

WinGo 财经文本数据平台

---操作指引---



深圳希施玛数据科技有限公司

地址：深圳市南山区学苑大道 1001 号南山智园 A4 栋 4 楼

网址：cn.gtadata.com

客服电话：400-639-8883

目录

一、平台介绍.....	1
二、中国上市公司文本数据库.....	1
2.1 词频模块.....	1
2.1.1 词频搜索.....	2
2.1.2 邻近搜索.....	3
2.1.3 高级搜索.....	3
2.2 相似词工具模块.....	3
2.3 文本特征模块.....	4
2.4 自定义特征模块.....	4
三、美国上市公司文本数据库.....	5
3.1 词频模块.....	5
3.1.1 单词词频.....	7
3.1.2 短语词频.....	7
3.1.3 高级搜索.....	7
3.2 相似词工具模块.....	7
3.3 文本特征模块.....	8
3.4 自定义特征模块.....	8
四、中国政府文本数据库.....	9
4.1 词频模块.....	9
4.1.1 词频搜索.....	10
4.1.2 邻近搜索.....	11

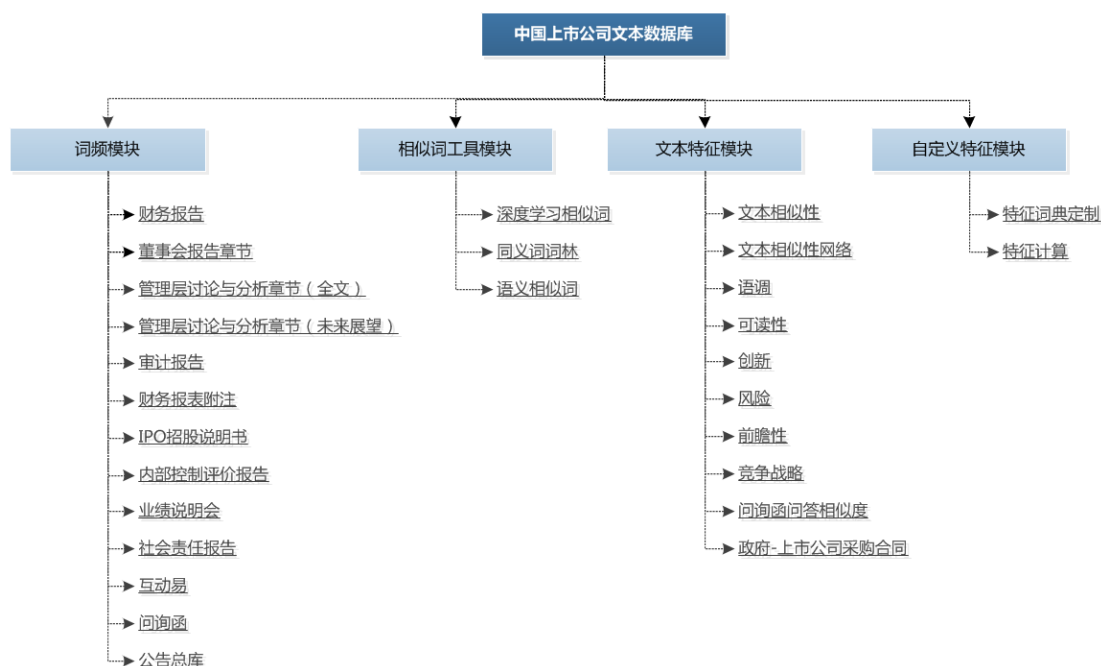
4.1.3 高级搜索.....	11
4.2 相似词工具模块.....	11
4.3 文本特征模块.....	12
4.4 自定义特征模块.....	12

一、平台介绍

WinGo 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，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。平台由中国上市公司、美国上市公司、中国政府文本三大数据库组成。

二、中国上市公司文本数据库

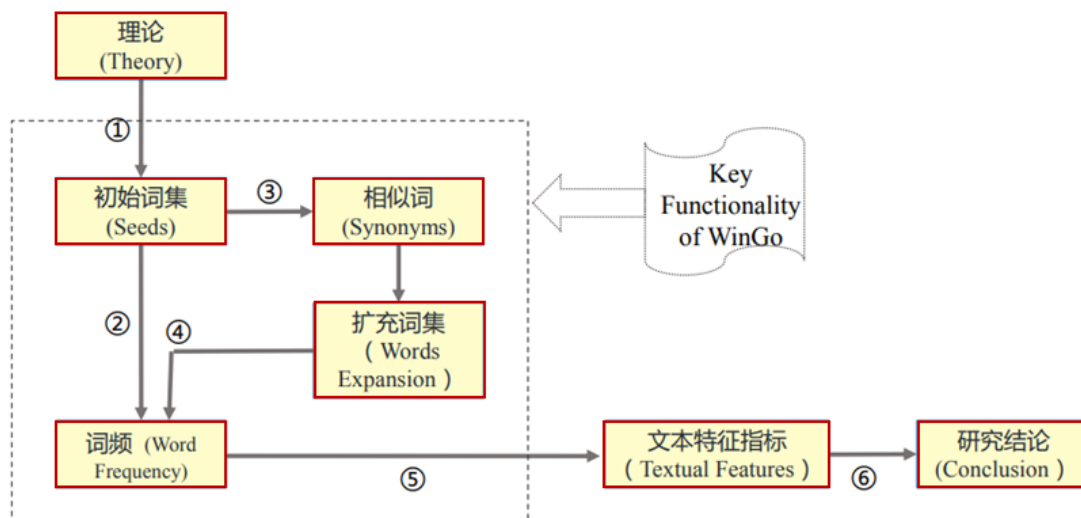
中国上市公司文本数据库由词频、相似词工具、文本特征、自定义特征四大模块组成。数据源涵盖范围广阔，囊括了上市公司披露的年度报告、季度报告、董事会报告、管理层讨论与分析（MD&A）、审计报告、财务报表附注、IPO 招股说明书、内部控制评价报告、业绩说明会、社会责任报告等，数据区间最早可追溯到 2001 年。



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：①③⑥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

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（即上图的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。

2.1.1 词频搜索

词频搜索是词频数据库实现的最基础功能。词频数据库除向用户提供词汇的原始词频之外，还向用户提供了语料的基本信息表，包括报告大小、总字数、总字数（去除数字）、总

字数（去除数字、英文）、总词数、总词数（去除数字）、总词数（去除数字、英文）、总句数、总句数（去除数字）、总句数（去除数字、英文）等数据内容。

2.1.2 邻近搜索

邻近搜索指的是将词语间的间隔字符数限定为 n 后（ n 的范围为 0-15），数据库将输出词语间间隔字符数小于等于 n 个字符数的词组的频率。词组需至少包含两个词汇，词汇字数无限制，词汇之间用空格隔开。例如，查找时输入“推迟 投资”，回车即可输入词组；若选择“指定相邻两词汇间隔字符数不超过 4”，则将输出“推迟”和“投资”间字符数小于等于 4 的词组的个数。

2.1.3 高级搜索

高级搜索指的是用户可对语料中的指定词汇设置搜索条件，包括 And, Not, Or, Near 等，从而获取出现符合搜索条件的指定词汇的文件列表和语句信息，包括公司名称、报告期间、报告链接、总词数、总句数等。对于包含英文字母的关键词，本系统默认为小写。此外，关键词只针对两个及两个以上字符的词汇（纯数字除外）。逻辑运算优先级顺序为：括号内为第一优先级，括号外 NOT>AND>OR。

2.2 相似词工具模块

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（即上图的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，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

2.4 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。

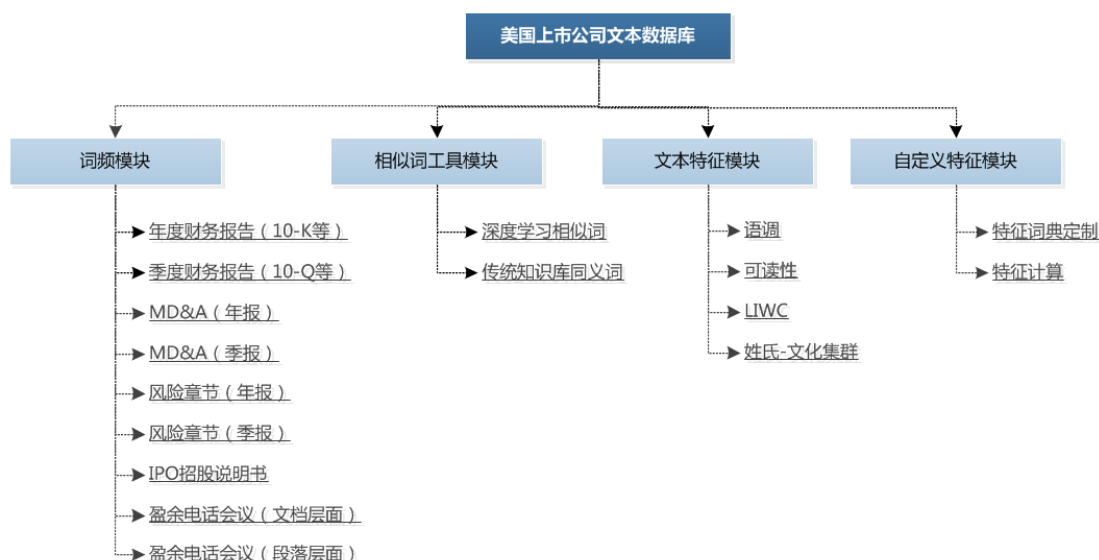
具体来讲，自定义特征指标的构建逻辑分为三步：首先，用户根据已有的研究理论，定义所构建指标的原始词集（又称种子词集）；其次，用户使用系统集成的 WinGo 深度学习相似词推荐功能对种子词集进行相似词扩充；最后，系统自动计算自定义指标词集中每个词的词频，加总归一化得到最终文本指标。

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。

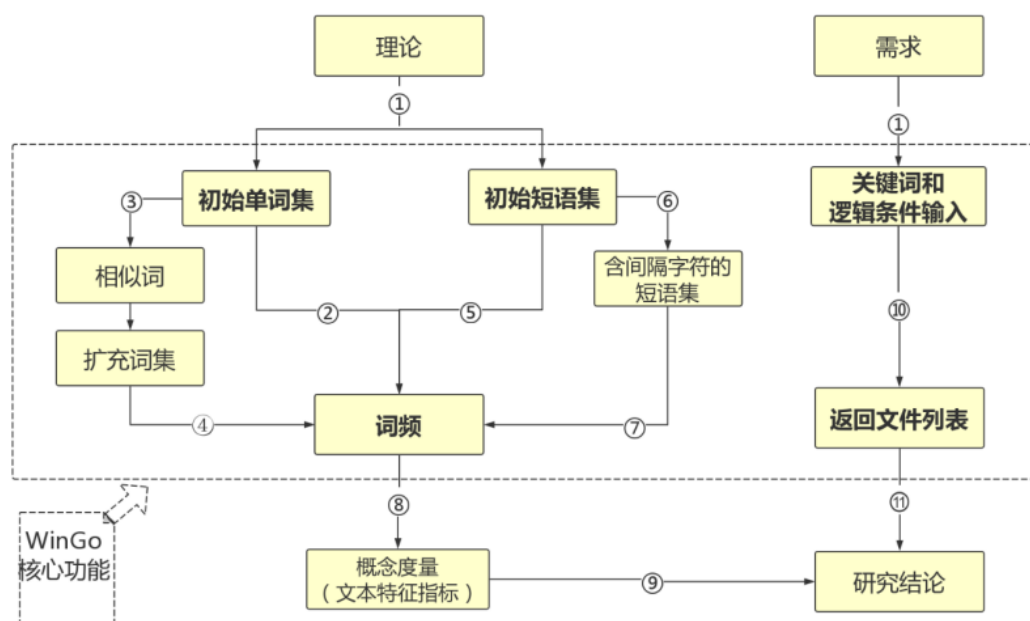
三、美国上市公司文本数据库

美国上市公司文本数据库由词频、相似词工具、文本特征、自定义特征四大模块组成。数据源涵盖范围广阔，囊括了美国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。



3.1 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：



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(即上图的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(即上图的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 美国上市公司文本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。

3.1.1 单词词频

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。

3.1.2 短语词频

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。

3.1.3 高级搜索

“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。

3.2 相似词工具模块

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（即上图的③④步骤）。

3.3 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

3.4 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。

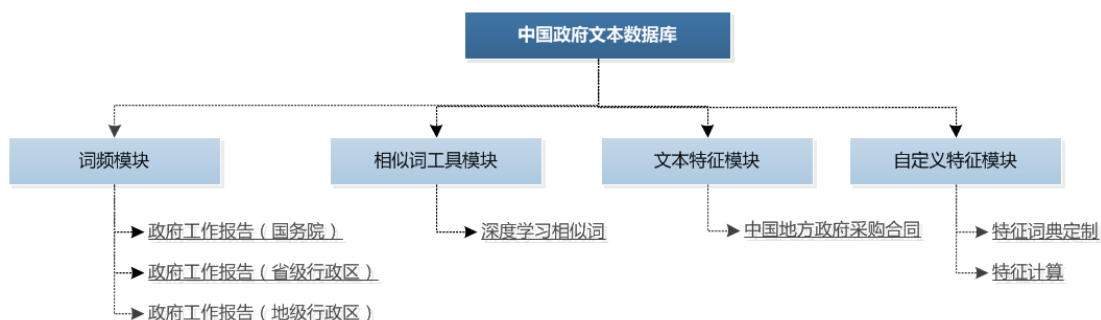
具体来讲，自定义特征指标的构建逻辑分为三步：首先，用户根据已有的研究理论，定义所构建指标的原始词集（又称种子词集）；其次，用户使用系统集成的 WinGo 深度学习

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。

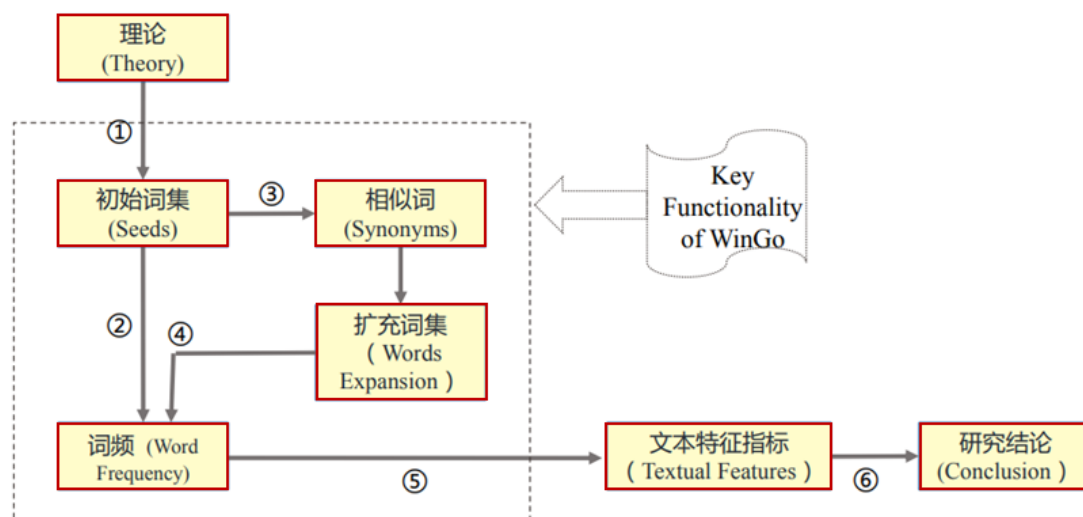
四、中国政府文本数据库

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 2001 年。



4.1 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

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（即上图的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。

4.1.1 词频搜索

词频搜索是词频数据库实现的最基础功能。词频数据库除向用户提供词汇的原始词频之外，还向用户提供了语料的基本信息表，包括报告大小、总字数、总字数（去除数字）、总

字数（去除数字、英文）、总词数、总词数（去除数字）、总词数（去除数字、英文）、总句数、总句数（去除数字）、总句数（去除数字、英文）等数据内容。

4.1.2 邻近搜索

邻近搜索指的是将词语间的间隔字符数限定为 n 后（ n 的范围为 0-15），数据库将输出词语间间隔字符数小于等于 n 个字符数的词组的频率。词组需至少包含两个词汇，词汇字数无限制，词汇之间用空格隔开。例如，查找时输入“推迟 投资”，回车即可输入词组；若选择“指定相邻两词汇间隔字符数不超过 4”，则将输出“推迟”和“投资”间字符数小于等于 4 的词组的个数。

4.1.3 高级搜索

高级搜索指的是用户可对语料中的指定词汇设置搜索条件，包括 And, Not, Or, Near 等，从而获取出现符合搜索条件的指定词汇的文件列表和语句信息，包括公司名称、报告期间、报告链接、总词数、总句数等。对于包含英文字母的关键词，本系统默认为小写。此外，关键词只针对两个及两个以上字符的词汇（纯数字除外）。逻辑运算优先级顺序为：括号内为第一优先级，括号外 NOT>AND>OR。

4.2 相似词工具模块

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（即上图的③④步骤）。

4.3 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，目前中国政府文本数据库的文本特征部分主要包括中国地方政府采购合同。中国地方政府采购合同数据是中国政府采购网中国政府购买服务信息平台披露的政府采购合同信息，主要囊括了采购人（甲方）、采购人所属行政区、供应商（乙方）以及合同金额等关键信息。

4.4 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。