1.《中国政府推进基本公共服务的注意力测量——基于中央政府工作报告(1954-2013)的文本分析》，吉林大学社会科学学报，2014 年 3 月 54 卷 2 期

文本分析工具：
本文应用 QSR NVivo9 这款自动文本分析工具，QSR NVivo 9 软件通常认为文本内容会反映当事人对某件事务的价值判断及认知情形，而文本中当事人认知模式的高低变化反映了当事者的价值判断与关注程度。QSR NVivo9 软件通常会用某一特定词汇或句子出现的频数或频率来测量决策者注意力的配置。
笔者通过 QSR NVivo 9 软件,运用文本分析方法,选择与基本公共服务均等化各个维度相关的关键词进行分类汇总,逐一分析特定词频及其语义环境所隐含的信息,挖掘与之相关的组织行为。在此基础上,通过分析种子词语在文本中出现的频率,测量政府对基本公共服务问题的关注度。

基本步骤：
1. POST CM 6 分词软件，分词后导出分词结果，得出分词的词频排名。
2. 通过专家投票的方式，确定一级关键词和二级关键词
3. 使用 QSR NVivo9 软件，按照关键词进行编码，得到每个关键词的词频。当文本中的一个句子出现多个关键词时，每个关键词只编码一次，当一个句子中关键词含义不同时，对该语句进行重复计数。

2.《改革开放以来中国政府职能转变的测量——基于国务院政府工作报告(1879-2015)的文本分析》2015.08.05

文本分析工具：
ROST Content Mining 6.0 文本挖掘软件

建立的数据库
1.关键词数据库，利用软件找出每年政府工作报告中几个关键词（例如经济建设、政治建设、文化建设、社会建设、环境保护）的词频数。
2.词频库（两个）：词频前 500 位的政府年度工作报告高频词，词频前 300 位的政府年度工作报告高频词。
3.段落字数数据库：将 38 年的政府工作报告内容分年按照经济建设、政治建设、文化建设、社会建设、生态文明建设、其他这几大类分类计算每类的段落字数

基本步骤：
1. 数据预处理
   第一步，计算 38 年政府年度工作报告词频数排名在前 500 的高频词平均每个高频词每年的词频数为 5.9 次。
   第二步，计算出每年政府年度工作报告词频数排名在前 300 的高频词的平均词频数。
   第三步，选取词频数高于每年词频数排名在前 300 的高频词的平均词频数，并且高于 5.9 次的高频词建立一个数据库。
   第四步，删除没有实质性意义的或专指程度比较低的高频词及频次，如“我国”、“今

年”、“百分之”、“一九”等，最终有 3406 个高频词成为分析对象，总词频数为 64187 次。

2. 数据分类编码统计

根据经济建设、政治建设、文化建设、社会建设、生态文明建设这几个核心概念的内涵，将 3406 个高频词中明显属于这五大类的分年进行分类编码，并将属于这五大类的高频词的词频数进行加总。

对高频词的分类是严格按照内容分析法规范执行的:提出研究问题或假设，界定样本框，抽取样本，建构类目并界定分析单位，建立量化系统，执行预测，建立信度，最后依照定义编码，分析资料。按照内容分析法来操作高频词的分类，能有效保证词频的选择范围和所要描述的指标之间的契合度和归类的合理性。但考虑到篇幅限制以及本论文研究的主要问题，在此仅介绍信度的处置问题。在这次编码中，采用了 2 个编码员进行编码。在预测中运用了信度统计公式获得信度系数，其中 M 为完全同意数目;N1 表示第一位编码员应有的同意数目;N2 表示第二位编码员应有的同意数目;n 表示编码员人数。经过两轮预测后，正式确定这些高频词的类别，第一轮预测取得的信度是 0.816;第二轮取得的信度为 0.943，具体计算见公式(1)和(2)。对第三类数据库数据编码为了提高信度也采用了此方法。

$$\text{相同同意度} = \frac{2M}{N_1+N_2} \quad (1)$$

$$\text{信度} = \frac{\text{n}\times(\text{平均相同互同意度})}{1+\left[(n-1)\times\text{平均相互同意度}\right]} \quad (2)$$

3.《中国赴澳大利亚游客的情感特征研究——基于大数据的文本分析》

文本分析工具:

ROST Content Mining 6.0 文本挖掘软件

基本步骤:

1. 构建旅游分析词库。

在数据处理上，本次研究首先构建了基础旅游分析词库。该词库以 HowNet（知网）词典为基础词库，再通过大量读取（超过 200 篇游记）和整理旅游在线评论、游记、旅游文献，提炼出旅游专属词库。该词库内容覆盖旅游景区、餐饮、交通、住宿、娱乐、购物 6 个方面，共包含 317 个正面词汇和 185 个负面词汇。与 HowNet 词典相比，本次人工筛选新增 298 个词汇，只有 40%与 Hownet 的词汇重合。与此同时，本文也对 HowNet 词典进行修正，删减部分只有在特定语境下才会表达出情感偏向的词汇和具有二义性的词汇。最终所构建的完整游客情感评价词库共包含 3507 个正面词汇和 3365 个负面词汇。

2. 对情感评价词前的程度副词、否定副词、转折词的作用进行梳理和解析。对于不同的词语赋予不同的系数。见表 2

表2 语义逻辑词类和系数判定表

Tab.2 Judging rules of semantic logic and coefficient

词类 Parts of speech	级别 Magnitude	个数 Number of the words	计分方法 Coefficients	示例 Examples
前置副词表 Pre-adverb	极其/最	42	3倍	“万分”“无比”
	超	16	2.5倍	“超”“何止”
	很	46	2倍	“大为”“多”
	较	18	1.5倍	“还”“较”
	稍/欠	28	0.5倍	“略”
后置副词表 Post-adverb	极其/最	12	3倍	“不得了”
	超	3	2.5倍	“不为过”
	很	4	2倍	“不少”
	较	1	1.5倍	“如斯”
	稍/欠	1	0.5倍	“些”
否定词表 Negative adverb		30	奇数重否定，表否定意义，系数-1；偶数重否定，表肯定意义，系数+1	“不”“没有”
转折词表 Adversative	第一类转折词	14	0.5倍	“虽然”“虽是”
	第二类转折词	14	2倍	“但是”“可是”

数据来源：本文作者根据HowNet整理。

3. 应用 ROST CM6 对已区分的正负面评论进行量化处理，生成关键词网络图，用于进一步分析正负面评价的结构于特征。

4.《中国易地扶贫搬迁政策的演进特征——基于政策文本量化分析》

文本分析工具：
QSR NVivo 9

基本步骤：

1．检索收集政策文件。
本文利用北大法宝———中国法律检索系统以及各部委的政府官方网站收集了政策文本，通过检索“易地扶贫搬迁”和“易地搬迁”两组关键词，共检索出 106 个。经过进一步剔除重复和无关文件，本文共筛选得出政策文本 90 个

2．应用 NVivo，归纳出 7 个政策维度：财政与金融政策、机制创新、工程项目监管、民生保障、管理体制、土地政策、生态环境政策。

3．统计不同阶段中每个维度的数量。

5.《技术创新政策对企业创新绩效影响研究——基于政策文本分析》

文本分析工具：
专家评审方法

基本步骤：

1. 收集技术创新政策文本。
本文主要以 2006-2015 以来国家及地方颁布的与北京市电子信息企业相关的技术创新政策为研究样本，通过浏览政策文件、查找与北京市电子信息企业创新相关的关键词，从国家工业和信息化部、国家发展与改革委员会、北京市发展与改革委员会及北京市经济和信息化委员会等部门网站的政策文件中收集到技术创新政策 61 项。

2. 制定政策分类标准
本文将政策分为政策力度、政策目标和政策作用对象 3 维度。每个维度又细分为几个方

面，每个方面建立了五个评价等级，形成量化矩阵。

3. 请专家根据量化标准对政策一一打分，将得分结果取平均值后录入量化矩阵中。

6.总结

基于文本分析的研究总的来说有以下几个步骤:

1．收集文本。

论文 1、2 收集的是政府工作报告，论文 3 是社交网站公开资料，论文 4、5 是国家颁布的政策文本。

2．制定评价维度（关键字）。

文章 3、4、5 应用专家评审的方式制定评价维度，文章 1、2 使用统计高频词和专家筛选的方式制定统计的关键词。

3．对步骤 1 收集的文本按评价维度进行统计。

词频统计软件有 QSR NVivo9（文章 1、4)和 ROST Content Mining 6.0（文章 2、3）两种，其次还有专家评审统计方法[5]。除此之外也可以使用 python 的 jieba 分词实现词频统计[1]。

4．根据统计结果进行进一步分析。

[1] https://github.com/fxsjy/jieba