数量化投资

2016 年 11 月 1 日

招金词酷

金融文本挖掘的分词工具

招金词酷是在开源分词工具 jieba 的基础上，通过整合搜狗金融词库、Wind 词汇、招商金工总结词汇三部分增量词库，很大程度地提升了金融类文本的分词效果！

- 搭建招金词酷：招金词酷是在开源分词工具 jieba 的基础上，通过整合搜狗金融词库、Wind 词汇、招商金工总结词汇三部分增量词库，通过设置四部分词汇的词频，使得金融词汇可以优先被识别，很大程度地提升了金融类文本的分词效果！
- 招金词酷优势：通过对比招金词酷与其他分词工具，总结出四大优势：

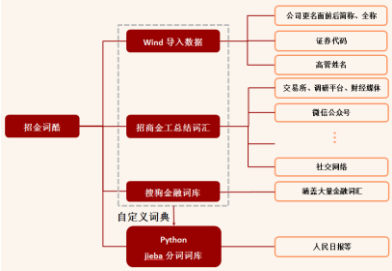
  1.可直接对 txt、pdf 文件分词；

  2.支持批量文件分词；

  3.分词精度高、信息保留度高；

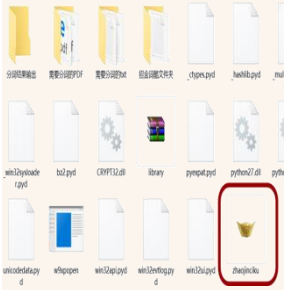
  4.操作便捷，无需安装 python。
- 手把手教您用招金词酷：我们已经把分词工具打包成 exe 文件，无需安装编程软件，一键运行；在本章中我们都用图来一步步演示招金词酷的具体使用方法。
- PDF 转换工具：目前很多金融类文本都是 PDF 格式，我们也为各位投资者额外提供了一款批量 PDF 转换小工具，这款工具支持把 PDF 软件批量转换成 txt 和 HTML 两种格式。我们之所以额外提供 HTML 格式的转换，是为了方便各位投资者对于 PDF 文件中表格信息的抓取。HTML 文件由于有很多标签，可以实现对表格中内容的快速定位，对于精确抓取表格中信息有很好的作用。

招金词酷组成部分



资料来源：天软科技、招商证券

招金词酷主体文件



资料来源：天软科技、招商证券

叶涛
021-68407343
yetao@cmschina.com.cn
S1090514040002

研究助理
赵月涓
zhaoyuejuan@cmschina.com.cn

正文目录

图表目录

一、搭建招金词酷

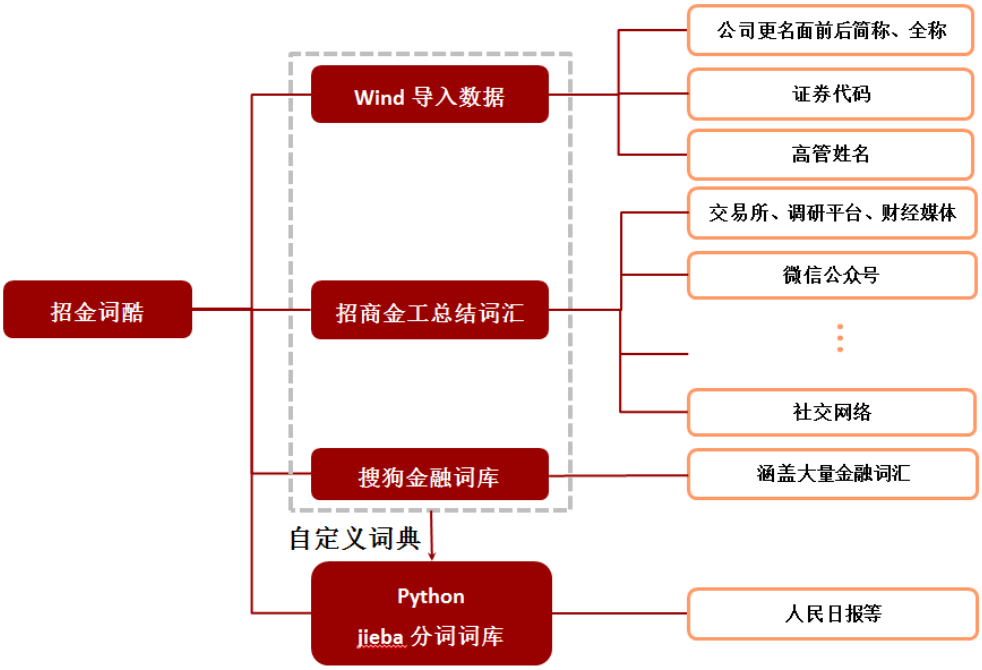
文本挖掘是事件驱动方向的一个重要分支，用于金融分析的文本源一般来自交易所、调研信息平台、财经媒体、微信公众号、社交网络（如雪球、股吧.....）等平台。

文本挖掘的方法有很多，但不论是文本分类、信息提取还是情感分析，任何一种分析方法，想要做到精细，都离不开分词，而分词的基础是需要有个适合且全面的词库。现在中文分词工具很多，但大都针对普通文本。金融类文本中往往会有诸多金融特有词汇出现，现有分词工具无法正确识别，导致现有分词工具对于金融文本的分词效果并不理想。

为了弥补这一块的空缺，招商金工团队在 Pyhon 的 jieba 分词工具的基础上，开发了一款为金融类文本专门打造的分词工具—招金词酷。

jieba 分词工具本身自带一个较为强大的词库，我们再 jieba 词典的基础上添加了 Wind 数据、搜狗金融词库以及我们人工自行总结的词汇，通过自定义词典的方式加载到 jieba 中，大大提升分词精度。

图 1 招金词酷组成部分

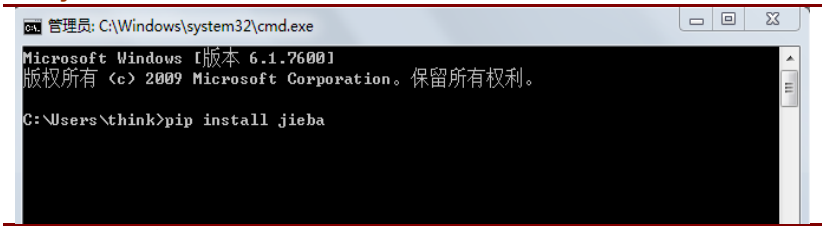


资料来源：Wind 资讯、招商证券

1.1 jieba 分词安装及基本命令

在 Python 中安装 pip 包，即可很方便快捷地利用 pip 包来加载其他的包。因此，若要安装 jieba 分词模块，只需要在命令提示符（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。

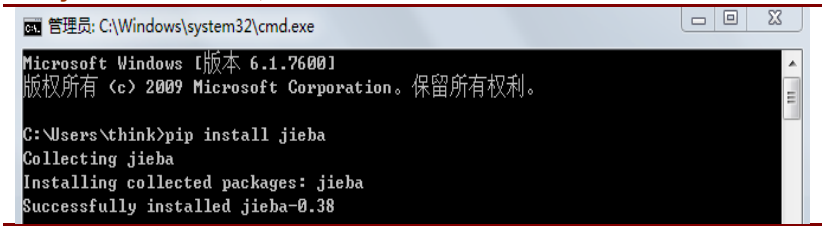
图 2 jieba 安装



资料来源：招商证券

安装成功后，系统出现下图提示，即安装成功。

图 3 jieba 安装成功界面



资料来源：招商证券

jieba 的基本命令是 jieba.cut()，三个参数分别为：

(1) 需要分词的字符串；

(2) cut_all：控制采用“全模式”还是“精确模式”；

下面我们举个例子，来说明这两个模式的区别：

全模式：

seg_list = jieba.cut("我来到北京清华大学", cut_all=True)

print(“Full Mode: ” + “ / ”.join(seg_list))

精确模式：

seg_list = jieba.cut(“我来到北京清华大学”, cut_all=False)

print(“Default Mode: ” + “ / ”.join(seg_list)) ）

这两种模式下相应的输出结果为：

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

从输出结果来看，虽然全模式有可以分拆出更多词的可能性，但是由于一些过度拆分，也会添加较多杂音；反观精确模式下，可以较为精准地把语句拆分开，并且不会出现过度拆分的现象。因此，招金词酷为了确保分词的精确程度，采用了 jieba 中的精确模式。

(3) HMM：控制是否使用 HMM 模型（隐马尔科夫模型）用于识别新词；

接下来我们举个例子，来说明是否启用 HMM 模式在分词效果上的区别：

默认新词识别模式：

```
seg_list = jieba.cut(“他来到了网易杭研大厦”)
print(“ / ".join(seg_list))
```

不使用新词识别模式:

```
seg_list = jieba.cut(“他来到了网易杭研大厦”,HMM = False)
print(“ / ".join(seg_list))
```

这两种模式下相应的输出结果为:

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

从输出结果也可以看出，如果使用新词识别模式，对于一个 jieba 词库中没有的词语，他会根据 HMM 模型来进行识别；如果未使用新词识别模式，遇到新词时 jieba 只会拆成自己词库中有的词。

1.2jieba 自定义词典加载方法

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。具体的导入命令为: jieba.load_userdict(file_name)，其中，file_name 为自定义词典的路径，注意在这里路径尽量使用英文，中文路径有可能会出现编码问题，使用中文路径需要先转换成 Unicode 码。

自定义词典的格式与 jieba 自带的词典文件 dict.txt 格式相同，一个词占一行。每一行分词汇、词频（可省略）、词性（可省略），用空格隔开，顺序不可颠倒。

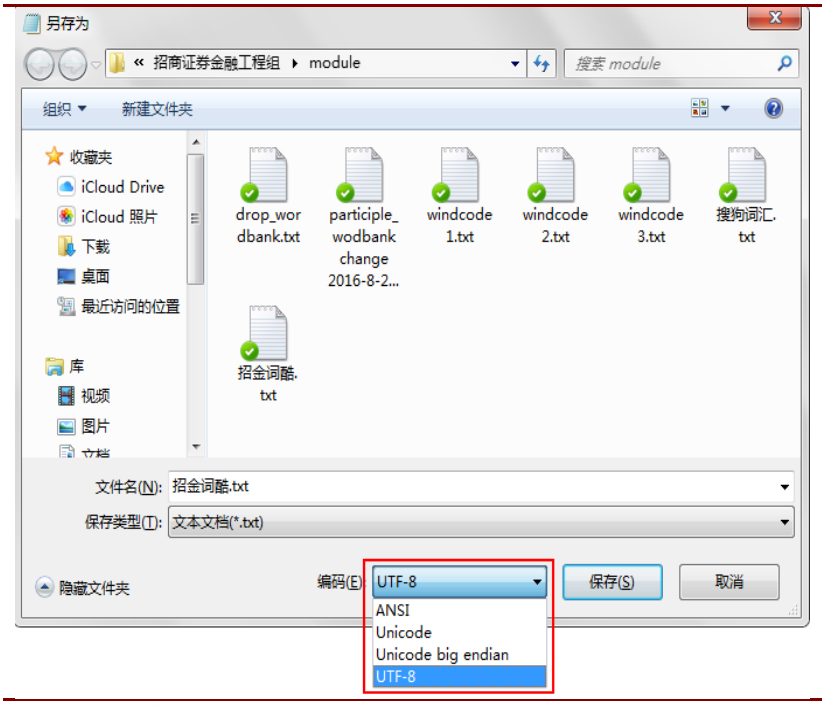
表 1: jieba 自定义词典格式

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

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 分词模块信息可以参见 https://github.com/fxsjy/jieba。

图 4 自定义词典保存方式



资料来源：招商证券

1.3 招金词酷组成部分

(1) jieba 词库

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。

表 2：jieba 词库示例

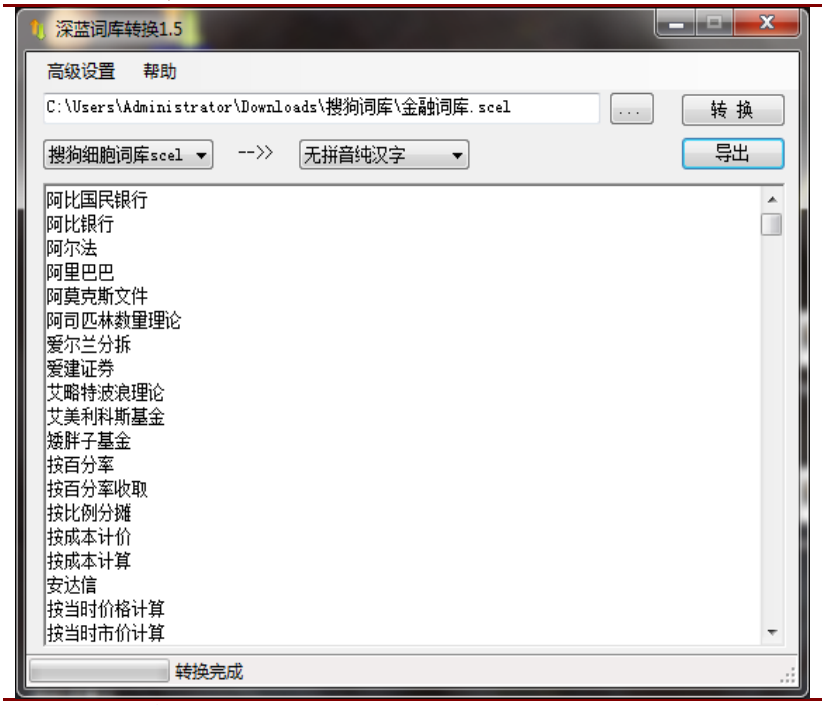
词汇	词频	词性
已达成	3	nrt
已过期	3	d
已近尾声	3	n
已远	3	d
已逝	3	v
已逸待劳	3	nz
已销	3	v
木牌	374	n
木牛流马	40	ns

资料来源：jieba、招商证券，截止 2016 年 9 月 20 日

(2) 搜狗金融词库

搜狗官网提供了很多种类的词库，我们挑选了金融词库，下载出的金融词库是特有的 scel 格式文件，下载网址为 http://pinyin.sogou.com/dict/detail/index/334。通过“深蓝词库转换工具”可将 scel 格式文件转换成 txt 文件。

图 5 深蓝词库转换工具



资料来源：招商证券

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。

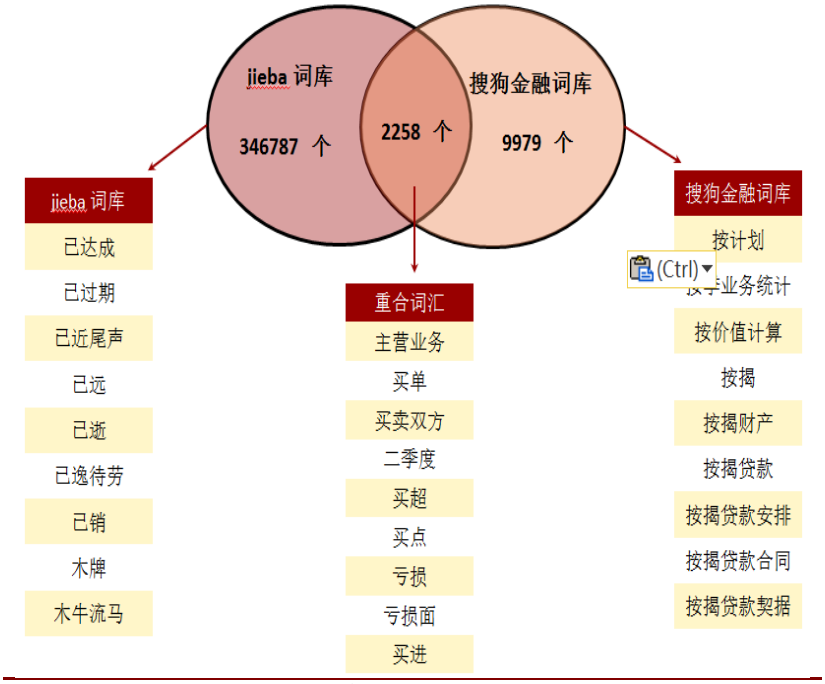
表 3：搜狗金融词库示例

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

我们也做了 jieba 词库和搜狗金融词库中所有词语的比较，搜狗金融词库中的词语净增量为 9979 个。

图 6 jieba 词库与搜狗金融词库的关系



资料来源：jieba、搜狗科技、招商证券

(3) 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", "")
```

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

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

表 4：Wind 词汇示例

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

(4) 招商金工总结词汇

这部分词汇是由我们人工总结加入的，主要包括从公告、财经媒体、微信公众号、社交网络等文章中实践经验总结出来的一些特词汇。

1.4 词频设置

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，应该适当提高新收录词汇的词频。

通过对 jieba 词典的词频统计，我们发现超过 70% 的词汇词频都集中在 10 以内。因此我们定义词频大于 10 的词汇为“招金词酷专属词”，属于较为重要的词汇；其余均为“普通词”。

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，由于 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。

图 7 招金词酷词频设置

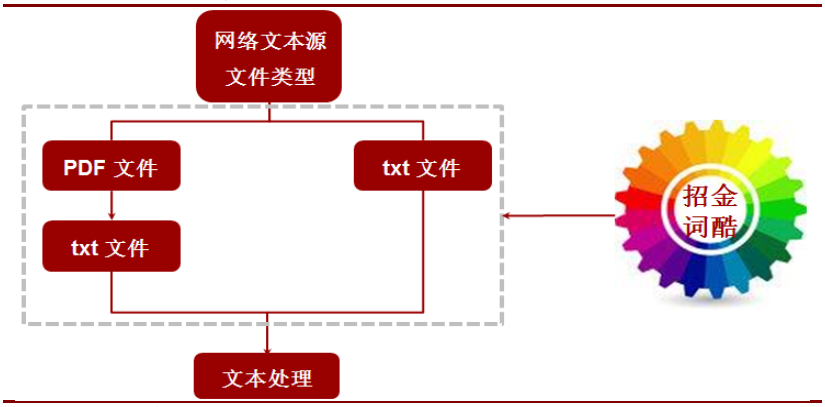
词库		词频
搜狗金融词库		11
Wind 词库词频	公司代码“000000”	15
	公司代码“000000.SH/SZ”	20
	公司简称	20
	公司全称	40
	高管姓名	40
招商金工总结词汇		200

资料来源：招商证券

二、招金词酷赢在精度

招金词酷可以支持的文件类型包括 PDF 文件和 txt 文件。对于 PDF 文件分词，内在的工作机理也是先把 PDF 转换成 txt 文件，再进行分词。

图 8 招金词酷工作原理



资料来源：招商证券

2.1 与 IK Analyzer 分词比较

接下来对于 txt 分词，我们将招金词酷与 IK Analyzer 进行分词精度的比较。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。我们从 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。

图 9 公告原文

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。

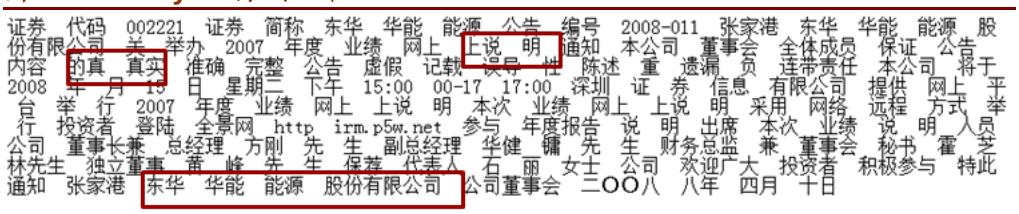
张家港东华能源股份有限公司董事会

二〇〇八年四月十日

资料来源：巨潮资讯、招商证券

下面为 IK Analyzer 对上述文章的分词结果：

图 10 IK Analyzer 分词效果



资料来源：招商证券

下面为招金词酷的分词结果：

图 11 招金词酷分词效果

资料来源：招商证券

从分词结果的对比可以看出，IK Analyzer 不仅不支持直接对 PDF 分词，并且在分词过程中存在过度拆分和不当拆分的问题。反观招金词酷的结果，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。

2.2 与某收费工具分词比较

2.2.1 公告爬取与抽样规则

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：

图 12 抽样规则

资料来源：招商证券

2.2.2 随机抽样公告

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。

表 5：抽样所得公告标题

公告	公告	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
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
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
26	20130509	600730	华仪电气：关于 2012 年现场业绩说明会暨投资者接待
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
34	20140521	601168	广誉远：关于举行 2013 年度报告网上业绩说明会的公
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
44	20150410	300010	立思辰：关于交易对手方对置入资产 2014 年度业绩承
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

公告	公告	公司代码	公告名称
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

资料来源：招商证券

2.2.3 得词率比较

我们定义了一个表征分词信息保留度的指标得词率如下：

得词率分子：分别使用某收费工具、招金词酷进行分词，所得分词 txt 文件中所有词语长度加总；

得词率分母：PDF 公告转化为 txt 文件，除去空格、换行符、标点的字符串总数；

得词率：得词率分子/得词率坟墓。

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。

图 13 PDF 转化失败示例

资料来源：招商证券

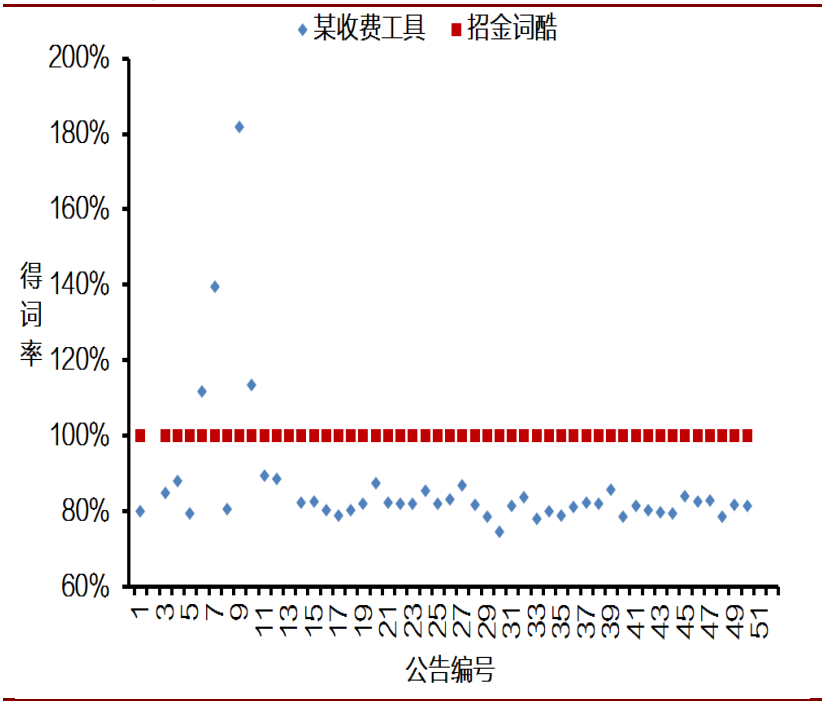
在 PDF 转化成 txt 文件成功的 49 篇 PDF 业绩类公告中，招金词酷全部分词成功，该收费工具有 1 篇分词失败。因此在之后得词率的计算比较中，该收费工具仅有 48 个点，招金词酷则有 49 个点。

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：

(1) 得词率普遍小于 1，说明文件类型转化时词语损失量较大；

(2) 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。

图 14 得词率



资料来源：招商证券

2.2.4 分词精度比较

我们分别选取了使用收费工具与招金词酷对一篇公告的分词效果。下图位该款收费工具的分词效果。

图 15 收费工具分词效果

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

我们可以发现，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。招金词酷的分词结果如下：

图 16 招金词酷分词效果

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

从上面分词效果的对比中，我们也可以看出招金词酷的优势凸显：

(1) 支持批量导入分词功能；

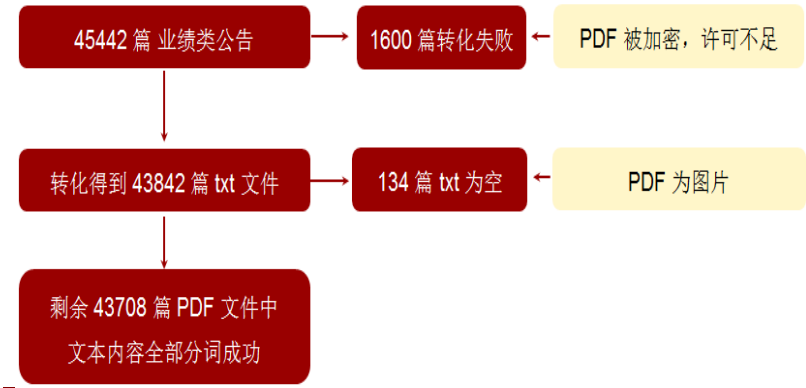
(2) 信息保留度高；

(3) 分词精度高，处理金融词汇具有明显优势。

2.3 压力测试

我们对 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%。

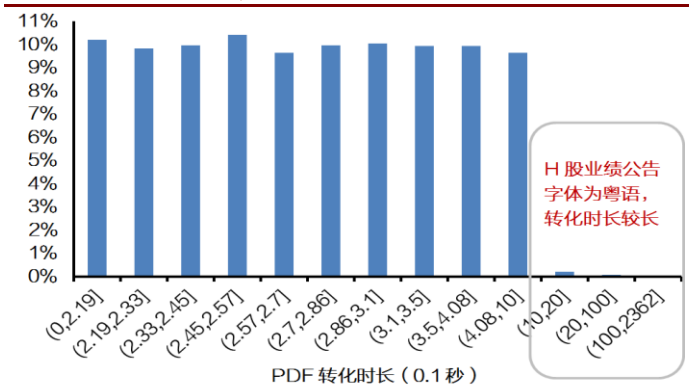
图 17 招金词酷文件转化及分词结果



资料来源：招商证券

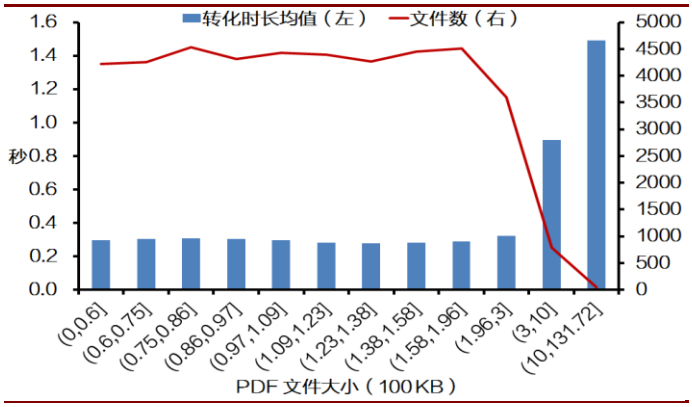
对于 PDF 文件分词，涉及到文件转化与分词两个过程，我们接下来也测试了文件转化与分词的时长分布。99% 的公告转化时长不超过 1 秒，分词时长不超过 0.1 秒。

图 18 转化时长分布



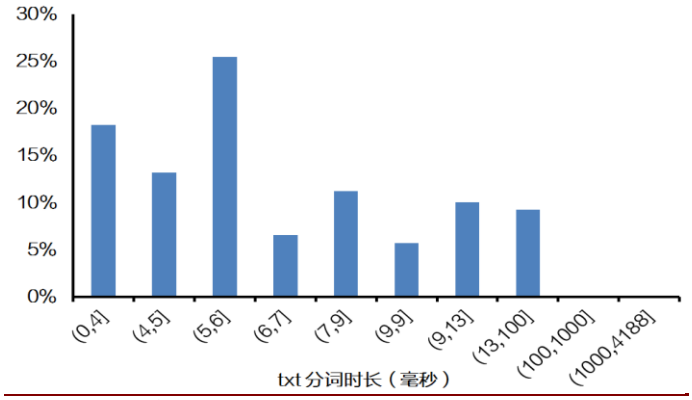
资料来源：招商证券

图 19 转化时长与文件大小



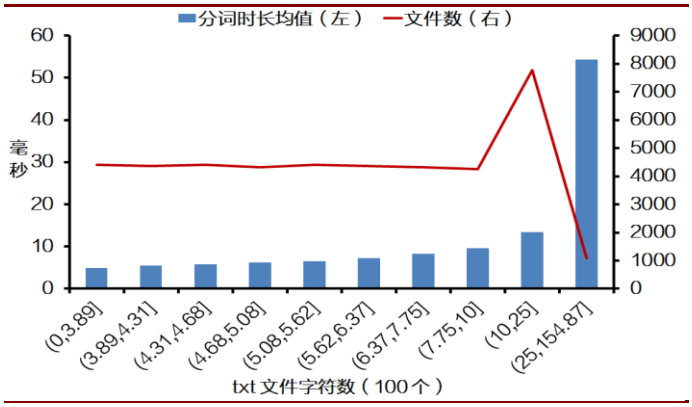
资料来源：招商证券

图 20 分词时长分布



资料来源：招商证券

图 21 分词时长与文件字符数



资料来源：招商证券

从这个结果可以看出，招金词酷对于 PDF 文件分词的成功率较高，并且也很稳健。

2.4 招金词酷维护方式

本次我们推出的招金词酷是第一个版本，之后我们也会不断地来更新、维护招金词酷。

(1) 增加招金词酷录入的词汇量，使得词库更为全面，适应更多文本；

(2) 完善词库中词汇的词性，满足信息提取、情感分析等文本挖掘工作的需要；

(3) 定期淘汰过时词汇，降低其词频，保持词库新鲜度

三、手把手教您用招金词酷

招金词酷的下载地址为 http://pan.baidu.com/s/1boKzTdt，密码为 35vf。我们分别为 32 位系统、64 位系统提供了对应的招金分词工具。不需要安装 Python 即可使用。

首先下载 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。

图 22 招金词酷文件明细

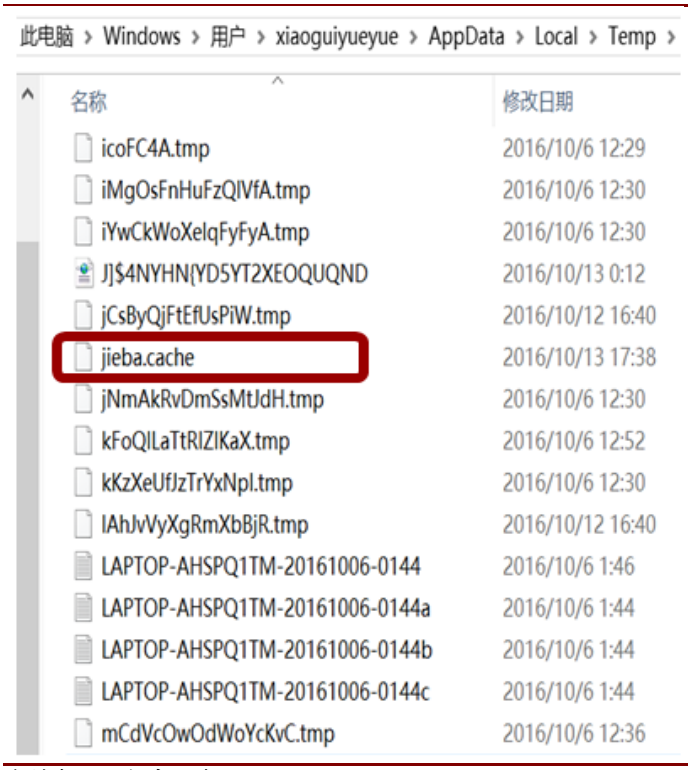
资料来源：招商证券

接下来配置招金词酷的环境，打开“文件夹选项”，暂时将文件的隐藏属性去掉，将“jieba.cache”拷贝到路径“C:\Users\~\AppData\Local\Temp”后，恢复文件夹选项设置。

图 23 修改文件夹选项

资料来源：招商证券

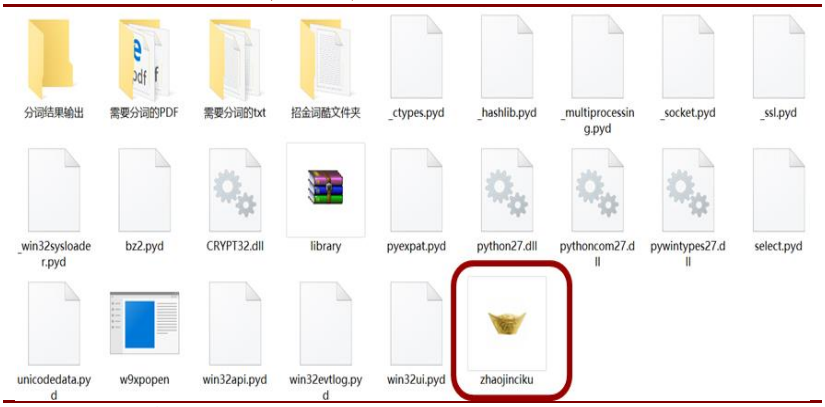
图 24 拷贝文件路径



资料来源：招商证券

目前我们所提供的招金词酷仅支持 Windows 系统。下面我们手把手教您使用我们的招金词酷。由于“【32 位】招金分词工具”与“【64 位】招金分词工具”使用方法类似，接下来重点介绍前者，后者也是类似操作。下图为解压后的界面，请注意请不要随意修改或移动上图所示文件夹及文件。

图 25 招金词酷工具解压后界面

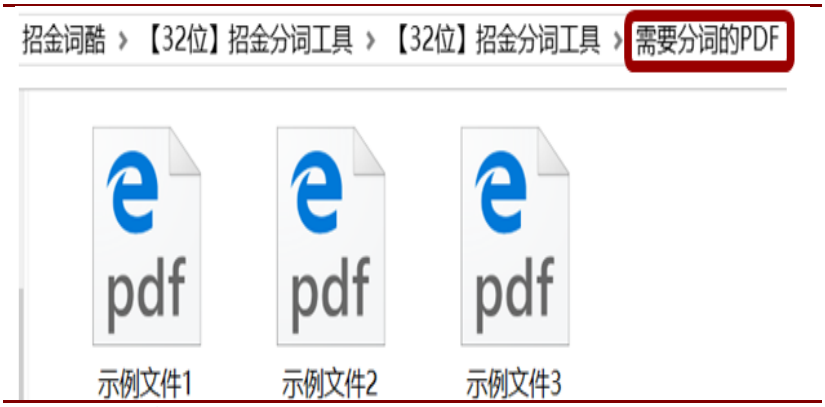


资料来源：招商证券

(1) 分词前放入需要分词的文件

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”文件夹中。

图 26 需要分词的 PDF



资料来源：招商证券

若对 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”文件夹中。

图 27 需要分词的 txt

资料来源：招商证券

这里需要注意，在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。

(2) 运行分词软件

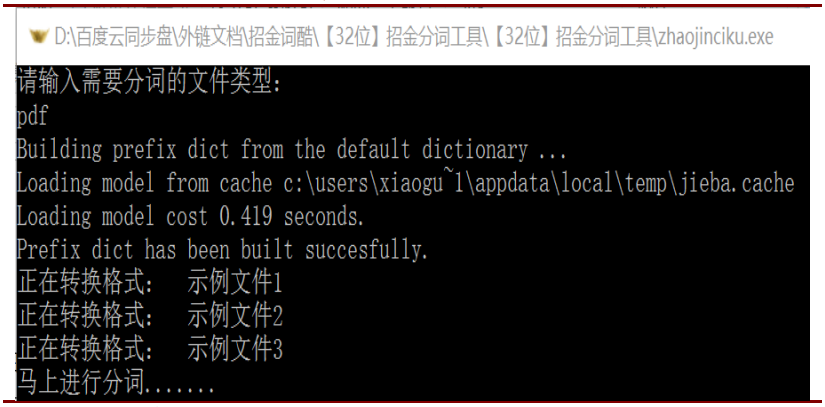
点击 “zhaojinciku.exe”，出现如下界面：

图 28 输入需要分词的文件类型

资料来源：招商证券

若对 PDF 文件分词，请输入 “pdf” 并回车；若对 txt 文件分词，请输入 “txt” 并回车。程序运行界面如下图：

图 29 招金词酷程序运行界面

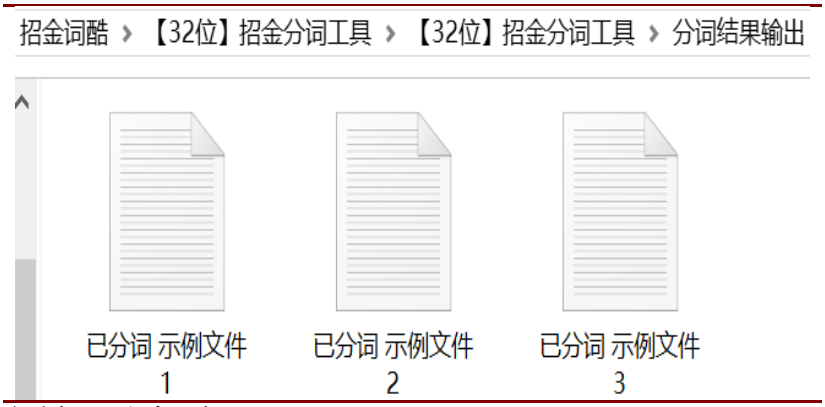


资料来源：招商证券

(3) 分词结果输出

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。储存分词结果的 txt 文件命名方式为 “已分词” +文件名

图 30 分词结果输出

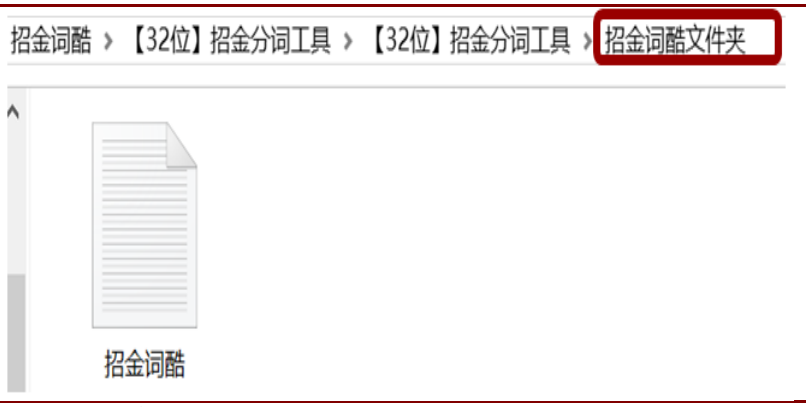


资料来源：招商证券

(4) 招金词酷

我们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。

图 31 招金词酷文件夹



资料来源：招商证券

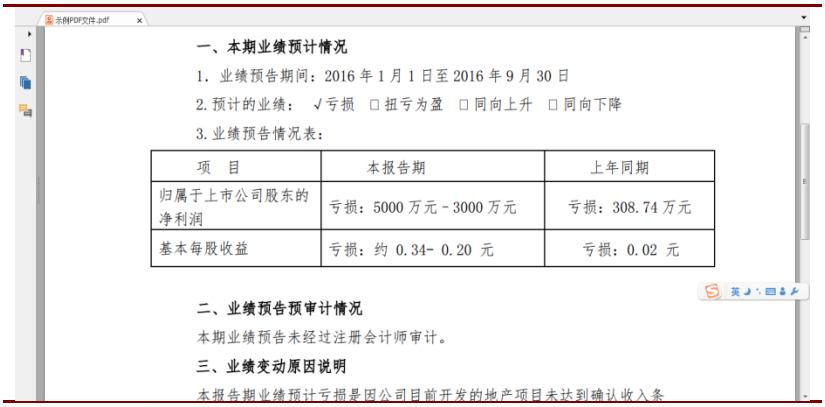
四、PDF 批量转 txt、HTML 工具

目前很多金融类文本都是 PDF 格式，我们也为各位投资者额外提供了一款批量 PDF 转换小工具，这款工具支持把 PDF 软件批量转换成 txt 和 HTML 两种格式。我们之所以额外提供 HTML 格式的转换，是为了方便各位投资者对于 PDF 文件中表格信息的抓取。

4.1 格式转换示例

接下来我们选取一篇业绩类 PDF 公告来做示例。

图 32 格式转换 PDF 公告示例



一、本期业绩预计情况

1. 业绩预告期间：2016 年 1 月 1 日至 2016 年 9 月 30 日

2. 预计的业绩： √亏损 □扭亏为盈 □同向上升 □同向下降

3. 业绩预告情况表：

项 目	本报告期	上年同期
归属于上市公司股东的净利润	亏损：5000 万元－3000 万元	亏损：308.74 万元
基本每股收益	亏损：约 0.34－0.20 元	亏损：0.02 元

二、业绩预告预审计情况

本期业绩预告未经过注册会计师审计。

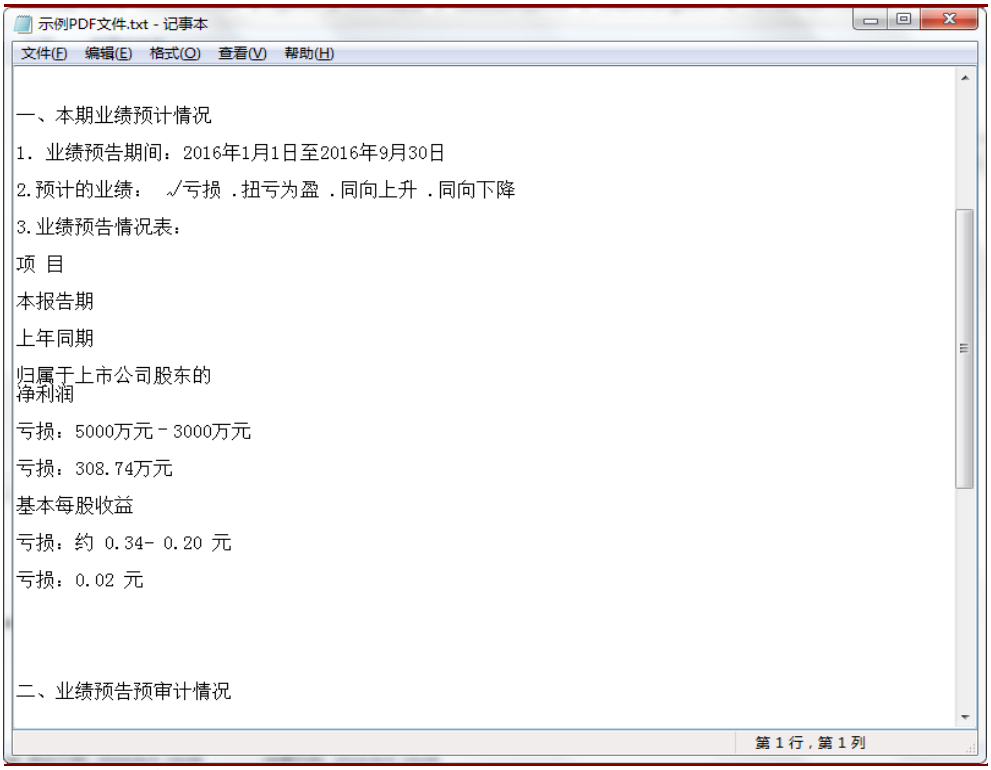
三、业绩变动原因说明

本报告期业绩预计亏损是因公司目前开发的地产项目未达到确认收入条

资料来源：招商证券

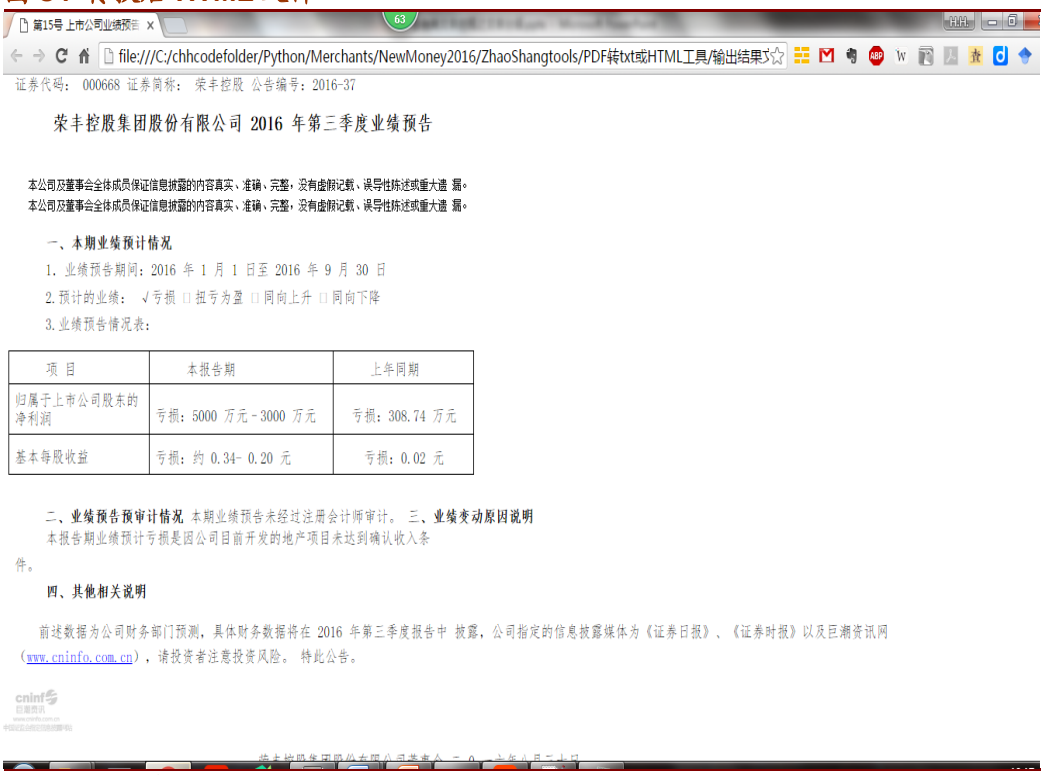
我们将这篇 PDF 公告分别转换成 txt 格式和 HTML 格式。

图 33 转换后的 txt 文件



资料来源：招商证券

图 34 转换后 HTML 文件



资料来源：招商证券

从上图可以看出，HTML 格式完美还原了原文中的表格。

4.2 与 PDF2TXT 比较

目前市场上还有一款收费的 PDF 转换工具—PDF2TXT 软件。以下为这款软件的收费标准。

图 35 PDF2TXT

资料来源：招商证券

接下来我们将我们提供的 PDF 转换工具与 PDF2TXT 进行比较。

表 6：PDF 批量转 txt、HTML 工具与 PDF2TXT 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

我们的 PDF 转换工具最大的优势是免费，且提供了转换成 HTML 格式的功能。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。

图 36 HTML 文件定位表格

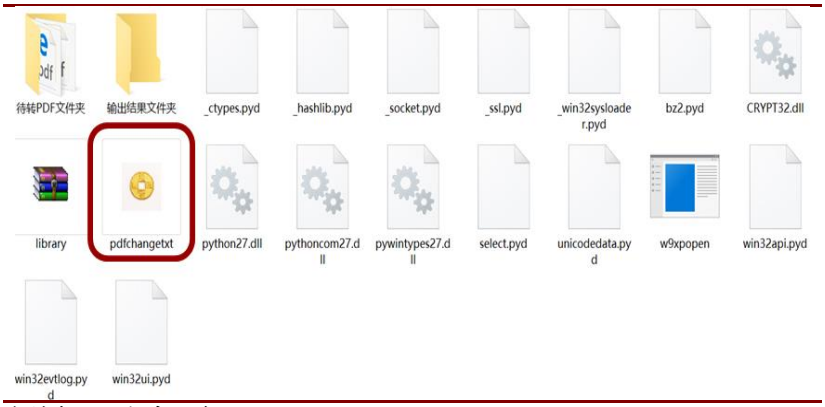
资料来源：招商证券

4.3 使用方法

安装 Adobe Acrobat DC，具体方法请见第三章。解压“【32 位】招金 PDF 转 txt 或 HTML 工具”。

在使用过程中要注意，请不要随意修改或移动上图所示文件夹及文件。

图 37 PDF 转换工具解压后界面

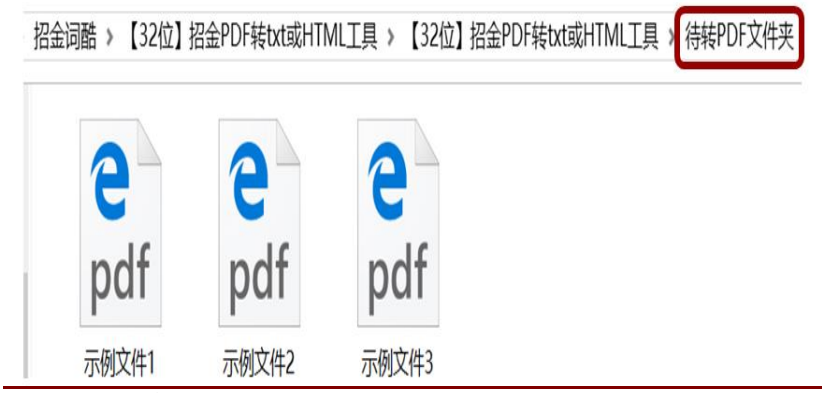


资料来源：招商证券

(1) 放入需要转换格式的 PDF 文件

将待转的 PDF 全部放入“待转 PDF 文件夹”，放入前保证该文件夹无其他类型的文件。

图 38 待转 PDF 文件夹



资料来源：招商证券

(2) 运行 PDF 转换工具

点击“pdfchangetxt.exe”，出现如下界面，若要将 PDF 文件转换为 txt，请输入“txt”，并回车；若要将 PDF 文件转换为 HTML，请输入“html”，并回车。

图 39 输入需要将 PDF 文件转换成的类型

资料来源：招商证券

下图为程序运行界面：

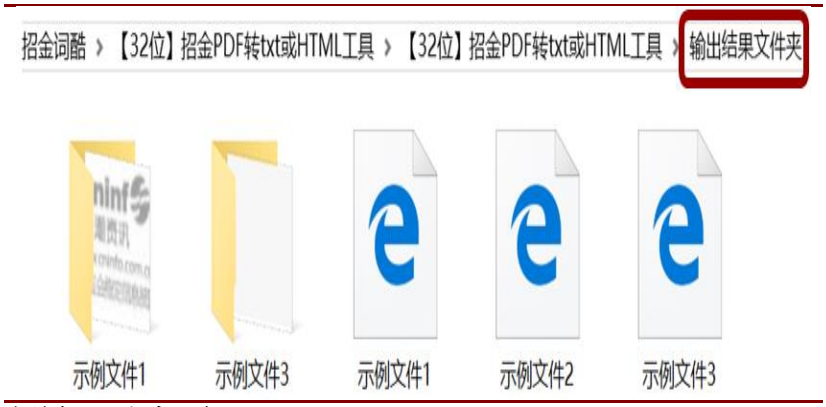
图 40 PDF 批量转换 txt、HTML 工具程序运行界面

资料来源：招商证券

(3) 转换结果输出

程序运行结束，转化成功的文件储存在“输出结果文件夹”。

图 41 将 PDF 转换为 HTML 格式的输出结果



资料来源：招商证券

图 42 将 PDF 转换为 txt 格式的输出结果

资料来源：招商证券

风险提示：

本文中所引入的假设以及基于假设所构建的模型，均是对所要研究问题的主要矛盾以及矛盾主要方面的一种抽象，因此模型以及基于模型所得出的相关结论并不能完全准确的刻画现实环境与预测未来。

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。

投资评级定义

公司短期评级

以报告日起 6 个月内，公司股价相对同期市场基准（沪深 300 指数）的表现为标准：

强烈推荐：公司股价涨幅超基准指数 20%以上
审慎推荐：公司股价涨幅超基准指数 5-20%之间
中性：公司股价变动幅度相对基准指数介于±5%之间
回避：公司股价表现弱于基准指数 5%以上

公司长期评级

A：公司长期竞争力高于行业平均水平
B：公司长期竞争力与行业平均水平一致
C：公司长期竞争力低于行业平均水平

行业投资评级

以报告日起 6 个月内，行业指数相对于同期市场基准（沪深 300 指数）的表现为标准：

推荐：行业基本面向好，行业指数将跑赢基准指数
中性：行业基本面稳定，行业指数跟随基准指数
回避：行业基本面向淡，行业指数将跑输基准指数

重要声明