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，聚焦于海量财经文本数据，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，为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。

# WinGo数据库内容

WinGo数据平台由词频、相似词、文本特征指标、会计与金融指标以及在线定制五大数据库组成。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、内部控制评价报告、社会责任报告、业绩说明会、IPO招股说明书、审计报告等。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数据库内容

**1.1 词频系列数据库**

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 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 相似词系列数据库**

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 文本指标数据库**

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，包括羊群效应、语调、文本相似性等特征。“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。

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
* 羊群效应、文本相似性与公司前景、投资者反应
* 风险、创新、诚信与公司绩效、市场反应

**1.4 其他数据库**

此外WinGo数据平台还推出了会计与金融指标系列数据库和在线定制服务：

会计与金融指标系列数据库包括“会计财务系列指标”和“金融系列指标”。其中，“会计财务系列指标”涵盖盈余质量、会计稳健性、会计可比性等相关指标。“金融系列指标”涵盖股价同步性、超额收益等金融指标。

在线定制服务可为用户提供基于文本的个性化服务，包括语料服务、模型服务和文本特征服务三大部分。其中语料服务涵盖在线分词、停用词管理、词频统计分析、PDF解析等。模型服务涵盖LDA主题模型、STM主题模型、Word2Vec模型和Doc2Vec模型。文本特征服务涵盖特征词典定制以及特征计算。

# WinGo数据平台优势

图3 WinGo数据平台业务流程图

**2.1 权威、丰富、海量的数据来源**

* 来自证监会官方网站、公司信息披露官方网站、巨潮资讯等
* 涵盖上市公司年度报告、季度报告、内部控制评价报告、社会责任报告、业绩说明会、IPO招股说明书、审计报告等
* 包括中国上市公司2001年以来披露的文本数据，共收录文档十六万余份，文字逾百亿

图4 WinGo数据平台文档收录部分简介（图示数据截至2018年08月）

**2.2 专业、严谨的PDF解析与数据清洗**

* 针对中国财经文档的特点，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

**2.3 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远超行业标准。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。



图5.1 通用分词系统分词结果



图5.2 WinGo平台分词系统分词结果

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。

**2.4 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：

**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 |