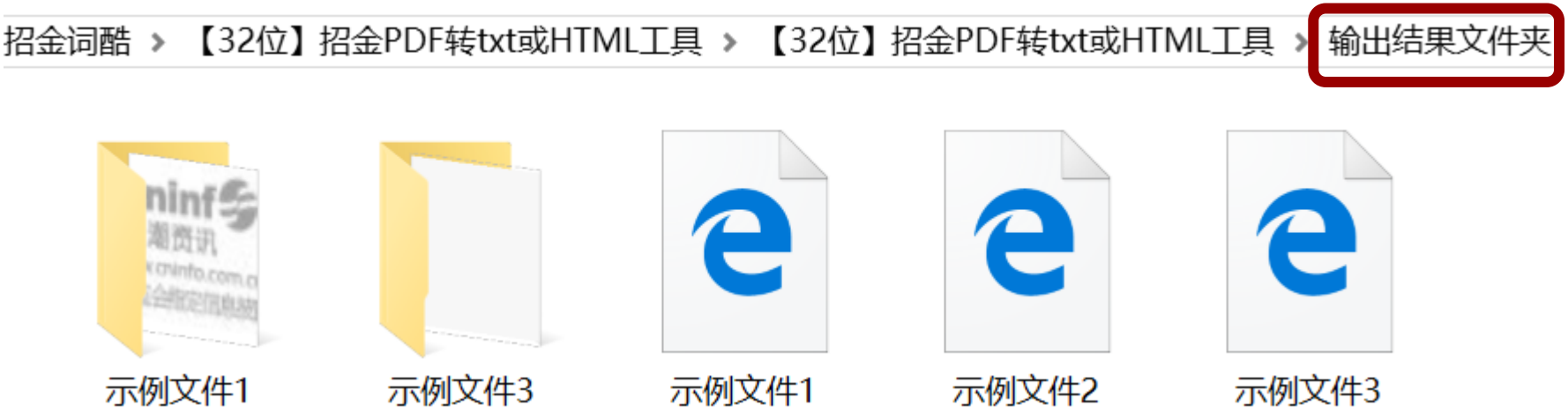
转换为 HTML，请输入“**html**”，并回车

- 程序运行界面：

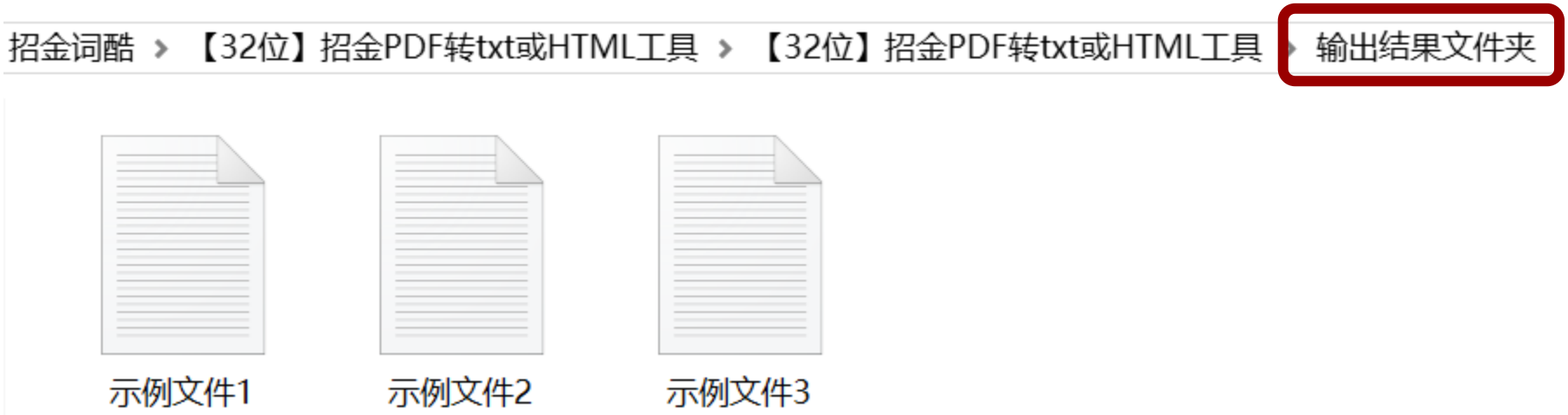
资料来源：招商证券

使用方法

- 程序运行结束，转化成功的文件储存在“输出结果文件夹”



资料来源：招商证券



资料来源：招商证券

分析师承诺

负责本研究报告的每一位证券分析师，在此申明，本报告清晰、准确地反映了分析师本人的研究观点。本人薪酬的任何部分过去不曾与、现在不与，未来也将不会与本报告中的具体推荐或观点直接或间接相关。

叶涛：首席分析师。上海交通大学管理学硕士，2005年起从事金融工程研究，曾先后任职于易方达基金机构投资部、上投摩根基金研究部、申万菱信基金投资管理总部、长江证券研究部、广发证券发展研究中心，2014年3月加盟招商证券研究发展中心。

欧阳廷婷：研究助理。上海交通大学信息工程硕士，2015年5月加盟招商证券研究发展中心。

赵月涓：研究助理。同济大学应用数学硕士，2015年5月加盟招商证券研究发展中心。

重要声明

本报告由招商证券股份有限公司（以下简称“本公司”）编制。本公司具有中国证监会许可的证券投资咨询业务资格。本报告基于合法取得的信息，但本公司对这些信息的准确性和完整性不作任何保证。本报告所包含的分析基于各种假设，不同假设可能导致分析结果出现重大不同。报告中的内容和意见仅供参考，并不构成对所述证券买卖的出价，在任何情况下，本报告中的信息或所表述的意见并不构成对任何人的投资建议。除法律或规则规定必须承担的责任外，本公司及其雇员不对使用本报告及其内容所引发的任何直接或间接损失负任何责任。本公司或关联机构可能会持有报告中所提到的公司所发行的证券头寸并进行交易，还可能为这些公司提供或争取提供投资银行业务服务。客户应当考虑到本公司可能存在可能影响本报告客观性的利益冲突。