WinGo Data 20190630

文构数据

WinGo 财经文本数据平台

---产品简介---

武汉文铸数据科技有限公司

湖北省武汉市东湖技术开发区高新大道 788 号沃德中心

电话：027-87419769　邮箱：sales@wingodata.com　网址：www.wingodata.cn

第一章 WinGo 产品简介

WinGo 财经文本数据平台（中文名为“文构财经文本数据平台”）是**中国首家基于中美上市公司披露文本的人工智能财经数据平台**。平台从学术研究和业界量化投资需求出发，聚焦于中美海量财经文本数据。针对两国截然不同的文本披露规则和财经文本特点，平台应用自然语言处理、深度学习和人工智能技术对财经文本进行深度加工，给用户提供财经文本的词频、相似词、文本特征等全新深度处理的数据，从而为学术研究、投资决策应用等提供多方位支持。

WinGo 数据平台包括**中国上市公司**和**美国上市公司**两大数据库，由业内专家和高校知名学者主持设计，打破了财经文本分析的技术壁垒，大幅降低研究成本，为广大研究和分析人员开辟出全新的研究模式。

1. WinGo 中国上市公司数据库内容

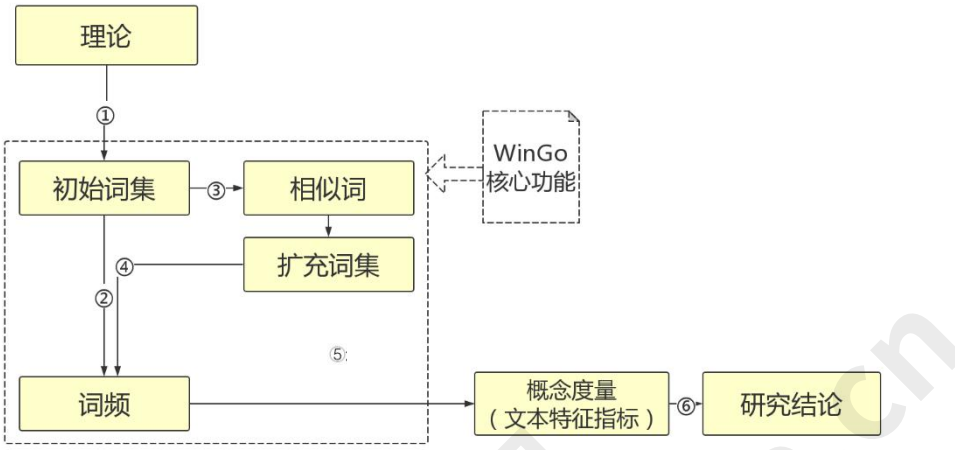
WinGo 平台中国上市公司数据库由词频、相似词、文本特征指标、自定义特征、会计金融指标以及在线服务六大模块组成。数据源涵盖范围广阔，囊括了上市公司披露的年度报告、季度报告、董事会报告、管理层讨论与分析（MD&A）、审计报告、财务报表附注、IPO 招股说明书、内部控制评价报告、业绩说明会、社会责任报告等。数据区间最早可追溯到 2001 年，共收录文档二十余万份，文字逾百亿。

图 1 WinGo 中国上市公司数据库内容

网址: www.wingodata.cn 地址：武汉市东湖新技术开发区高新大道 788 号沃德中心
电话: 027-87419769（武汉） 18612801983（北京）
邮箱: sales@wingodata.com（武汉） liufeng@wingodata.com（北京）

1.1 词频模块

词频指某个词汇或某类词汇在文本中出现的频率。作为文本分析的基石，词频可以有效帮助研究人员实现各类特征指标的构建，具体应用过程如下图所示：



注：①⑤⑥步骤为用户操作；②④步骤使用 WinGo 词频功能；③步骤使用 WinGo 相似词功能

图 2 WinGo 中文词频应用流程图

首先，研究人员根据理论或文献确定度量某个经济管理概念的初始关键词词集；然后，通过 WinGo 词频数据库获取目标词集在多种财经文本语料中的词频；接下来，便可基于词频信息进行相关概念的测度，并可以进一步构建自己独特的文本特征指标，从而得到新的因子用以更深层次的研究（即图 2 的①②⑤⑥步骤）。

目前，基于文本词频的概念测度是经济管理研究的学术前沿。例如，Loughran and McDonald（2011）通过计算财经专用积极消极词汇的词频比率构建适用于年报的语调测度方法，并研究了语调和股票收益率、交易量以及股票波动率等的关系。姜付秀等（2015）通过计算“诚信”等关键词在年报、内部控制评价报告等文本中出现情况构建了企业诚信文化指标，发现以诚信作为文化的企业盈余管理水平更低。王雄元等（2017）通过计算“风险”、“不确定性”等词汇的词频比例，构建了企业风险指标，研究得出企业风险披露水平与分析师预测精确度正相关。

1.2 相似词模块

构建特定的文本指标时我们一般需要用到语义相似的多个词汇，在现有的学术研究中，扩充词集的方法主要有两种：第一是通过同近义词词典人工查找来对词集进行扩充，第二是通过人工阅读所要研究的语料来扩充词集。然而，人工查找的方式往往会忽略文本语境，而且存在较强的主观性偏差，因而不能全面、准确、客观地衡量文本特征。

在此情况下，WinGo 平台推出了“深度学习相似词”数据库，采用 Word Embedding（词向量）模型对海量财经文本语料进行训练，构建词汇相似度计算模型，成功提取基于财经语

网址: www.wingodata.cn 地址：武汉市东湖新技术开发区高新大道 788 号沃德中心
电话: 027-87419769（武汉） 18612801983（北京）
邮箱: sales@wingodata.com（武汉） liufeng@wingodata.com（北京）

料的语义相似词集。这种方法打破了传统的技术壁垒，克服了现有方法的缺陷，大幅降低了研究成本。因此，在确定好初始词集后，研究人员可使用 WinGo 相似词产品（深度学习相似词）进行词集扩充（即图 2 的③④步骤）。

1.3 文本特征模块

为了提升研究效率，降低研究成本，WinGo 平台还推出了专业团队构建的文本特征指标，包括“与内容无关的特征”和“与内容有关的特征”两部分内容。其中，“与内容无关的特征”指的是与文本内容不相关的一类文本特征，包括文本相似性、语调、可读性等特征。“与内容有关的特征”指的是与文本内容相关的一类文本特征，包括创新、诚信、风险以及前瞻性等特征。每一个文本特征，都设有财务报告、审计报告、财务报表附注、IPO 招股说明书、内部控制评价报告、业绩说明会、社会责任报告等数据文本，旨在为广大研究人员提供简便高效的上市公司披露文本分析平台。

学者可基于文本特征数据库开展的热点研究包括但不限于以下方向：

- 文本语调与公司业务特征、高管信息披露动机
- 可读性、文本相似性与公司绩效、投资者反应
- 风险、创新、诚信与公司绩效、市场反应

1.4 自定义特征模块

自定义特征数据库集成了 WinGo 词频数据库与 WinGo 深度学习相似词数据库中的两大基础功能，旨在为用户提供便捷、高效的与内容有关文本指标的构建系统。具体来讲，自定义特征指标的构建逻辑分为三步：首先，用户根据已有的研究理论，定义所构建指标的原始词集（又称种子词集）；其次，用户使用系统集成的 WinGo 深度学习相似词推荐功能对种子词集进行相似词扩充；最后，系统自动计算自定义指标词集中每个词的词频，加总归一化得到最终文本指标。

根据以上自定义指标的构建逻辑，自定义数据库分别提供“特征词典定制”以及“特征计算”两大功能。针对特征词典定制功能，用户可以创建全新的与内容有关的文本指标，定义并修改指标对应的种子词集。此外，针对种子词集中的每个词汇，用户可以调用 WinGo 深度学习相似词推荐功能，查找对应相似词推荐结果，系统支持相似词迭代查找，能够最大程度满足用户对词集扩充的需求。针对特征计算功能，用户可以选择系统任意数据源作为指标计算的载体，进行简单的股票代码选择以及起始日期选择后，系统即可自动计算形成最终文本指标供用户下载使用。

1.5 会计金融指标模块

会计金融指标数据库从理论界与实务界关注的热点话题出发，提供会计与金融学科研究领域中被广泛使用的经典指标，每个指标有多种计算方法供用户选择，旨在帮助研究人员提升研究效率，促进不同学科研究领域的融合与发展。

网址: www.wingodata.cn 地址：武汉市东湖新技术开发区高新大道 788 号沃德中心
电话: 027-87419769（武汉） 18612801983（北京）
邮箱: sales@wingodata.com（武汉） liufeng@wingodata.com（北京）

会计金融指标数据库的指标构建过程如下：首先由 WinGo 专业研究团队通过阅读大量经典文献，总结与梳理各类指标的计算方法及使用场景；然后，研究团队成员通过程序独立计算指标，实现指标计算代码与计算结果的交叉验证；最后，将计算结果与权威文献比对核验，保证指标计算的可靠性。

目前会计金融指标模块下设有四类子数据库，分别是盈余质量系列数据库、分析师系列数据库、股票市场系列数据库以及中美上市公司估值对比系统。其中，盈余质量系列数据库包括应计盈余管理、真实盈余管理、会计稳健性和会计信息可比性等指标。分析师系列数据库包括分析师跟踪数量、分析师预测误差和分析师预测分歧度等指标。股票市场系列数据库包括事件研究、股价同步性和股价崩盘风险等指标。中美上市公司估值对比系统，将提供中国上市公司所在细分类行业中对应的美国上市公司估值指标数据（如市盈率、市净率、市销率、总市值等）以及指标对应的时间趋势变化图表，旨在为投资者或分析师提供可比的美国上市公司估值数据，提高估值准确度与投资决策效率。

此外 WinGo 平台还推出了在线服务：

在线服务可为用户提供基于文本的个性化服务，包括语料服务和模型服务两大部分。其中语料服务涵盖中文在线分词、停用词管理、词频统计分析、PDF 解析等。模型服务涵盖文本相似性、LDA 主题模型、STM 主题模型、Word2Vec 模型和 Doc2Vec 模型。

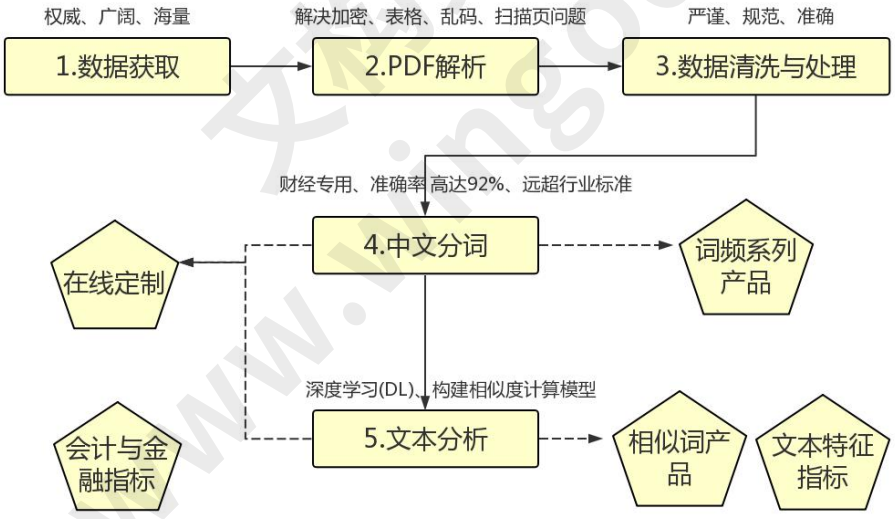


图 3 WinGo 数据平台业务流程图

2. WinGo 美国上市公司数据库内容

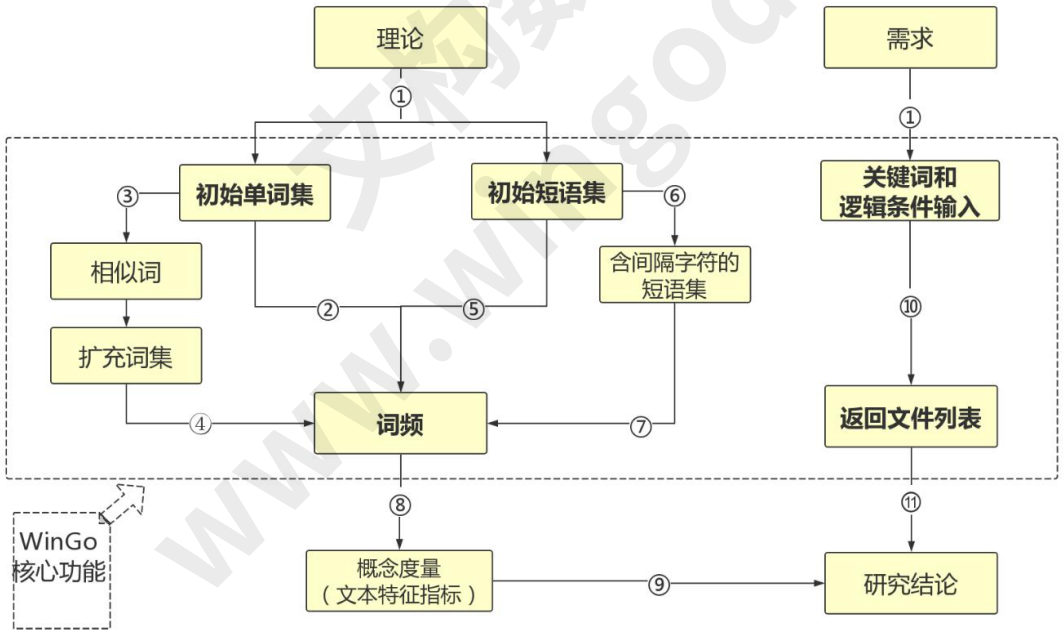
WinGo 平台美国上市公司数据库由词频、相似词、文本特征指标、自定义特征四大模块组成。数据源涵盖范围广阔，囊括了上市公司披露的年度报告、季度报告、管理层讨论与分析（MD&A）、风险章节（Risk factors）、IPO 招股说明书等。数据区间最早可追溯到 1993 年，共收录文档逾百万，文字逾百亿。

网址: www.wingodata.cn 地址：武汉市东湖新技术开发区高新大道 788 号沃德中心
电话: 027-87419769（武汉） 18612801983（北京）
邮箱: sales@wingodata.com（武汉） liufeng@wingodata.com（北京）

图 4 WinGo 美国上市公司数据库内容

2.1 词频模块

词频指某个词汇或某类词汇在文本中出现的频率。作为文本分析的基石，词频可以有效帮助研究人员实现各类特征指标的构建，具体应用过程如下图所示：



注：

①⑧⑨⑪步骤为用户操作；②④步骤使用 WinGo 单词词频功能；③步骤使用 WinGo 相似词功能；⑤⑥⑦步骤使用 WinGo 短语词频功能；⑩步骤使用 WinGo 高级搜索功能；

图 5 WinGo 英文词频应用流程图

首先，研究人员根据理论或文献确定度量某个经济管理概念的初始单词词集或短语词集；然后，通过 WinGo 词频数据库获取目标词集或短语集在多种财经文本语料中的词频；接下

网址: www.wingodata.cn　　地址：武汉市东湖新技术开发区高新大道 788 号沃德中心
电话: 027-87419769（武汉）　　18612801983（北京）
邮箱: sales@wingodata.com（武汉）　　liufeng@wingodata.com（北京）

来，便可基于词频信息进行相关概念的测度，并可以进一步构建自己独特的文本特征指标，从而得到新的因子用以更深层次的研究（即图 5 的①②⑤⑥⑦⑧⑨步骤）。除此之外，用户还可根据理论或文献确定关键词和逻辑条件，得到需要的返回文件列表，从而进行下一步的研究（即图 5 的①⑩⑪步骤）。

目前，基于文本词频的概念测度是经济管理研究的学术前沿。例如，Loughran and McDonald（2011）通过计算财经专用积极消极词汇的词频比率构建适用于年报的语调测度方法，并研究了语调和股票收益率、交易量以及股票波动率等的关系。Hoberg and Maksimovic（2015）构造了两组“推迟投资”词语列表，其中一组是有推迟、延期、搁置含义的动词词表，另一组是与投资、项目、计划等意思相近的名词词表。并通过计算文本中两组词表中的词语、短语同时出现且其相隔不超过 12 词的频率，研究了企业的融资约束程度和企业的盈利能力以及企业的资本支出等之间的相关关系。

WinGo 美国上市公司数据库词频模块包括单词词频、短语词频以及高级搜索三大子数据库。以下为各个数据库的基本功能介绍。

2.1.1 单词词频数据库

单词词频是词频数据库实现的最基础功能。单词词频数据库除向用户提供单词的原始词频之外，还向用户提供了经过词干化处理的单词词频。词干化处理（Stem）主要是指将名词的复数变为单数和将动词的其他形态变换为基本形态。利用对单词的词干化处理，可以得到更详尽的词频信息。

2.1.2 短语词频数据库

短语是在一起使用时具有特定含义的单词的组合。短语词频数据库除向用户提供英文短语的词频之外，还兼具邻近搜索和通配符搜索功能。邻近搜索是指用户将单词间的间隔字符个数限定为 n 后（n 的范围为 0-8），数据库将输出单词间间隔字符小于等于 n 个字符的短语的频率。通配符搜索是指搜索时在单词里加入通配符‘*’和‘？’，‘*’代表零个、单个或多个字符（*前需至少有三个字符），‘？’代表单个字符。

2.1.3 高级搜索数据库

“高级搜索”功能可实现对语料中的指定词汇或短语设置搜索条件，包括 And, Not, Or, Near 等，从而获取符合搜索条件的指定词汇或短语的文件列表信息，包括公司名称、报告期间、报告链接、总词数、总句数等。从而为用户提供文件列表预览和下载，满足广大研究人员的需求。

2.2 相似词模块

构建特定的文本指标时我们一般需要用到语义相似的多个词汇，在现有的学术研究中，扩充词集的方法主要有两种：第一是通过同近义词词典人工查找来对词集进行扩充，第二是通过人工阅读所要研究的语料来扩充词集。然而，人工查找的方式往往会忽略文本语境，而

网址: www.wingodata.cn 地址：武汉市东湖新技术开发区高新大道 788 号沃德中心
电话: 027-87419769（武汉） 18612801983（北京）
邮箱: sales@wingodata.com（武汉） liufeng@wingodata.com（北京）

且存在较强的主观性偏差，因而不能全面、准确、客观地衡量文本特征。

在此情况下，WinGo 平台推出了“深度学习相似词”数据库，采用 Word Embedding（词向量）模型对海量财经文本语料进行训练，构建词汇相似度计算模型，成功提取基于财经语料的语义相似词集。这种方法打破了传统的技术壁垒，克服了现有方法的缺陷，大幅降低了研究成本。因此，在确定好初始词集后，研究人员可使用 WinGo 相似词产品（深度学习相似词）进行词集扩充（即图 5 的③④步骤）。

2.3 文本特征模块

为了提升研究效率，降低研究成本，WinGo 平台还推出了专业团队构建的文本特征指标，目前美国上市公司数据库的文本特征部分主要包括与内容无关的特征。“与内容无关的特征”指的是与文本内容不相关的一类文本特征，包括可读性、语调、文本相似性等特征。每一个文本特征，都设有上市公司披露的年度报告、季度报告、管理层讨论与分析（MD&A）、风险章节（Risk factors）、IPO 招股说明书等数据文本，旨在为广大研究人员提供简便高效的上市公司披露文本分析平台。

学者可基于文本特征数据库开展的热点研究包括但不限于以下方向：

- 文本语调与公司业务特征、高管信息披露动机
- 文本相似性与公司绩效、投资者反应
- 可读性与公司绩效、市场反应

2.4 自定义特征模块

自定义特征模块集成了 WinGo 词频模块与 WinGo 相似词模块中的两大基础功能，旨在为用户提供便捷、高效的与内容有关文本指标的构建系统。具体来讲，自定义特征指标的构建逻辑分为三步：首先，用户根据已有的研究理论，定义所构建指标的原始词集（又称种子词集）；其次，用户使用系统集成的 WinGo 深度学习相似词推荐功能对种子词集进行相似词扩充；最后，系统自动计算自定义指标词集中每个词的词频，加总归一化得到最终文本指标。

根据以上自定义指标的构建逻辑，自定义模块分别提供“特征词典定制”以及“特征计算”两大功能。针对特征词典定制，用户可以创建全新的与内容有关的文本指标，定义并修改指标对应的种子词集。此外，针对种子词集中的每个词汇，用户可以调用 WinGo 深度学习相似词推荐功能，查找对应相似词推荐结果，系统支持相似词迭代查找，能够最大程度满足用户对词集扩充的需求。针对特征计算，用户可以选择系统任意数据源作为指标计算的载体，进行简单的股票代码选择以及起始日期选择后，系统即可自动计算形成最终文本指标供用户下载使用。

网址: www.wingodata.cn 地址：武汉市东湖新技术开发区高新大道 788 号沃德中心
电话: 027-87419769（武汉） 18612801983（北京）
邮箱: sales@wingodata.com（武汉） liufeng@wingodata.com（北京）

3. WinGo 数据平台优势

3.1 权威、丰富、海量的数据来源

- 来自中国证监会官方网站、公司信息披露官方网站、中国巨潮资讯、美国证券交易委员会官方网站等
- 涵盖中国上市公司年度报告、季度报告、董事会报告、管理层讨论与分析（MD&A）、审计报告、财务报表附注、IPO 招股说明书、内部控制评价报告、业绩说明会、社会责任报告等。涵盖美国上市公司财务报告、MD&A、风险章节、IPO 招股说明书等
- 包括中国上市公司 2001 年以来披露的文本数据，共收录文档二十万余份，文字逾百亿。包括美国上市公司 1993 年以来披露的文本数据，收录文档近三百万。

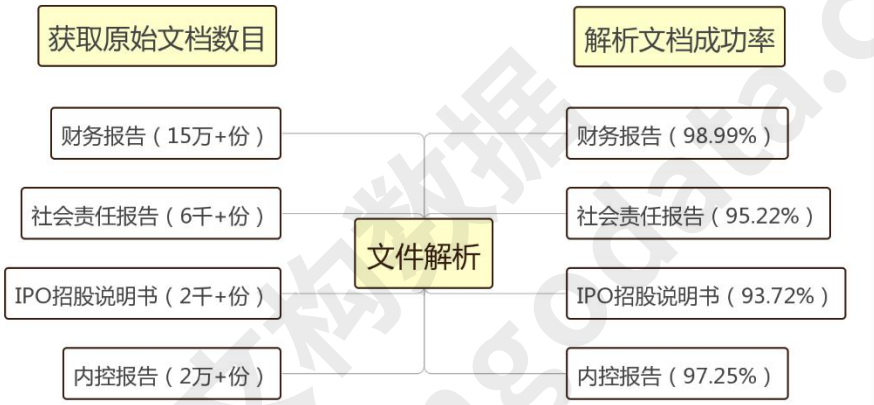


图 6 WinGo 数据平台中国上市公司文档收录部分简介（图示数据截至 2019 年 6 月）

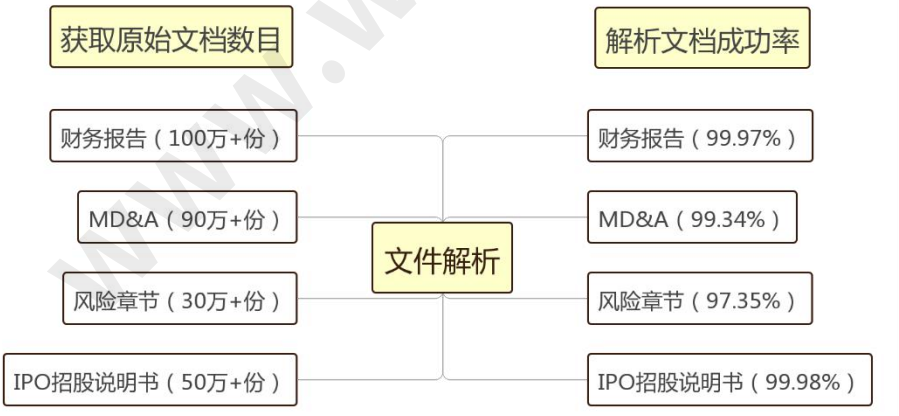


图 7 WinGo 数据平台美国上市公司文档收录部分简介（图示数据截至 2019 年 6 月）

网址: www.wingodata.cn 地址：武汉市东湖新技术开发区高新大道 788 号沃德中心
电话: 027-87419769（武汉） 18612801983（北京）
邮箱: sales@wingodata.com（武汉） liufeng@wingodata.com（北京）

3.2 专业、严谨的 PDF 解析与数据清洗

- 针对中美财经文档的不同特点，研发出独有的 PDF 解析组件，成功攻克 PDF 解析的各种技术难关，如加密 PDF 的解析、表格的识别与去除、扫描文件的解析（融合 OCR 技术）等，获取更加完整的数据
- 团队运用专业领域知识，深度研究证监会文件、权威文献以及报告语料内容，以确保报告章节数据提取准确无误（如，分析判断在不同年份、不同市场的公司财务报告中，董事会报告和管理层分析讨论章节的提取规则等）
- 紧跟学术研究前沿，严格审查原始数据，交叉检验录入数据，多重校验成品数据，以确保数据清洗严谨、数据质量高
- 团队具备多年文本数据获取及处理经验，所处理数据已被运用于大量国内外权威期刊论文

3.3 独特、智能的中文财经专用分词系统

中文词语博大精深，如何对财经专业类文本进行准确分词，一直以来都是文本挖掘的难点，这需要财经领域和语言学领域的专业人员进行判断。本平台已自主开发出适用于中文财经文本的分词系统，以及分词所需的专用财经词典。目前，WinGo 平台针对财经文本的分词效果领先于行业标准。下图是通用分词系统与 WinGo 平台分词系统的分词结果对比图。

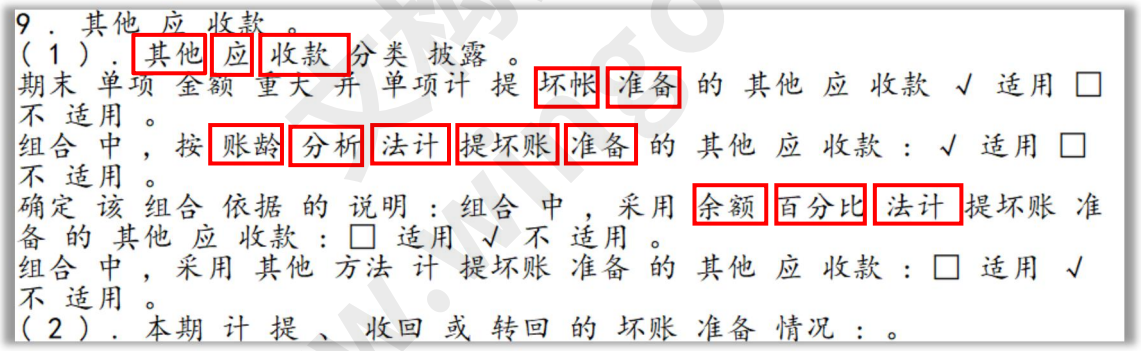

9 . 其他 应 收款 。
(1) . 其他 应 收款 分类 披露 。
期末 单项 金额 重大 并 单项计 提 坏帐 准备 的 其他 应 收款 √ 适用 □ 不 适用 。
组合 中 ， 按 账龄 分析 法计 提坏账 准备 的 其他 应 收款 ： √ 适用 □ 不 适用 。
确定 该 组合 依据 的 说明 ： 组合 中 ， 采用 余额 百分比 法计 提坏账 准备 的 其他 应 收款 ： □ 适用 √ 不 适用 。
组合 中 ， 采用 其他 方法 计 提坏账 准备 的 其他 应 收款 ： □ 适用 √ 不 适用 。
(2) . 本期 计 提 、 收回 或 转回 的 坏账 准备 情况 ： 。

图 8.1 通用分词系统分词结果

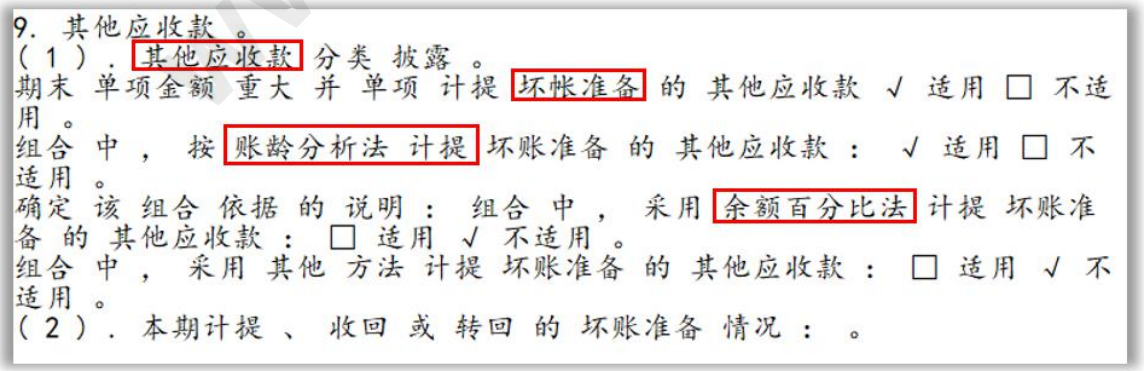

9. 其他应收款 。
(1) . 其他应收款 分类 披露 。
期末 单项金额 重大 并 单项 计提 坏帐准备 的 其他应收款 √ 适用 □ 不适用 。
组合 中 ， 按 账龄分析法 计提 坏账准备 的 其他应收款 ： √ 适用 □ 不适用 。
确定 该 组合 依据 的 说明 ： 组合 中 ， 采用 余额百分比法 计提 坏账准备 的 其他应收款 ： □ 适用 √ 不适用 。
组合 中 ， 采用 其他 方法 计提 坏账准备 的 其他应收款 ： □ 适用 √ 不适用 。
(2) . 本期计提 、 收回 或 转回 的 坏账准备 情况 ： 。

图 8.2 WinGo 平台分词系统分词结果

网址: www.wingodata.cn　　地址：武汉市东湖新技术开发区高新大道 788 号沃德中心
电话: 027-87419769（武汉）　　18612801983（北京）
邮箱: sales@wingodata.com（武汉）　　liufeng@wingodata.com（北京）

由上图可以看出，通用分词系统无法识别财经金融专业术语，对财经术语存在不当拆分（如：账龄 分析 法计）和过度拆分（如：坏账 准备、其他 应收款）的问题。而 WinGo 平台分词系统的分词结果表明，会计金融等财经专业术语均被较好地识别，不存在不当拆分和过度拆分的问题。此外，与通用分词系统相比，WinGo 分词系统还可更准确地识别新兴行业的通用词汇（如:大数据、网络游戏）、法律文件名称（如:《证券法》、《公司法》）和公司名、人名等实体名称。经专业对比计算，WinGo 中文财经专用分词系统的分词准确率达到 92%，领先于行业标准。

3.3 多功能集成的英文词频数据智能搜索服务

- 针对英文单词构词特点，采用 NLP 技术对单词进行**词干化处理**
- 短语词频向用户提供英文短语词频的同时，兼具**邻近搜索**和**通配符搜索**功能
- 高级搜索可根据**用户指定的搜索条件**提供文件列表信息，最大程度满足用户个性化需求

3.4 基于深度学习的相似词推荐系统

- 采用**深度学习（DL）**技术，训练海量上市公司披露的财经语料
- 构建词语相似度计算模型，为用户提供相似词词集以及对应相似度大小
- 不同于传统的同近义词产品，WinGo 深度学习相似词推荐系统能客观、综合地反映词语在语义、句法、上下文环境等方面的特征。具体示例结果如下表所示:

网址: www.wingodata.cn 地址：武汉市东湖新技术开发区高新大道 788 号沃德中心
电话: 027-87419769（武汉） 18612801983（北京）
邮箱: sales@wingodata.com（武汉） liufeng@wingodata.com（北京）

示例结果 1——“成本管理”

关键词	相似词	相似度	词频
成本管理	成本控制	0.854	20685
成本管理	成本管控	0.776	2371
成本管理	预算管理	0.731	13367
成本管理	全面预算管理	0.726	8167
成本管理	精细化管理	0.704	11389
成本管理	目标成本	0.673	1370
成本管理	费用控制	0.669	3730
成本管理	降本增效	0.648	8315
成本管理	过程控制	0.645	5301
成本管理	精益管理	0.642	2815

示例结果 2——“一带一路”

关键词	相似词	相似度	词频
一带一路	一路一带	0.818	278
一带一路	长江经济带	0.653	886
一带一路	京津冀一体化	0.640	406
一带一路	经济带	0.633	780
一带一路	走出去	0.628	5323
一带一路	丝绸之路	0.628	892
一带一路	中国制造 2025	0.586	2105
一带一路	经济走廊	0.571	111
一带一路	京津冀	0.560	1798
一带一路	城镇化	0.558	9088

3.5 资深、知名的跨学科团队

- 美国前 Capital One 首席统计专家、国内外知名高校教授、商业数据分析专家、自然语言处理专家全程指导产品开发
- 专业研究型金融团队和掌握前沿技术的数据分析团队通力合作，依托自主搭建的高性能计算平台和云计算技术，匠心打造国内首家财经文本人工智能研究平台
- 长期与华尔街以及清华大学等团队密切合作，紧跟业界和学术界的研究热点，为产品的可持续发展保驾护航

网址: www.wingodata.cn 地址：武汉市东湖新技术开发区高新大道 788 号沃德中心
电话: 027-87419769（武汉） 18612801983（北京）
邮箱: sales@wingodata.com（武汉） liufeng@wingodata.com（北京）

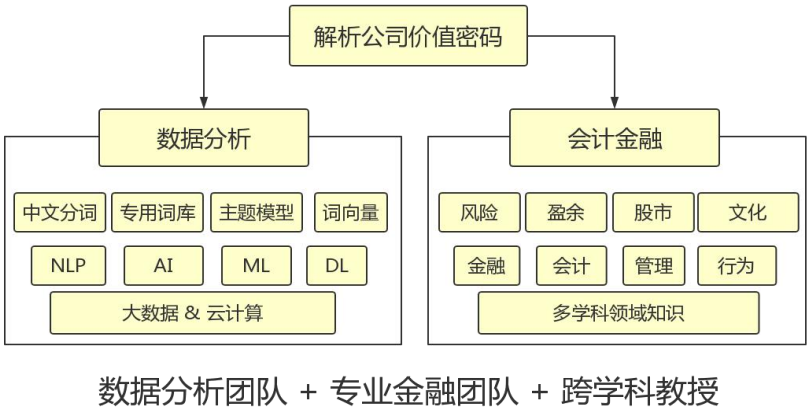


图 9 WinGo 跨学科团队

4. WinGo 平台应用

本平台针对的主要用户是经济管理领域的高等院校学者和研究人员，以及相关金融机构和专业人士。平台的应用场景主要包括学术研究、交易策略研发和验证、实践教学数据平台、业界市场调研等。

4.1 学术研究

(1) 学术研究应用场景介绍

近年来，文本信息逐渐成为国内外实证研究的热点。许多学者开始致力于运用文本分析方法来解决会计、财务、金融、经济和管理学科中的研究问题。文本信息相比于传统的数值型信息，具有以下几点优势：首先，公司披露的数值型信息受限于会计准则，难以全面、准确、客观地反映公司实际财务状况和经营成果。其次，文本信息蕴藏着数值型信息难以体现的丰富内涵。在最初阶段，采用人工阅读的方法进行文本分析不仅需要大量时间和人力，还会产生较强的主观性，因而未得到应有的重视，其信息价值未得到深度挖掘。但随着计算机科学和自然语言处理技术的发展，有关文本信息的研究已得到越来越多的关注。

目前，国外学者已经可以借助于人工智能程序来阅读海量财经专业文本，通过专家式的信息解读、特征识别与变量构建来解决众多的金融、财务和管理问题，并取得了诸多有价值的研究成果。与之相比，国内的文本研究目前尚处于起步阶段。中国人民大学姜付秀等(2015)通过计算“诚信”等关键词在年报、内部控制评价报告等文本中出现情况构建了企业诚信文化指标，发现具有诚信文化特征的企业盈余管理水平更低。

本数据库的设计、开发和核查均从前沿学术研究热点出发，具体来讲，本数据库在财务、金融和管理等领域可有以下应用：

- 文本语调与公司业务特征、高管信息披露动机
- 羊群效应、文本相似性与公司绩效、投资者反应
- 风险、创新、诚信与公司绩效、市场反应

网址: www.wingodata.cn 地址：武汉市东湖新技术开发区高新大道 788 号沃德中心
电话: 027-87419769（武汉） 18612801983（北京）
邮箱: sales@wingodata.com（武汉） liufeng@wingodata.com（北京）

(2) 国内外文本研究文献列表

国内文本相关文献列表

发表年份	论文标题	期刊名称	相关库
2019	分析师能降低股价同步性吗——基于研究报告文本分析的实证研究	中国工业经济	
2018	年报风险信息披露与审计费用——基于文本余弦相似度视角	审计研究	财务报告系列词频数据库
2018	年报文本信息复杂性与管理者自利——来自中国上市公司的证据	管理世界	财务报告系列词频数据库
2018	年报语调与内部人交易：“表里如一”还是“口是心非”?	管理世界	财务报告系列词频数据库
2018	社会关系与企业信息披露质量——基于中国上市公司年报的文本分析	南开管理评论	财务报告系列词频数据库
2018	应计操纵与年报文本信息语气操纵研究	会计研究	财务报告系列词频数据库
2018	年报风险披露与权益资本成本	金融研究	财务报告词频数据库
2017	多个大股东与企业融资约束——基于文本分析的经验证据	管理世界	财务报告系列词频数据库
2017	管理层讨论与分析披露的信息含量与股价崩盘风险——基于文本向量化方法的研究	中国工业经济	财务报告系列词频数据库
2017	年报风险信息披露有助于提高分析师预测准确度吗?	会计研究	财务报告词频数据库
2017	自愿性信息披露质量评判方法的架构与实现	统计与决策	财务报告、社会责任报告、内部控制评价报告词频数据库
2017	企业社会责任信息披露是否客观——基于文本挖掘的我国上市公司实证研究	南开管理评论	社会责任报告词频数据库
2017	分析师荐股更新利用管理层语调吗?——基于业绩说明会的文本分析	管理世界	业绩说明会词频数据库
2016	创新注意力转移、研发投入跳跃与企业绩效——来自中国 A 股上市公司的经验证据	南开管理评论	财务报告词频数据库
2016	语义分析方法在企业环境信息披露研究中的应用	会计研究	财务报告、社会责任报告词频数据库
2016	内部控制缺陷披露的经济后果分析——基于上市公司内部控制强制实施的视角	会计研究	内部控制评价报告词频数据库
2015	“诚信”的企业诚信吗?——基于盈余管理的经验证据	会计研究	财务报告、内部控制评价报告词频数据库
2015	内部控制信息披露能够降低股价崩盘风险吗?	金融研究	内部控制评价报告词频数据库
2015	管理层语调能预示公司未来业绩吗?——基于我国上市公司年度业绩说明会的文本分析	会计研究	业绩说明会词频数据库

2014	MD&A 语言特征、管理层预期与未来财务业绩——来自中国创业板上市公司的经验证据	中国软科学	财务报告词频数据库
2014	创业板上市公司文本惯性披露、信息相似度与资产定价——基于 Fama-French 改进模型的经验分析	中国管理科学	财务报告词频数据库
2012	社会责任信息披露的清晰性、第三方鉴证与个体投资者的投资决策——一项实验证据	审计研究	社会责任报告词频数据库
2010	管理层讨论与分析是否提供了有用信息?——基于亏损上市公司的实证探索	管理世界	财务报告词频数据库
2002	深市 B 股发行公司年度报告可读性特征研究	会计研究	财务报告词频数据库

国外文本相关文献列表

发表年份	论文标题	期刊名称	相关库
2019	Manager sentiment and stock returns	Journal of Financial Economics	财务报告系列词频数据库
2018	Customers'risk factor disclosures and suppliers'investment efficiency	Contemporary Accounting Research	财务报告系列词频数据库
2018	Examination of CEO-CFO social interaction through Language Style Matching: Outcomes for the CFO and the organization	Academy of Management Journal	业绩说明会词频数据库
2018	Linguistic complexity in firm disclosures: Obfuscation or information	Journal of Accounting Research	业绩说明会词频数据库
2018	Linguistic information quality in customers' forward-looking disclosures and suppliers' investment decisions	Contemporary Accounting Research	
2018	Management earnings forecasts and other forward-looking statements	Journal of Accounting and Economics	
2018	Manager-analyst conversations in earnings conference calls	Review of Accounting Studies	业绩说明会词频数据库
2018	Managers' cultural background and disclosure attributes	The Accounting Review	业绩说明会词频数据库
2018	Qualitative similarity and stock price comovement	Journal of Banking and Finance	
2018	Readability of 10-K reports and stock price crash risk	Contemporary Accounting Research	财务报告系列词频数据库
2018	Text-based industry momentum	Journal of Financial and Quantitative Analysis	财务报告系列词频数据库
2018	The influence of business strategy on annual report readability	Journal of Accounting and Public Policy	财务报告系列词频数据库

网址: www.wingodata.cn 地址: 武汉市东湖新技术开发区高新大道 788 号沃德中心
电话: 027-87419769（武汉） 18612801983（北京）
邮箱: sales@wingodata.com（武汉） liufeng@wingodata.com（北京）

发表年份	论文标题	期刊名称	相关库
2018	The effect of mandatory CSR disclosure on firm profitability and social externalities: Evidence from China	Journal of Accounting and Economics	社会责任报告词频数据库
2017	A plain English measure of financial reporting readability	Journal of Accounting and Economics	财务报告词频数据库
2017	The impact of narrative disclosure readability on bond ratings and the cost of debt	Review of Accounting Studies	财务报告词频数据库
2017	Benefits and costs of Sarbanes-Oxley Section 404 (b) exemption: Evidence from small firms' internal control disclosures	Journal of Accounting and Economics	内部控制评价报告、财务报告词频数据库
2016	Internal control over financial reporting and the safeguarding of corporate resources: Evidence from the value of cash holdings	Contemporary Accounting Research	内部控制评价报告、财务报告词频数据库
2015	Forward-looking MD&A disclosures and the information environment	Management Science	财务报告词频数据库
2015	CSR reporting practices and the quality of disclosure: An empirical analysis	Critical Perspectives on Accounting	社会责任报告词频数据库
2015	Legal opportunism, litigation risk, and IPO underpricing	Journal of Business Research	IPO 招股说明书词频数据库
2015	Do sophisticated investors interpret earnings conference call tone differently than investors at large? Evidence from short sales	Journal of Corporate Finance	业绩说明会词频数据库
2015	The effect of manager-specific optimism on the tone of earnings conference calls	Review of Accounting Studies	业绩说明会词频数据库
2014	The information content of mandatory risk factor disclosures in corporate filings	Review of Accounting Studies	财务报告词频数据库
2014	Corporate culture and CEO turnover	Journal of Corporate Finance	财务报告词频数据库
2014	Does ineffective internal control over financial reporting affect a firm's operations? Evidence from firms' inventory management	The Accounting Review	内部控制评价报告、财务报告词频数据库
2014	Knowledge, compensation, and firm value: An empirical analysis of firm communication	Journal of Accounting and Economics	业绩说明会词频数据库
2013	Individual investors and financial disclosure	Journal of Accounting and Economics	财务报告词频数据库
2013	A Measure of Competition Based on 10-K Filings	Journal of Accounting Research	财务报告词频数据库
2013	Narrative disclosure and earnings performance: Evidence from R&D disclosures	The Accounting Review	财务报告词频数据库

网址: www.wingodata.cn 地址: 武汉市东湖新技术开发区高新大道 788 号沃德中心
电话: 027-87419769（武汉） 18612801983（北京）
邮箱: sales@wingodata.com（武汉） liufeng@wingodata.com（北京）

发表年份	论文标题	期刊名称	相关库
2013	The relevance of environmental disclosures: are such disclosures incrementally informative?	Journal of Accounting and Public Policy	社会责任报告词频数据库
2012	Processing fluency and investors' reactions to disclosure readability	Journal of Accounting Research	财务报告词频数据库
2012	Litigation risk, strategic disclosure and the underpricing of initial public offerings	The Journal of Finance	IPO 招股说明书词频数据库
2012	Detecting deceptive discussions in conference calls	Journal of Accounting Research	业绩说明会词频数据库
2011	The effect of annual report readability on analyst following and the properties of their earnings forecasts	The Accounting Review	财务报告词频数据库
2011	When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks	The Journal of Finance	财务报告词频数据库
2010	Management's tone change, post earnings announcement drift and accruals	Review of Accounting Studies	财务报告词频数据库
2010	The information content of forward-looking statements in corporate filings—A naïve Bayesian machine learning approach	Journal of Accounting Research	财务报告词频数据库
2010	The effects of disclosure complexity on small and large investor trading.	The Accounting Review	财务报告词频数据库
2010	The effects of ambiguous information on initial and subsequent IPO returns	Financial Management	IPO 招股说明书词频数据库
2010	The information content of IPO prospectuses	Review of Accounting Studies	IPO 招股说明书词频数据库
2010	Does silence speak? An empirical analysis of disclosure choices during conference calls	Journal of Accounting Research	业绩说明会词频数据库
2009	How does financial reporting quality relate to investment efficiency?	Journal of Accounting and Economics	财务报告词频数据库
2009	Financial reporting complexity and investor underreaction to 10-K information	Review of Accounting Studies	财务报告词频数据库
2008	Annual report readability, current earnings, and earnings persistence	Journal of Accounting and Economics	财务报告词频数据库
2008	Evidence on the audit risk model: Do auditors increase audit fees in the presence of internal control deficiencies?	Contemporary Accounting Research	内部控制评价报告、财务报告词频数据库

网址: www.wingodata.cn 地址: 武汉市东湖新技术开发区高新大道 788 号沃德中心
电话: 027-87419769（武汉） 18612801983（北京）
邮箱: sales@wingodata.com（武汉） liufeng@wingodata.com（北京）

4.2 交易策略的研发和验证

金融市场的量化交易者可基于数据库所提供的词频、文本特征指标和会计金融指标等数据构建更加全面、可靠的交易策略，例如将风险词频或者风险文本特征指标引入到多因子模型中对交易资产的未来收益和波动进行估计。本数据库所提供长达近 18 年的历史量化财务报告数据则为这些交易者的策略研究，以及回测验证都提供了丰富可靠的数据样本。

4.3 实践教学数据平台

如何应用文本挖掘和机器学习技术解决金融经济领域的问题是当前学界和业界的前沿趋势，而本数据平台则为经济管理类专业的学生在相关课程实践环节中学习和应用文本挖掘技术提供了有力的支持。例如，结合本数据平台和股市历史行情数据，金融工程专业的学生可学习如何制定和评测基于上市公司披露文本信息的交易策略。此外，本数据平台还提供了所有文本的原始网页链接，部分经济管理类专业的进阶课程乃至计算机相关专业的学生可据此进一步学习并实践数据采集、清洗和分析过程中的相关技术，并根据数据库中的数据分析结果进行比对查验。

4.4 业界市场调研

上市公司自身在进行市场研究、制定风险管理策略，或者投资机构在进行投资决策之前都希望对相关上市公司及其竞争对手披露的财经专业文本中蕴含的非财务信息进行全面持久的了解和追踪。通过本数据平台，相关从业人员能方便快捷地检索并统计整理出海量上市公司的财经专业文本数据，并据此撰写调研分析报告。

联系我们

电话：027-87419769（武汉），18612801983（北京）
邮件：sales@wingodata.com（武汉），liufeng@wingodata.com（北京）
网址：www.wingodata.cn
地址：湖北省武汉市东湖新技术开发区高新大道 788 号沃德中心