

# 政府工作报告文本分析

21 国贸 张兆阳

22 金工 郭欣雨

22 信管 刘慧泓

# 项目介绍

对江苏省的政府工作报告进行文本分析，并对比其与长三角地区其他省份（安徽、浙江、上海）的区别。

- 根据TF-IDF算法，提取历年江苏省政府工作报告中的关键词，观察其变化趋势。
- 构建政府环境关注度和数字化关注度指标，计算长三角地区各省份的政府关注度，并对比江苏省与长三角其他省份的区别。

# 数据来源

- 省政府官网（以江苏省政府为例）



# 文本预处理

- 获取政府工作报告原始文本（爬虫+正则表达式匹配）
  - 爬取历年政府工作报告的网页链接
  - 从网页源代码提取政府工作报告内容并保存为txt文件
- 对原始文本进行清洗
  - 分词（jieba分词）
  - 去除停用词（中文停用词表+手动补充）

# 文本预处理

	year	tokens
1	2005	代表 代表 江苏省人民政府 大会 报告 政协委员 意见 回顾过去 改革开放 经济社会 发展 ...
2	2006	政府 报告 人民代表大会 会议 省长 代表 代表 江苏省人民政府 大会 报告 步入 发展 时...
3	2007	政府 报告 人民代表大会 会议 省长 代表 代表 江苏省人民政府 大会 报告 政协委员 意见...
4	2008	政府 报告 人民代表大会 会议 代省长 志军 代表 代表 江苏省人民政府 大会 报告 建议 ...
5	2009	政府 报告 人民代表大会 会议 省长 志军 代表 代表 人民政府 大会 报告 政协委员 意见...
6	2010	政府 报告 人民代表大会 会议 省长 志军 代表 代表 人民政府 大会 报告 政协委员 意见...
7	2011	政府 报告 人民代表大会 会议 代表 代表 江苏省人民政府 大会 报告 政协委员 意见 时期...
8	2012	政府 报告 人民代表大会 会议 省长 代表 代表 江苏省人民政府 大会 报告 政协委员 意见...
9	2013	政府 报告 人民代表大会 会议 省长 代表 代表 江苏省人民政府 大会 报告 政协委员 意见...

## 关键词分析

- 根据TF-IDF算法，提取每年政府工作报告中的关键词
  - 生成语料库的TF-IDF矩阵（sklearn库TfidfVectorizer函数）

[illegible]

# 关键词分析

- 根据TF-IDF算法，提取每年政府工作报告中的关键词
  - 生成语料库的TF-IDF矩阵（sklearn库TfidfVectorizer函数）
  - 提取每年TF-IDF值最高的词汇

2005	发展 建设 经济 社会 政府	2015	发展 建设 改革 经济 社会
2006	发展 建设 经济 农村 社会	2016	发展 建设 改革 经济 社会
2007	发展 建设 经济 社会 农村	2017	发展 建设 企业 改革 经济
2008	发展 建设 社会 政府 经济	2018	发展 建设 改革 社会 时代
2009	发展 建设 经济 企业 社会	2019	发展 建设 经济 企业 改革
2010	发展 经济 建设 社会 重点	2020	发展 建设 高质量 时代 着力
2011	发展 建设 经济 社会 基本	2021	发展 建设 疫情 安全 现代化
2012	发展 建设 经济 社会 文化	2022	发展 建设 国家 高质量 疫情
2013	发展 建设 社会 文化 经济	2023	建设 发展 国家 高质量 现代化
2014	发展 建设 改革 经济 社会	2024	发展 建设 高质量 国家 时代

# 关键词分析

- 根据TF-IDF算法，提取每年政府工作报告中的关键词
  - 生成语料库的TF-IDF矩阵（sklearn中TfidfVectorizer）
  - 提取每年TF-IDF值最高的词汇
  - 剔除“发展”、“建设”，生成词云图



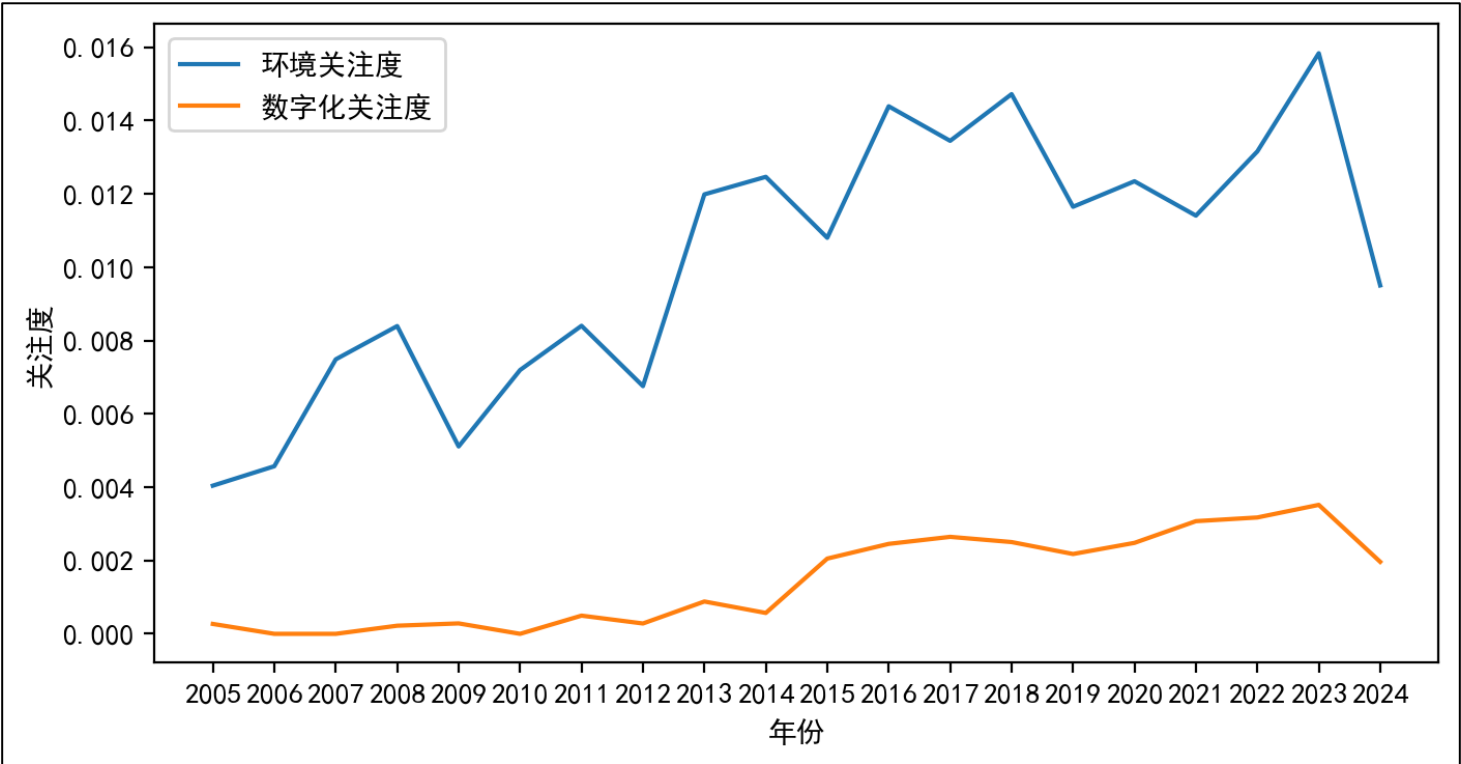


# 政府关注度分析

- 指标构建方法
  - 主题X：关键词x1，关键词x2，...，关键词xn
  - 对主题X的关注度： $(x1出现次数 + x2出现次数 + \dots + xn出现次数) / 词汇总数$
- 关注度指标
  - 数字化关键词：数字经济、数据安全、智慧化、智能经济、数据服务、信息经济、互联网、云平台、智能制造、物联网、数字化、数字产业、信息技术、区块链、数字基础设施、大数据、数据治理、云计算、数字生活、智能化、人工智能、数据共享、机器人、工业互联网、数字技术  
《数字经济政策何以影响企业数字技术创新》 (周剑明, 2024)
  - 环境关键词：生态、污染、生态环境、节能、新能源、绿化、环境保护、环保、绿色、可持续发展、排放、环境治理  
《环境质量因地方政府的重视得到改善吗? 》 (陈安琪、李永友, 2021)

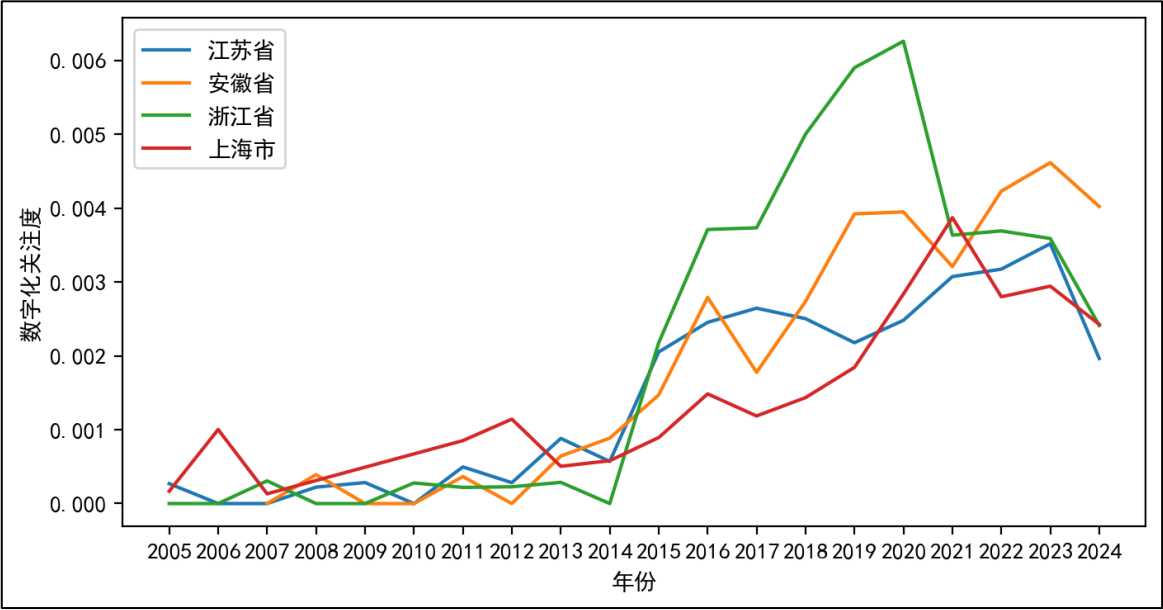
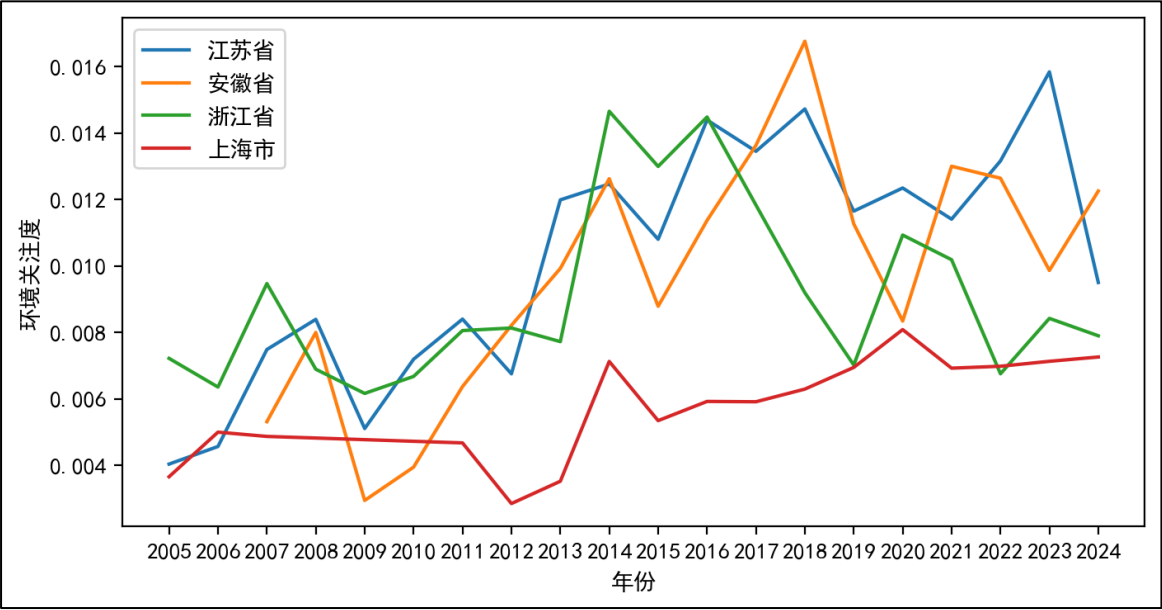
# 政府关注度分析

- 江苏省政府环境关注度与数字化关注度



# 政府关注度分析

- 长三角地区政府关注度的对比



谢谢