WinGo Data 20190915

文构数据

WinGo 财经文本数据平台

---中国政府文本数据库---

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

WinGo 财经文本数据平台——中国政府文本数据库简介

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及中国政府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，聚焦于中美海量文本数据。应用自然语言处理、深度学习和人工智能技术对目标文本进行深度加工，给用户提供目标文本的词频、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、美国上市公司和中国政府文本三大数据库，由业内专家和高校知名学者主持设计，打破了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。以下是平台中国政府文本数据库简介：

1. 中国政府文本数据库内容

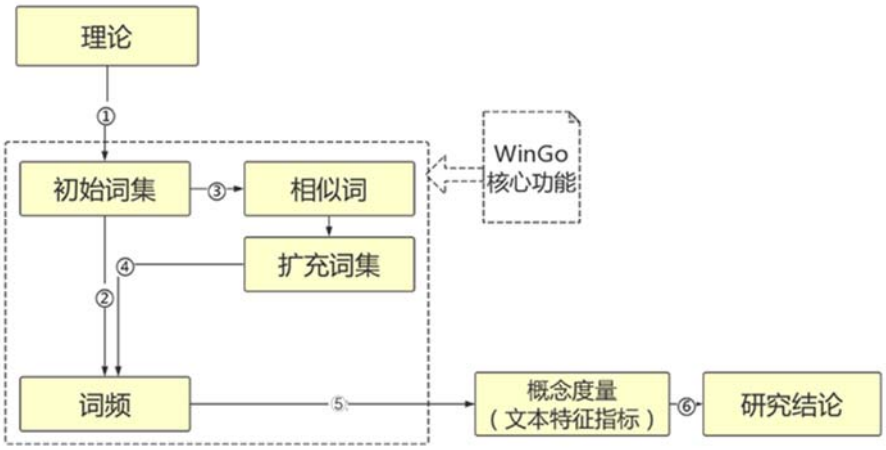
WinGo 平台中国政府文本数据库由词频、相似词、自定义特征三大模块组成。数据源涵盖了国务院、省级、地级的政府工作报告，国务院政府工作报告数据区间最早可追溯到 1954 年，省、地级政府工作报告数据区间可追溯到 2010 年，2001 年-2009 年数据也即将上线。

图 1 WinGo 中国政府文本数据库内容

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：

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）



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 相似词功能使用 WinGo 短语词频功能；⑤步骤使用 WinGo 高级搜索功能；

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政府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，余永泽（2019）以 2002-2014 年中国 230 个地级市政府工作报告为研究对象，以经济增长目标用语是否包含“之上”、“确保”、“力争”等词汇的条件来定义其是否属于具有较强约束力的经济增长目标设定方式，检验了地方经济增长目标的制定对全要素生产率的影响。邓雪琳（2015）采用“经济建设、政治建设、文化建设、社会建设、生态文明建设”中国政府“五位一体”的职能分析框架，通过对国务院政府工作报告中的关键词、高频词、关键段落字数开展计量，回溯性地测量了改革开放以来中国政府职能转变的特点并预测了中国政府职能未来转变的趋势。

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）模型对海量政府文本语料进行训练，构建词汇相似度计算模型，成功提取基于政府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）。

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）

1.3 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。

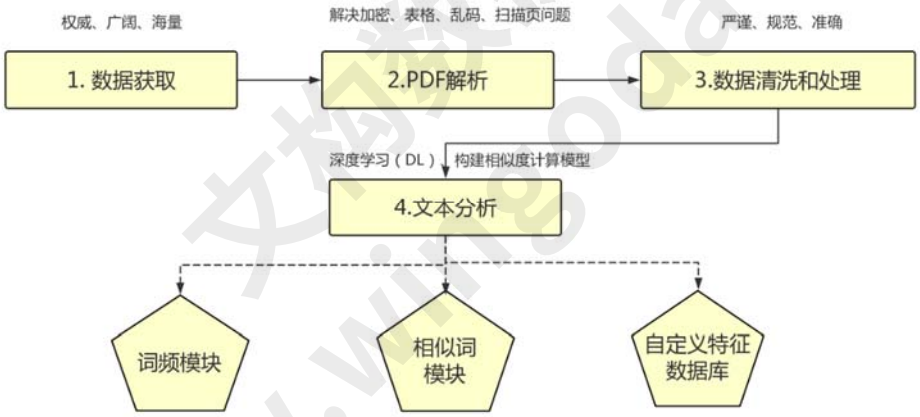


图 3 WinGo 数据平台业务流程图（中国政府文本数据库）

2. 中国政府文本数据库优势

2.1 权威、丰富、全面的数据来源

- 来自政府官方门户网站、地方年鉴、地方机关报纸。
- 涵盖国务院、省级、地级政府工作报告。
- 数据全面，收录了超过 99%的政府工作报告。

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）

2.2 专业、严谨的 PDF 解析与数据清洗

- 针对中文档的特点，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
- 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 词频数据智能搜索，提供多功能集成服务

中文语言博大精深，如何对不同类型文本进行准确分词，一直以来都是文本挖掘的难点，这需要经济管理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政府文本的分词系统，以及分词所需的政府文本专用词典。目前，WinGo 平台针对政府文本的分词效果领先于行业标准。

通用分词系统无法识别政府文本专业术语，对术语存在不当拆分和过度拆分的问题。而 WinGo 平台分词系统的分词结果表明，政府文本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、法律文件名称和公司名、地名等实体名称。经专业对比计算，WinGo 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，训练海量政府披露的文本语料
- 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

示例结果 1——“互联网+”

关键词	相似词	相似度	词频
互联网+	互联网	0.7736	3918
互联网+	数字家庭	0.5032	41
互联网+	大数据	0.4815	3240
互联网+	智慧农业	0.4649	123
互联网+	互联网+农业	0.458	34
互联网+	云服务	0.4544	115
互联网+	物联网	0.4515	1169
互联网+	智慧医疗	0.4441	156
互联网+	中国制造 2025	0.4284	598

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）

互联网+	信息化	0.4283	6030

示例结果 2——“三去一降一补”

关键词	相似词	相似度	词频
三去一降一补	去产能	0.6081	747
三去一降一补	供给侧结构性改革	0.5567	3513
三去一降一补	降成本	0.5303	624
三去一降一补	去库存	0.5173	680
三去一降一补	供给侧改革	0.4964	91
三去一降一补	去杠杆	0.4741	355
三去一降一补	结构性改革	0.4469	232
三去一降一补	经济体制改革	0.4409	431
三去一降一补	稳增长	0.4378	3478
三去一降一补	僵尸企业	0.4344	522

示例结果 3——“农业”

关键词	相似词	相似度	词频
农业	农牧业	0.7885	2619
农业	现代农业	0.6928	8083
农业	种养业	0.6681	334
农业	农业生产	0.648	1574
农业	农业产业	0.6277	1132
农业	种植业	0.6258	742
农业	农业产业化	0.6036	6167
农业	渔业	0.575	909
农业	生态农业	0.5749	1272
农业	畜牧业	0.5719	2999

2.5 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、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）

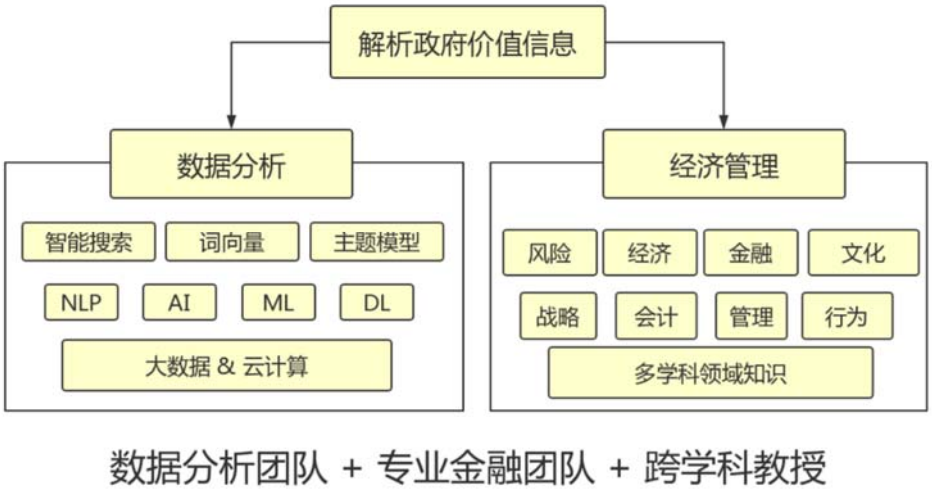


图 4 WinGo 跨学科团队

3. 中国政府文本数据库应用

本平台针对的主要用户是经济管理以及社会科学领域的高等院校学者和研究人员，以及相关企业用户和专业人士。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、企业的战略和实施等。

3.1 学术研究

近年来，政府工作报告逐渐成为国内外社会科学实证研究的热点。之前的相关研究集中于从某一角度出发对历年政府工作报告进行回顾，或仅仅使用政府工作报告中提及的数值指标来进行研究。随着计算机技术的发展，目前在经管、财务、社会科学和情报学等领域，许多学者基于中国政府报告等文本数据源，通过文本分析提取关键词汇或关键句，构建“生态”、“环保”、“住房调控”、“外交”、“海洋事务”等等指标，进行相关研究。

目前，国内学者已经可以借助于人工智能程序来阅读中国政府文本，通过专家式的信息解读、特征识别与变量构建来研究众多的社科和经管问题，并取得了诸多有价值的研究成果。例如，余永泽（2019）以 2002-2014 年中国 230 个地级市政府工作报告为研究对象，以经济增长目标用语是否包含“之上”、“确保”、“力争”等**词汇**的条件来定义其是否属于具有较强约束力的经济增长目标设定方式，检验了地方经济增长目标的制定对全要素生产率的影响。邓雪琳（2015）采用“经济建设、政治建设、文化建设、社会建设、生态文明建设”中国政府“五位一体”的职能分析框架，通过对国务院政府工作报告中的**关键词**、**高频词**、**关键段落字数**开展计量，回溯性地测量了改革开放以来中国政府职能转变的特点并预测了中国政府职能未来转变的趋势。

本数据库的设计、开发和核查均从前沿学术研究热点出发，旨在打破政府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。

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 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、相似词等数据构建更加全面、可靠的交易策略，例如将生态、住房调控、外交等构建所得的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。本数据库所提供丰富、完备的政府文本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。

3.3 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以及社会科学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、社会科学、情报学等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，本数据平台可为部分专业的进阶课程提供数据帮助。

3.4 企业战略的制定和实施

依据 PEST 模型，企业在制定长短期战略时需要虑政治环境（P）、经济环境（E）、社会环境（S）以及技术环境（T）因素的影响。政府工作报告中包含了与对过去一年工作业绩与问题的回顾性总结以及对未来一年政府工作的预期目标，以及具体实施方案与安排，因此包含了关于法律、经济、技术、人文、民生等多方面的内容，是在进行 PEST 分析时不可或缺的信息来源之一。本数据平台为企业提供近十几年内的地市级政府工作报告，并提供数据查找对比平台，方便企业制定适合的长短期战略。

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