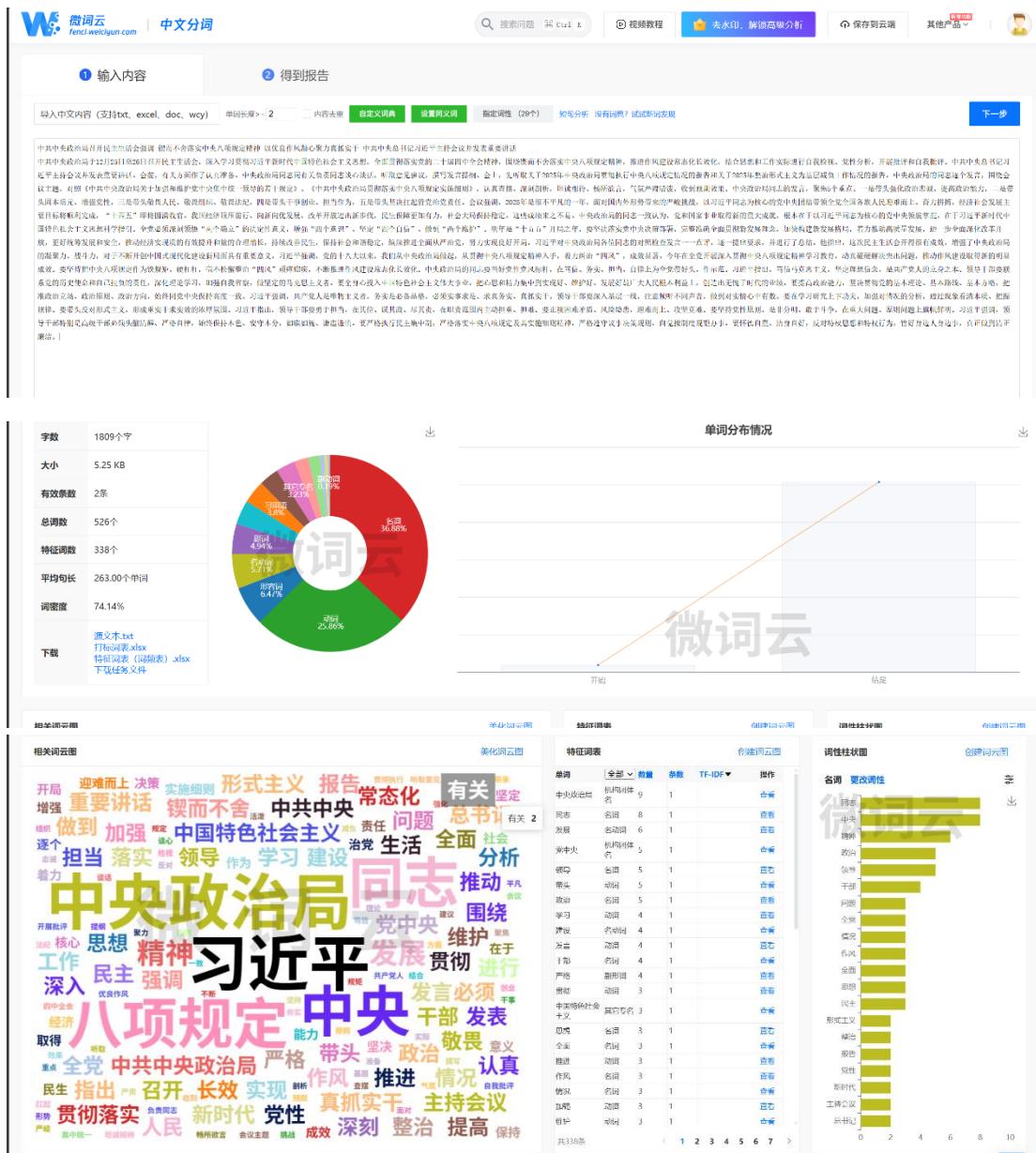
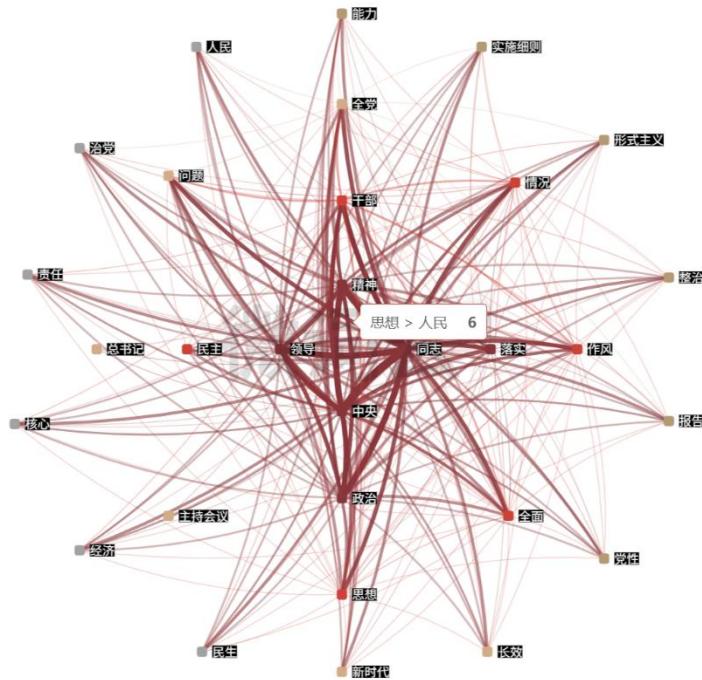


第一讲:课程导演与分词

作业 1:





	习近平	同志	精神	领导	落实	全面	干部	思想	全面	作风	情况	全面	问题	白书记	主持会议	新时代	长治	民主	报告	政治	新时代	落实细则	形式主义	民主	治党	责任	核心	经济	民生
习近平	14	16	10	10	10	6	8	4	6	6	6	6	2	2	4	4	4	4	4	4	4	4	4	4	4	4	4	4	
经济	14	16	10	10	10	6	8	4	6	6	6	6	2	2	4	4	4	4	4	4	4	4	4	4	4	4	4	4	
核心	14	16	10	10	10	6	8	4	6	6	6	6	2	2	4	4	4	4	4	4	4	4	4	4	4	4	4	4	
责任	14	16	10	10	10	6	8	4	6	6	6	6	2	2	4	4	4	4	4	4	4	4	4	4	4	4	4	4	
人民	14	16	10	10	10	6	8	4	6	6	6	6	2	2	4	4	4	4	4	4	4	4	4	4	4	4	4	4	
能力	14	16	10	10	10	6	8	4	6	6	6	6	2	2	4	4	4	4	4	4	4	4	4	4	4	4	4	4	
实施细则	14	16	10	10	10	6	8	4	6	6	6	6	2	2	4	4	4	4	4	4	4	4	4	4	4	4	4	4	
形式主义	14	16	10	10	10	6	8	4	6	6	6	6	2	2	4	4	4	4	4	4	4	4	4	4	4	4	4	4	
整治	14	16	10	10	10	6	8	4	6	6	6	6	2	2	4	4	4	4	4	4	4	4	4	4	4	4	4	4	
监督	14	16	10	10	10	6	8	4	6	6	6	6	2	2	4	4	4	4	4	4	4	4	4	4	4	4	4	4	
党风	14	16	10	10	10	6	8	4	6	6	6	6	2	2	4	4	4	4	4	4	4	4	4	4	4	4	4	4	
长效	14	16	10	10	10	6	8	4	6	6	6	6	2	2	4	4	4	4	4	4	4	4	4	4	4	4	4	4	
新风	14	16	10	10	10	6	8	4	6	6	6	6	2	2	4	4	4	4	4	4	4	4	4	4	4	4	4	4	
新风会议	14	16	10	10	10	6	8	4	6	6	6	6	2	2	4	4	4	4	4	4	4	4	4	4	4	4	4	4	
新时代	14	16	10	10	10	6	8	4	6	6	6	6	2	2	4	4	4	4	4	4	4	4	4	4	4	4	4	4	
主持会议	8	8	6	5	5	4	4	3	3	3	3	3	3	3	3	2	2	2	2	2	2	2	2	2	2	2	2	2	2
总书记	8	8	6	5	5	4	4	3	3	3	3	3	3	3	3	2	2	2	2	2	2	2	2	2	2	2	2	2	2
问题	21	24	15	15	15	9	12	6	9	9	9	9	9	9	9	3	3	6	6	6	6	6	6	6	6	6	6	6	6
金党	21	24	15	15	15	9	12	6	9	9	9	9	9	9	9	3	3	6	6	6	6	6	6	6	6	6	6	6	6
作风建设	21	24	15	15	15	9	12	6	9	9	9	9	9	9	9	3	3	6	6	6	6	6	6	6	6	6	6	6	6
作风	21	24	15	15	15	9	12	6	9	9	9	9	9	9	9	3	3	6	6	6	6	6	6	6	6	6	6	6	6
全面	21	24	15	15	15	9	12	6	9	9	9	9	9	9	9	3	3	6	6	6	6	6	6	6	6	6	6	6	6
全面思想	21	24	15	15	15	9	12	6	9	9	9	9	9	9	9	3	3	6	6	6	6	6	6	6	6	6	6	6	6
民主	15	16	11	10	10	7	8	6	6	6	6	6	6	6	6	3	3	4	4	4	4	4	4	4	4	4	4	4	4
干部	28	32	20	20	20	12	8	12	12	12	12	12	4	4	4	8	8	8	8	8	8	8	8	8	8	8	8	8	
落实	22	24	16	15	15	12	7	9	9	9	9	9	4	4	4	6	6	6	6	6	6	6	6	6	6	6	6	6	
政治	35	40	25	25	25	25	20	10	15	15	15	15	5	5	5	10	10	10	10	10	10	10	10	10	10	10	10	10	
领导	35	40	25	25	25	25	20	10	15	15	15	15	5	5	5	10	10	10	10	10	10	10	10	10	10	10	10	10	
精神	36	40	25	25	25	20	11	15	15	15	15	15	6	6	10	10	10	10	10	10	10	10	10	10	10	10	10	10	
同志	56	40	40	40	40	24	32	16	24	24	24	24	8	8	8	16	16	16	16	16	16	16	16	16	16	16	16	16	
中央	56	36	35	35	22	28	15	21	21	21	21	21	8	8	8	14	14	14	14	14	14	14	14	14	14	14	14	14	

作业 2:

```
[1]: print("Hello, World. Hello gk")
Hello, World. Hello gk
```

作业 3-001:

代码 1:

▼ 1. 基本分词

```
[26]: import jieba  
[46]: seg_list1 = jieba.cut("黄旭华, 1926年3月12日出生于广东省汕尾市, 原籍广东省揭阳市。1949年毕业于上海交通大学。历任北京海军核潜艇研究室副总工程师、中船重工集团公司核潜艇总工程师等职务。1994年当选为中国工程院院士。")  
[47]: print("\n".join(seg_list1))  
黄旭华$,$1926$年$3$月$12$日出$生$于$广$东$省$汕$尾$市$,$原籍$广$东$省$揭$阳$市$,$1949$年$毕$业$于$上$海$交$通$大$学$,$历$任$北$京$海$军$核$潜$艇$研$究$室$副$总$工$程$师$,$中$船$重$工$集$团$公$司$核$潜$艇$总$体$研$究$设$计$所$研$究$员$,$名$誉$所$长$,$1994$年$当$选$为$中$国$工$程$院$院$士$。
```

分得好的地方

人名、地名、机构名等专有名词：

“黄旭华”“广东省”“汕尾市”“揭阳市”“上海交通大学”“北京”“海军”“核潜艇研究室”“中船重工集团公司”“核潜艇总体研究设计所”“中国工程院”等专有名词都分得很准确，没有出现错误拆分或合并的情况。这些专有名词是文本中的关键信息，正确分词有助于明确其指代内容。

日期和时间表达：

“1926 年 3 月 12 日”“1949 年”“1994 年”等日期表达被正确分词，年份、月份和日期都被合理地分隔开来，符合常见的日期表达习惯，有助于理解时间信息。

常见的词语和短语：

“出生于”“毕业于”“历任”“当选为”等常见的词语和短语也分得较好，这些词语在文本中起到关键的连接和表意作用，正确分词有助于把握句子的结构和语义。

标点符号的处理：

标点符号（如逗号、句号）被正确地分隔出来，没有与文本内容混淆，有助于清晰地划分句子结构。

分得不好的地方

职务名称的拆分：“副\$总工程师”被拆分成了“副”和“总工程师”，这是不合理的。职务名称“副总工程师”应该作为一个整体来分词，因为“副”是修饰“总工程师”的，拆分后会让人误解为“副”和“总工程师”是两个不同的概念或职务。

部分词语的拆分：“出 所长”被拆分成了“名誉”和“所长”，虽然“名誉”和“所长”可以单独作为词语，但在这里它们共同组成一个短语“名誉所长”，应该作为一个整体来分词，这样更符合语言习惯和语义表达的准确性。

数字和单位的组合：“3\$月 日”中的“3 月 12 日”被正确分词，但“3 日”之间的拆分可以更紧凑一些。虽然这种分词没有错误，但更紧凑的分词（如“3 月 12 日”作为一个整体）会更符合常见的日期表达习惯。

代码 3

```
2.加入词典，是针对第二个片段的，希望是能够完整把“长短期记忆网络”这个术语整体分割出来
[31]: jieba.load_userdict('dict.txt')

[32]: seg_list_dict = jieba.cut("LSTM (Long Short-Term Memory) 是长短期记忆网络，是一种时间递归神经网络，适合于处理和预测时间序列中间隔和延迟相对较长的重要事件。")
[33]: print('/'.join(seg_list_dict))
LSTM/ (/Long/ /Short-/Term/ /Memory/) /是/长短期记忆网络/，/是/一种/时间递归神经网络/，/适合/于/处理/和/预测/时间/序列/中/间隔/和/延迟/相对/较长/的/重要/事件/。

3.加入停用词，针对第一个片段，希望的结果是，结果中不会出现“的”、“等虚词”
[34]: stopwords = [line.strip() for line in open('stop_words.txt', 'r', encoding='utf-8').readlines()]

[41]: seg_list_stopw = jieba.cut("黄旭华，1926年3月12日出生于广东省汕尾市，原籍广东省揭阳市。1949年毕业于上海交通大学。历任北京海军核潜艇研究室副总工程师、中船重工集团公司核
[42]: final = ''
[43]: #这是一行注释，进行分词结果的过滤
for seg in seg_list_stopw:
    if seg not in stopwords:
        final += seg + '/'

[44]: print(final)
黄旭华/1926/年/3/月/12/日出/生于/广东省/汕尾市/原籍/广东省/揭阳市/1949/年/毕业/于/上海交通大学/历任/北京/海军/核潜艇/研究室/副/总工程师/中船重工集团公司/核潜艇/总体/研
究/设计所/研究员/名誉/所长/1994/年/当选/为/中国工程院院士/
```

分得好的地方

人名、地名、日期、机构名等专有名词：如“黄旭华”“广东省”“汕尾市”“揭阳市”“上海交通大学”“北京”“海军”“核潜艇研究室”“中船重工集团公司”“核潜艇总体研究设计所”“中国工程院”等专有名词都分得很准确，没有出现错误拆分或合并的情况，这有助于明确这些特定概念的边界，方便理解其在文本中的指代含义。

一些常见的词语和短语：像“出生于”“毕业于”“历任”“当选为”等常见的词语和短语也分得较好，这些词语在文本中起到关键的连接和表意作用，正确分词有助于把握句子的结构和语义。

分得不好的地方

数字和年份的分词：将“1926”“1949”“1994”等年份拆分成了单个数字，这是不合理的。年份应该作为一个整体来分词，因为它们在文本中表示一个完整的时间概念，拆分后会破坏其原有的意义和完整性。

部分短语的拆分：例如“原籍广东省揭阳市”被拆分成了“原籍/广东省/揭阳市”，虽然从字面上看这种分法似乎也说得过去，但从语义完整性和表达习惯上来说，更合理的分法应该是“原籍/广东省揭阳市”，因为“原籍”是一个限定词，后面应该紧跟着完整的籍贯信息，这样更符合语言习惯和语义表达的准确性。

“副/总工程师”：这种分法不太合理，应该合并为“副总工程师”，因为这是一个完整的职务名称，拆分后会让人误解为“副”和“总工程师”是两个不同的概念或职务，而实际上它们是连在一起表示一个特定的职位。

作业 3-002:

功勋科学家-黄旭华-传记文本分词

现在，可以开启你的小组项目的第一个小小任务啦！就是对一小段有关“功勋科学家”的文本进行分词处理。

```
[1]: # 简单分词
[2]: import jieba
[3]: seg_list_huang = jieba.cut('黄旭华, 1926年3月12日出生于广东省汕尾市, 原籍广东省揭阳市, 1949年毕业于上海交通大学。历任北京海军核潜艇研究室副总工程师、中船重工集团公司核
[4]: print('/'.join(seg_list_huang))

Building prefix dict from the default dictionary ...
Loading model from cache C:\Users\Lenovo\AppData\Local\Temp\jieba.cache
Loading model cost 1.481 seconds.
Prefix dict has been built successfully.

黄旭华/, 1926/年/3/月/12/日/出/生/于/广/东省/汕尾市/, /原籍/广/东省/揭阳市/, /1949/年/毕业/于/上/海交/通大/学/, /历任/北京/海/军/核/潜/艇/研/究室/副/总工/程师/、/中/船/重/工/集/团/公/司/核/潜/艇/总/体/研/究/设/计/所/研/究/员/、/名/誉/所/长/。/1994/年/当/选/为/中/国/工/程/院/院/士/。
```



```
[1]: print('/'.join(seg_list_huang))

黄旭华/, 1926/年/3/月/12/日/出/生/于/广/东省/汕尾市/, /原籍/广/东省/揭阳市/, /1949/年/毕业/于/上/海交/通大/学/, /历任/北京/海/军/核/潜/艇/研/究室/副/总工/程师/、/中/船/重/工/集/团/公/司/核/潜/艇/总/体/研/究/设/计/所/研/究/员/、/名/誉/所/长/。/1994/年/当/选/为/中/国/工/程/院/院/士/。
```



```
[1]: # 加入词典之后，哪些词汇被分出来了呢？
[2]: 
[3]: # 使用停用词表
[4]: 
[5]: stopwords = [line.strip() for line in open('stop_words.txt', 'r', encoding='utf-8').readlines()]
[6]: 
[7]: stopwords = open('stop_words.txt','r', encoding='utf-8').read()
[8]: stopwords = stopwords.split('\n')
[9]: 
[10]: stopwords
[11]: ['的', '了', '是', '啊', '、', '、', '。', '。', '停用']
[12]: 
[13]: seg_list_huang = jieba.cut('黄旭华, 1926年3月12日出生于广东省汕尾市, 原籍广东省揭阳市, 1949年毕业于上海交通大学。历任北京海军核潜艇研究室副总工程师、中船重工集团公司核
[14]: final = ''
[15]: 
[16]: for seg in seg_list_huang:
[17]:     if seg not in stopwords:
[18]:         final+= seg+'/'
[19]: 
[20]: print(final)
[21]: 黄旭华/1926/年/3/月/12/日/出/生/于/广/东省/汕尾市/原籍/广/东省/揭阳市/1949/年/毕业/于/上/海交/通大/学/历任/北京/海/军/核/潜/艇/研/究室/副/总工/程师/中/船/重/工/集/团/公/司/核/潜/艇/总/体/研/究/设/计/所/研/究/员/名/誉/所/长/1994/年/当/选/为/中/国/工/程/院/院/士/
```

作业 3-003:

```
[11]: # 输出所有词汇的词频（按频率降序）
print("\n所有词汇词频统计（前20个）：")
for word, count in word_counts.most_common(20):
    print(f'{word}: {count}次')
```

所有词汇词频统计（前20个）：

```
', ': 13次
':
': 9次
'管理': 5次
'、': 5次
'与': 4次
'“': 3次
'企业': 3次
'”': 3次
'体系': 3次
'智能化': 3次
'经营': 3次
'打造': 3次
'能力': 3次
'供应链': 3次
'建设': 2次
'通过': 2次
'自动化': 2次
'数字化': 2次
'升级': 2次
'生产': 2次
```

提问

- 1，以上仅仅是统计了文本中的词的频次，这能反映什么呢？
 - 2，词频统计本身存在什么问题呢？去停用词没有完成？
 - 3，仅仅统计词频有什么不足呢？能够确定“数字化”与“安全”之间的关系吗？那如何改进呢？
-

解答：

1. 词频统计能反映什么？

词频统计可以提供以下信息：

文本内容的重点：高频词汇往往反映了文本的主要主题和关注点。例如，在你的统计结果中，“管理”“企业”“体系”等词汇出现频率较高，这可能表明文本主要围绕企业管理、体系建设等内容展开。

文本的风格和领域：高频词汇可以帮助判断文本的风格和所属领域。例如，“数字化”“智能化”等词汇表明文本可能与科技、企业管理等领域相关。

文本的结构和语言特点：高频词汇中可能包含一些常见的连接词、标点符号等，这些可以帮助了解文本的结构和语言风格。例如，“，”“”“”等标点符号的高频出现表明文本中存在较多的列举和并列结构。

2. 词频统计本身存在什么问题？去停用词没有完成？

词频统计本身存在以下问题：

标点符号和停用词的干扰：在你的统计结果中，“，”“”“”等标点符号和

“与”“的”等常见连接词（停用词）的频率较高。这些词汇虽然在文本中出现频率高，但通常不具有实际的语义价值，会干扰对文本主题和重点的分析。

未完成停用词过滤：从你的输出结果来看，停用词（如“，”“与”等）没有被过滤掉。停用词过滤是文本预处理的重要步骤，可以去除这些不具有实际语义价值的词汇，从而更准确地反映文本的核心内容。

上下文信息缺失：词频统计只关注词汇的出现频率，而忽略了词汇之间的上下文关系。例如，虽然“数字化”和“安全”都出现在文本中，但无法确定它们之间的具体关系。

3. 仅仅统计词频有什么不足？如何改进？

仅仅统计词频的不足：

无法捕捉上下文关系：词频统计无法确定词汇之间的语义关系。例如，虽然“数字化”和“安全”都出现了2次，但无法确定它们是否在同一段落中出现，或者它们之间是否存在因果、并列等关系。

无法处理同义词和多义词：词频统计无法区分同义词和多义词。例如，“数字化”和“智能化”可能在某些情况下是同义词，但词频统计无法将它们合并处理。

无法捕捉语义结构：词频统计无法捕捉文本的语义结构，如句子、段落的逻辑关系等。

改进方法

为了克服这些不足，可以采用以下方法：

停用词过滤：在词频统计之前，去除常见的停用词（如“，”“与”“的”等），以减少干扰，更准确地反映文本的核心内容。

上下文分析：

共现分析：统计词汇之间的共现频率，即两个词汇在一定范围内（如同一句子或段落）同时出现的次数。例如，统计“数字化”和“安全”在同段落中出现的次数，可以初步判断它们之间的关联。

依存句法分析：通过依存句法分析，确定词汇之间的语法和语义关系。例如，使用自然语言处理工具（如 Stanford NLP、Spacy 等）分析句子结构，确定“数字化”和“安全”之间是否存在修饰、因果等关系。

语义分析：

词嵌入 (Word Embedding)：使用词嵌入技术（如 Word2Vec、GloVe 等）将词汇映射到高维向量空间，通过计算向量之间的相似度来判断词汇之间的语义关系。

主题建模: 使用主题建模技术（如 LDA）对文本进行分析，提取文本中的主题，确定“数字化”和“安全”是否属于同一主题。

深度学习方法：

BERT 等预训练模型: 使用 BERT 等预训练模型对文本进行分析，提取上下文信息，判断词汇之间的语义关系。BERT 模型能够捕捉词汇在上下文中的动态语义，适用于复杂的语义分析任务。

作业 3-004:

```
提取到的实体和专业术语:  
```json  
{
 "理论": [
 "肿瘤免疫微环境",
 "T细胞耗竭",
 "免疫编辑理论"
],
 "方法": [
 "单细胞RNA测序",
 "细胞亚群聚类",
 "轨迹分析",
 "pseudotime推断",
 "细胞间通讯网络构建"
],
 "工具": [
 "Seurat",
 "Monocle3",
 "CellChat"
],
 "专业术语": [
 "TIME",
 "scRNA-seq",
 "非小细胞肺癌",
 "PD-1/PD-L1",
 "TGF-β路径",
 "免疫抑制信号通路",
 "个体化免疫治疗"
]
}
```

解答:

#### 1. 关于使用 DeepSeek 开展工作的感觉:

我并不是一个实体工具，所以并没有“使用 DeepSeek 开展工作”的实际体验。不过从技术角度来看，DeepSeek 这类先进的工具通常能够提供强大的数据分析和处理能力，尤其是在生物医学领域，它可以帮助研究人员更好地理解复杂的生物系统，比如肿瘤免疫微环境 (TIME)。通过使用 DeepSeek，研究人员可以更高效地处理单细胞 RNA 测序 (scRNA-seq) 数据，从而更深入地研究细胞亚群的聚类、细胞间通讯

网络等，这对于理解非小细胞肺癌等疾病的机制非常有帮助。

## 2. 大语言模型的工作表现：

大语言模型（如我）在处理语言相关的任务上表现得相当出色。我可以帮助用户理解复杂的理论，比如肿瘤免疫微环境和免疫编辑理论。同时，我也可以辅助用户理解各种专业术语，例如 PD-1/PD-L1、TGF- $\beta$  路径等。在方法方面，我可以解释单细胞 RNA 测序、细胞间通讯网络构建等概念，帮助用户更好地应用这些方法。不过，大语言模型也有局限性，例如在处理非常具体的数据分析任务时，可能需要与专门的工具（如 Seurat、Monocle3、CellChat）结合使用，才能达到最佳效果。

## 3. 识别关系的能力：

大语言模型可以通过上下文理解来识别实体之间的关系。例如，从你提供的实体和术语中，我可以识别出以下关系：

**肿瘤免疫微环境 (TIME) 与非小细胞肺癌：** TIME 是非小细胞肺癌等肿瘤疾病的重要特征，它影响着肿瘤的进展和治疗反应。

**T 细胞耗竭与免疫抑制信号通路：** T 细胞耗竭是免疫抑制信号通路（如 PD-1/PD-L1、TGF- $\beta$  路径）过度激活的结果，这会导致 T 细胞功能下降，影响抗肿瘤免疫反应。

**单细胞 RNA 测序 (scRNA-seq) 与细胞亚群聚类：** scRNA-seq 是一种强大的方法，用于研究细胞亚群的特征和功能，通过聚类分析可以发现不同细胞亚群在肿瘤免疫微环境中的作用。

**细胞间通讯网络构建与免疫编辑理论：** 细胞间通讯网络的构建有助于理解免疫细胞与肿瘤细胞之间的相互作用，这与免疫编辑理论密切相关，因为免疫编辑理论描述了肿瘤如何逃避免疫系统的监视。

**Seurat、Monocle3、CellChat 与单细胞 RNA 测序：** 这些工具（Seurat 用于细胞聚类和数据整合，Monocle3 用于轨迹分析和 pseudotime 推断，CellChat 用于细胞间通讯网络构建）是单细胞 RNA 测序数据分析的关键工具，它们帮助研究人员从不同角度分析和理解细胞行为和相互作用。

## 作业 4：

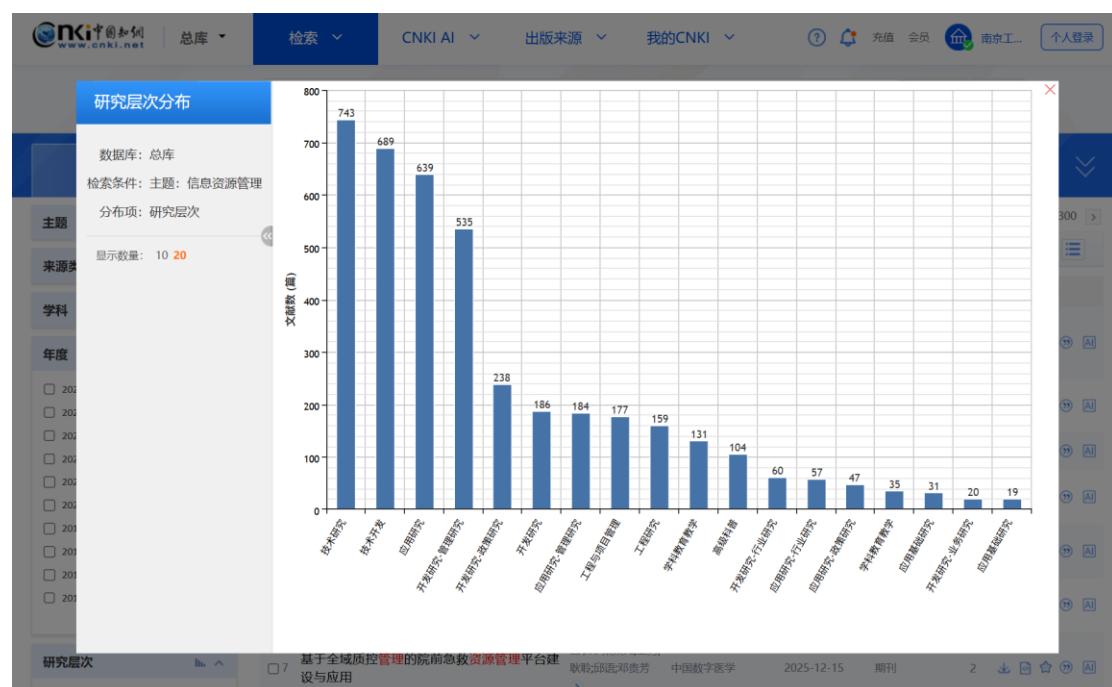
随着科技文献数量的快速增长，研究方法作为文献中的核心内容，其快速准确的提取对于科研工作者具有重要意义。研究方法句的汇总可以帮助科研人员快速搜寻合适的研究方法，提升研究效率。然而，现有方法多集中于从摘要中提取，忽略了全文中更

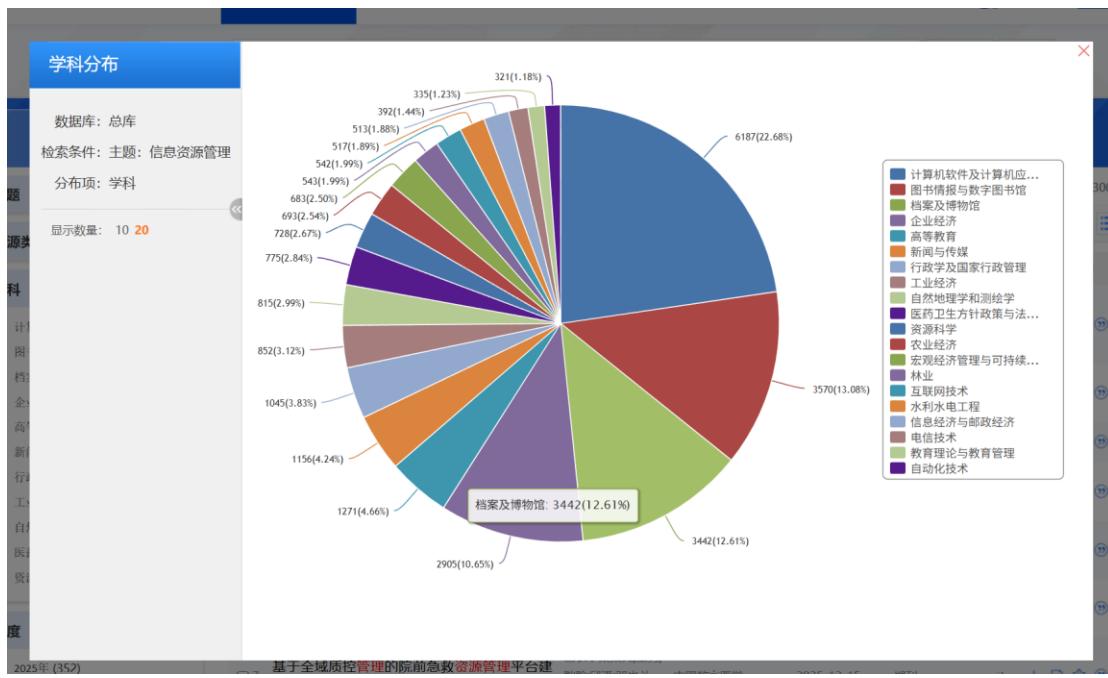
详细的研究方法描述。因此，本文提出从学术论文全文中自动抽取研究方法句的方法，并分析其在情报学领域的应用。

本文提出的基于 BERT 的单层次 BiLSTM 模型能够有效从学术论文全文中抽取研究方法句，并揭示了《情报学报》在情报学理论研究方面的重视。未来工作将引入半监督学习以减少对标注数据的依赖，并扩大研究范围至更多情报学期刊论文。

## 第二讲:词频统计

### 作业 1:





## 作业 2 和 3:

## 代码 1:

```
[38]: # 输出词频的前n个
for i in range(100):
 print(articlelist[i])

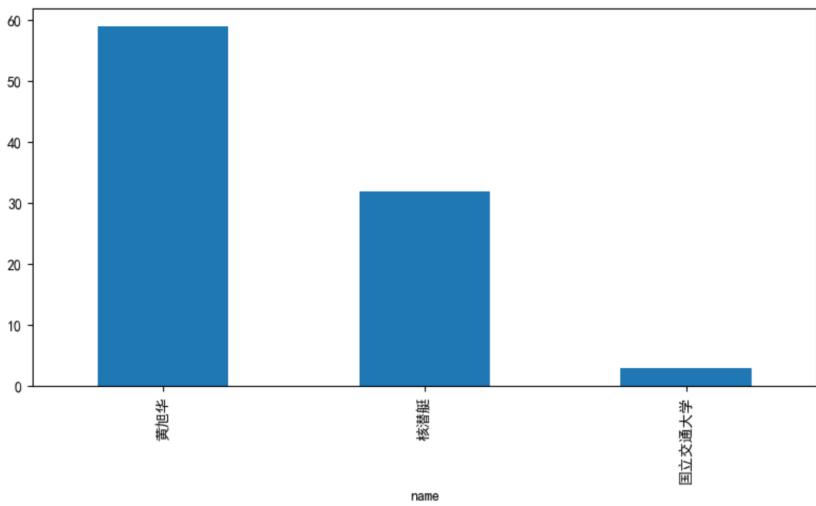
('黄旭华', 53)
('核潜艇', 32)
('采集', 29)
('学术', 22)
('资料', 21)
('工作', 17)
('成长', 15)
('小组', 14)
('进行', 13)
('专业', 13)
('院士', 13)
('我国', 12)
('技术', 12)
('研制', 12)
('工程', 11)
('访谈', 10)
('主要', 8)
('介绍', 8)
('科学', 8)
('第一代', 8)
('人生', 7)
('及其', 7)
('思想', 7)
('传记', 7)
('历史', 7)
('要求', 6)
('过程', 6)
('一生', 6)
('设计', 6)
('成就', 6)
('传主', 6)
('研究', 6)
```

```
('求学', 5)
('实现', 5)
('精神', 5)
('重点', 5)
('实物', 5)
('完成', 4)
('时间', 4)
('照片', 4)
('描述', 4)
('完整', 4)
('事件', 4)
('学习', 4)
('反映', 4)
('船舶', 4)
('任务', 4)
('保密', 4)
('重要', 4)
('依据', 4)
('成熟', 4)
('轨迹', 4)
('09', 4)
('其中', 4)
('中国', 4)
('系统', 4)
('还原', 4)
('形成', 3)
('制造', 3)
('关于', 3)
('回顾', 3)
('客观', 3)
('大学', 3)
('各种', 3)
('生活', 3)
('国家', 3)
('参加', 3)
('直接', 3)
('档案', 3)
('规定', 3)
('背景', 3)
('本章', 3)
('突破', 3)
```

## 代码 2

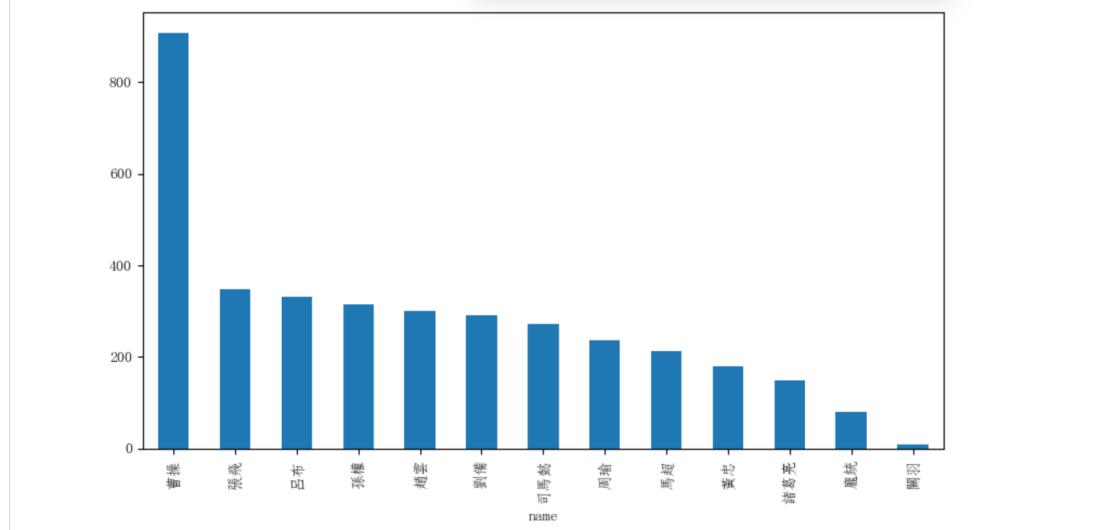
```
[14]: articlelist = sorted(articleDict.items(),key = lambda x:x[1], reverse = True) # 对词典中的词排序
[15]: # 输出词频的前N个
for i in range(100):
 print(articlelist[i])
('董卓', 97)
('吕布', 60)
('曹操', 59)
('袁绍', 57)
('天下', 53)
('玄德', 48)
('貂蝉', 37)
('太守', 36)
('朝廷', 32)
('孙坚', 31)
('不可', 31)
('次日', 26)
('引兵', 25)
('商议', 25)
('李儒', 25)
('天子', 24)
('左右', 23)
('太师', 22)
('玄德曰', 22)
('军士', 21)
('大喜', 21)
('校尉', 21)
('今日', 21)
('太后', 20)
('何进', 20)
('王允', 20)
('二人', 19)
('司徒', 18)
('郭汜', 18)
('公孙瓒', 18)
('不能', 18)
('诸侯', 18)
('朱俊', 17)
```

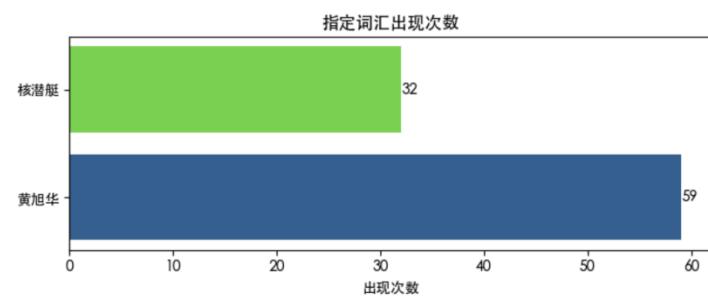
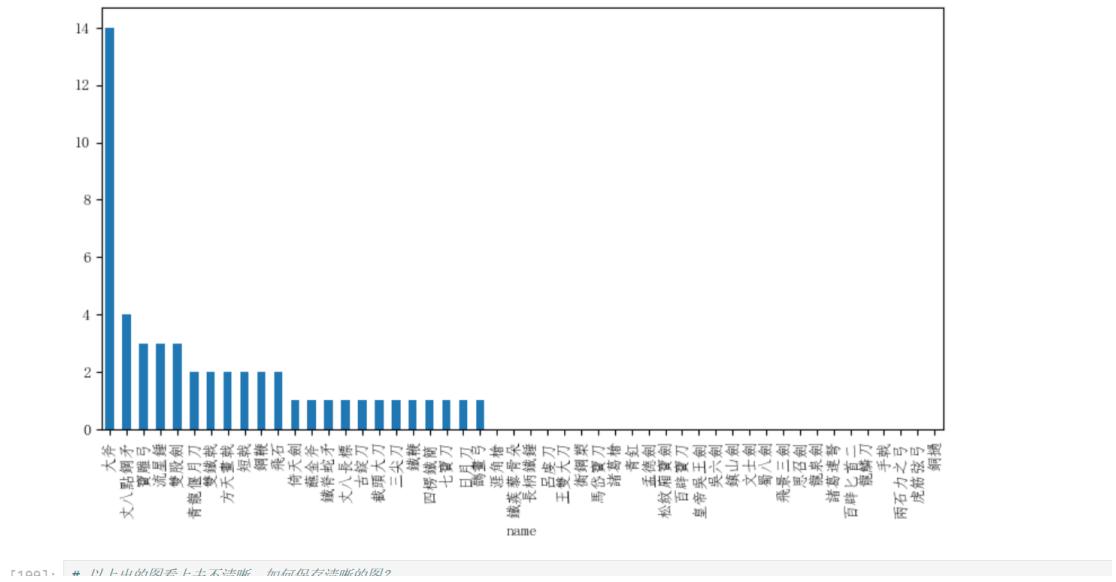
## 代码 3:



代码 4:

```
[186]: draw_dict(name_dict)
```





## 作业 4:

## 研究背景

文章旨在通过分析全球数字化的书籍和文章，来考察物理学领域最具影响力科学家的科学声誉。科学研究不仅推动了人类知识的进步，也对社会产生了深远的影响。然而，传统上对科学家贡献的评估主要集中在学术圈内的引用等指标上，而对于科学家在更广泛社会中的影响力（即科学声誉）的研究相对较少。随着谷歌图书（Google Books）和谷歌学术（Google Scholar）等数字化资源的出现，为研究科学家的长期社会影响提供了新的可能性。

研究方法

**数据来源：**研究主要基于谷歌图书和谷歌学术的数据，涵盖了超过 3600 万本全球数字化书籍和 9100 万篇学术文章，时间跨度从 16 世纪到 21 世纪。

**研究对象选择：**文章以艾萨克·牛顿（Isaac Newton）和阿尔伯特·爱因斯坦（Albert Einstein）为例，这两位科学家都是物理学领域的杰出代表，且在科学史上具有极高的声誉，他们的比较能够揭示科学声誉的演变和差异。

**测量科学声誉：**通过统计书籍和文章中提到科学家全名的频率来衡量其科学声誉。此外，还利用谷歌图书的 Ngram Viewer 工具，分析不同语言的书籍中对

这两位科学家的提及情况，以探讨“本群体偏好”（own-group preference）现象，即科学家在其母国或使用相同语言的学者中可能享有更高的声誉。

**共现分析：**为了深入了解牛顿和爱因斯坦最著名的科学成就对其声誉的贡献，文章还进行了共现分析，统计了与他们名字同时出现的特定科学概念（如牛顿的万有引力定律、爱因斯坦的相对论等）的频率。

## 研究发现

**科学家声誉的持久性：**研究表明，伟大的科学家虽然已经去世，但他们对人类历史的科学影响却持续了数百年。牛顿和爱因斯坦等科学家的名字在现代社会中仍然被广泛提及，他们的科学贡献成为了人类知识体系的基石。

**本群体偏好现象：**研究发现存在本群体偏好现象。例如，在英国英语的书籍中，牛顿的声誉一直高于爱因斯坦，这与 2005 年英国皇家学会的民意调查结果一致；而在美国英语和德语的书籍中，爱因斯坦自 20 世纪中后期以来被提及的频率远高于牛顿。这表明科学家的声誉在一定程度上受到其所属群体（如国家、语言群体）的影响。

**科学家声誉的演变：**文章通过分析牛顿和爱因斯坦的声誉随时间的变化趋势发现，牛顿在 1680 年至 1880 年间达到了其历史影响力的高峰，而爱因斯坦的声誉则在 20 世纪中后期逐渐超越了牛顿。这种变化可能与物理学领域的重大理论突破（如量子理论和相对论）以及两次世界大战对科学的影响有关。

**科学家最著名的成就：**共现分析显示，牛顿的声誉与其万有引力定律和运动定律密切相关，而爱因斯坦的声誉则主要与相对论和量子理论相关。这表明科学家的某些特定成就对其长期声誉有着显著的贡献，但同时也存在其他因素（如科学精神、轶事等）对其声誉的影响。

## 研究意义

**拓展科学声誉研究：**这是首次尝试结合谷歌图书和谷歌学术的数据来探索科学家的科学声誉，为评估科学家的社会影响力提供了新的视角和方法。

**揭示科学家声誉的复杂性：**研究不仅揭示了科学家声誉的持久性和本群体偏好现象，还通过共现分析深入了解了科学家声誉与其科学成就之间的关系，丰富了对科学家声誉形成机制的认识。

**为科学传播提供启示：**通过比较科学家在学术界和公众中的声誉，文章为科学传播和科学教育提供了启示，强调了科学家在促进科学知识传播和公众科学素养提升方面的重要作用。

## 研究局限性

**数据来源的局限性：**谷歌图书和谷歌学术的数据存在一定的局限性，如语言偏

差（非英语书籍的代表性不足）、书籍质量的异质性以及对科学家声誉测量的适用性可能仅限于具有重大影响力的科学家。

**测量方法的局限性：**在测量科学家声誉时，可能会出现虚假正例（如将科学家自己撰写的书籍计入其声誉）和虚假负例（如只提及科学家的成就而不提其名字）的情况，这可能会影响研究结果的准确性。

未来研究方向

**进一步区分科学家声誉的来源：**未来研究可以进一步区分科学家在学术界和公众中的声誉差异，以及不同类型的书籍（如教科书、小说、学术著作等）对科学家声誉的影响。

**跨学科研**：将研究范围扩展到其他学科领域，探讨不同学科中科学家声誉的差异和特点，以及跨学科研对科学家声誉的影响。

**长期声誉与短期引用的对比：**对比科学家的长期声誉（通过谷歌图书等数据衡量）与短期学术引用之间的关系，深入理解科学家贡献的长期影响和短期学术认可之间的差异。

**社交媒体等新数据源的应用：**随着社交媒体等新兴数据源的发展，未来可以探索利用这些数据源来评估科学家的声誉，以更全面地了解科学家在当代社会中的影响力。

### 第三讲:词云与可视化

作业 1

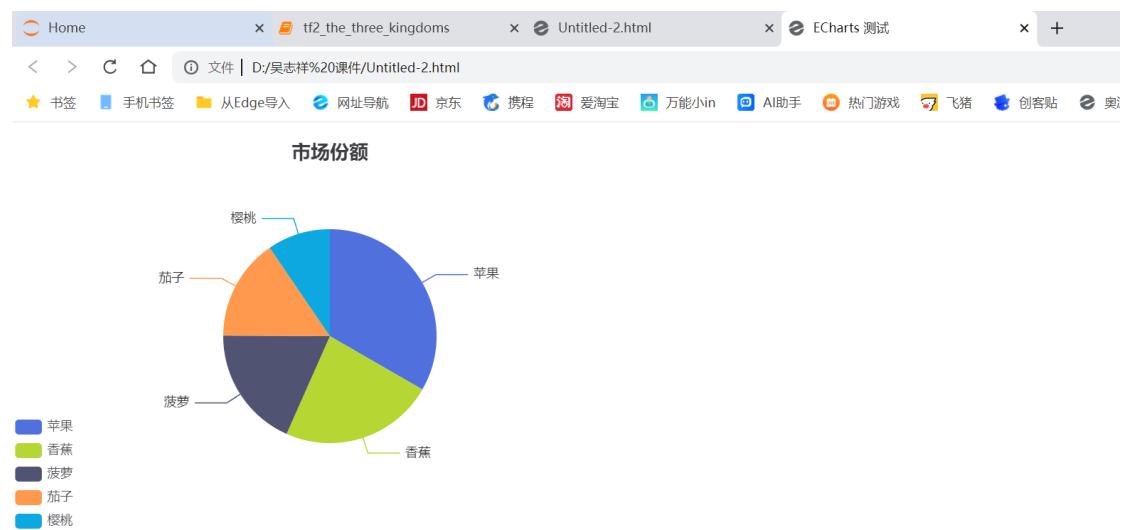
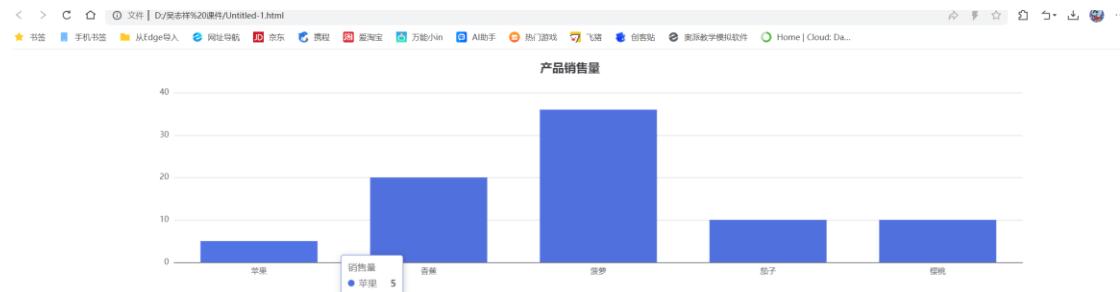


这张词云图以直观的方式突出了文本中的关键词汇，其中“中央政治局”、“习近平”和“八项规定”等词汇的字体较大，显示它们在文本中的重要性和高频出现。这些词汇主要围绕中国共产党的领导、政策和党的建设等主题，反映了文本的核心内容和关注点。

**中央政治局和习近平**: 字体最大, 表明这些词汇在文本中最为重要或出现频率最高。  
中央政治局是中国共产党的核心决策机构, 习近平是中国的国家领导人。

**八项规定：**这是中国共产党在 2012 年提出的规定，旨在改进党的作风，反对形式主义、官僚主义、享乐主义和奢靡之风

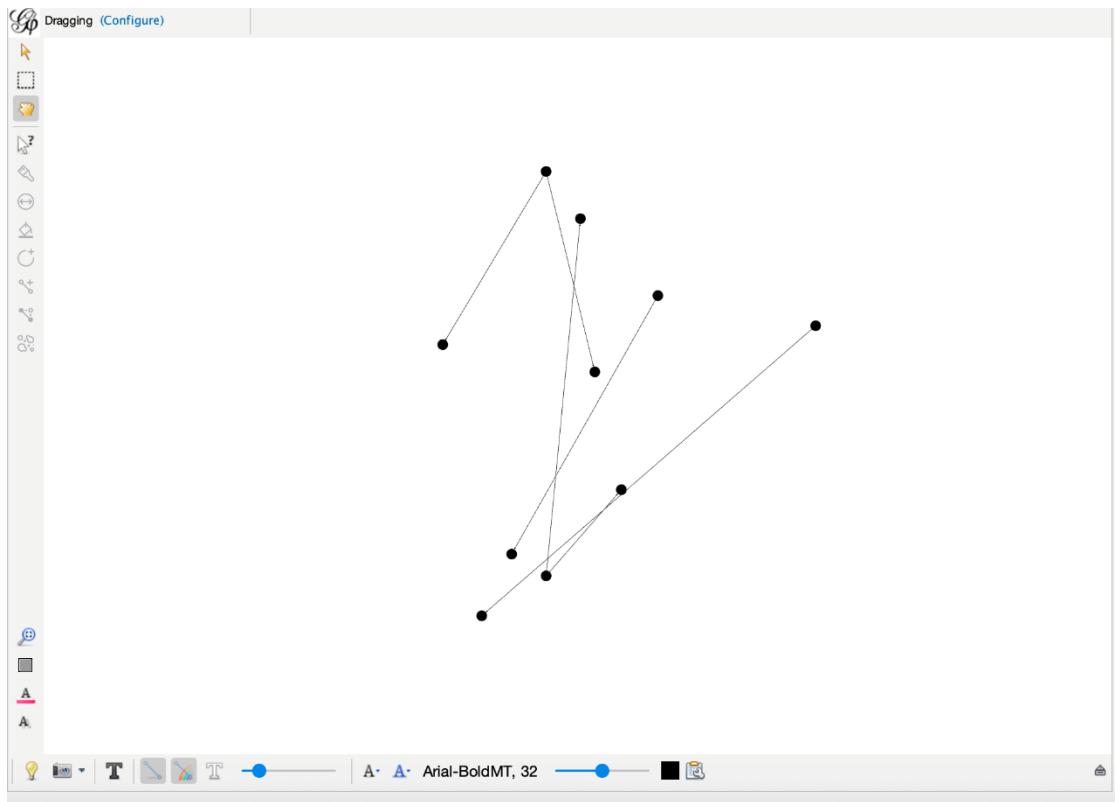
## 作业 2：



## 社交网络关系图



作业 3:



作业 4:



第4讲:情感分析

作业 1:

## 情感分析

请输入一段中文文本：

同志为核心的党中央团结带领全党全国各族人民迎难而上、奋力拼搏，经济社会发展主要目标将顺利完成，“十四五”即将圆满收官。我国经济顶住前行、向新向优发展的改革步伐，民生保障更加有力，社会大局保持稳定。这些成绩来之不易。中央政治局的同志一致认为，党和国家事业取得新的重大成就，根本在于以习近平同志为核心的党中央领航掌舵，在于习近平新时代中国特色社会主义思想科学指引。全党必须深刻领悟“两个确立”的决定性意义，增强“四个意识”、坚定“四个自信”、做到“两个维护”。明年是“十五五”开局之年，要坚决落实党中央决策部署，完整准确全面贯彻新发展理念，加快构建新发展格局，着力推动高质量发展，进一步全面深化改革开放，更好统筹发展和安全，推动经济实现质的有效提升和量的合理增长，持续改善民生，保持社会和谐稳定，纵深推进全面从严治党，努力实现良好开局。习近平对中央政治局各位同志的对照检查发言一一点评、逐一提出要求，并进行了总结。他指出，这次民主生活会开得很有成效，增强了中央政治局的凝聚力、战斗力，对于不断开创中国式现代化建设新局面具有重要意义。习近平强调，党的十八大以来，我们从中央政治局做起，从责

1000/1000

## 此页内容

[简介](#)  
[调用方法](#)  
[创建客户端](#)  
[情感分析](#)  
[本地调用](#)  
[多语种支持](#)

情感分析

## 情感极性



## 作业二：

## 代码 1：

```
例如
text_taobao_1 = "显示效果：挺好的 运行速度：目前来说很流畅 拍照效果：拍照效果挺好的 电池续航：一天一冲 总结：目前没啥毛病，用了
taobao_1 = SnowNLP(text_taobao_1)
taobao_1.sentiments
0.999947261146611

text_taobao_2 = "总结：这是我买过最不满意的一款手机！两千多元的手机这样，真的很不值！"
taobao_2 = SnowNLP(text_taobao_2)
for sentence in taobao_2.sentences:
 print(sentence)

总结：这是我买过最不满意的一款手机
两千多元的手机这样
真的很不值

taobao_2.sentiments
0.889005139666256
```

```

[1] 给不了语音，得重新开机后才能发，才买半个月的手机就这样，客服态度也很差！拍照效果：拍照不清晰！电池续航：手机不蓄电，充满电后用半
[2] taobao_3 = SnowNLP(text_taobao_3)
[3] taobao_3.sentiments
[4] 5.707609422256343e-05

```

## 代码 2：



```
[18]: plt.savefig('timeline.png') # 看不到? 改一改?
```

在图中，我们发现许多正面评价情感分析数值极高的。同时，我们也清晰地发现了那几个数值极低的点。对应评论的情感分析数值接近于0。这几条评论，被Python判定为基本上没有正面情感了。

从时间上看，最近一段时间，几乎每隔几天就会出现一次比较严重的负面评价。

作为经理，你可能如坐针毡。希望尽快了解发生了什么事儿。你不用在数据框或者Excel文件里面一条条翻找情感数值最低的评论。Python数据框Pandas为你提供了非常好的排序功能。假设你希望找到所有评论里情感分析数值最低的那条，可以这样执行：

```
[19]: df.sort_values(['sentiments'])[:1]
```



情感分析结果数值几乎就是0啊！不过这里数据框显示评论信息不完全。我们需要将评论整体打印出来。

```
[20]: print(df.sort_values(['sentiments']).iloc[0].comments)
```

这次是在情人节当天过去的，以前从来没在情人节正日子出来过，不是因为没有男朋友，而是感觉哪哪都多，所以特意错开，这次实在是傍A餐厅了，所以赶在正日子也出来了，从下午四点多的时候我看排号就排到一百多了，我从家开始过去得堵的话一个小时，我一看提前两个小时就在网上先排着号了，差不多我们是六点半到的，到那的时候我看号码前面还有才三十多号，我想着肯定没问题了，等一会就能吃上的，没想到悲剧了，就从我们到那坐到等位区开始，大约是十分二十分一叫号，中途多次我都想走了，哈哈，哎，等到最后早上九点才吃上的。服务员感觉也没以前准时时候到了，不过这肯定的，一人负责好几桌，今天节日这么多人，肯定是很累的，所以大多都是我自己跑腿，没让服务员给弄太多，就赶紧让服务员下的，然后环境来说感觉卫生方面是不错的，就是有些太吵了，味道还是一如既往的那个味道，不过A餐厅最人性化的就是看我们等了两个多小时，上来送了我们一张打折卡，而且当次就可以使用，这点感觉还是挺好的，不愧是A餐厅，就是比一般的要人性化。不过这次是选错日子了，以后还是得提前预约，要不就别赶节日去，太火爆了！

### 代码 3:

```
细粒度情感实体抽取结果:
{
 "实体": [
 {
 "部位": "全身",
 "症状": "失眠",
 "情感": "失去重心"
 },
 {
 "部位": "头部",
 "症状": "头痛",
 "情感": "疲乏无力"
 },
 {
 "部位": "皮肤",
 "症状": "异常敏感, 触碰如针扎疼痛",
 "情感": "不适"
 },
 {
 "部位": "心脏",
 "症状": "心慌",
 "情感": "惊恐"
 },
 {
 "部位": "胸部",
 "症状": "胸闷",
 "情感": "沉重"
 },
 {
 "部位": "背部",
 "症状": "沉重如压石头",
 "情感": "压抑"
 },
 {
 "部位": "感官",
 "症状": "对光线和声音极度敏感",
 "情感": "惊恐"
 },
 {
 "部位": "行为",
 "症状": "",
 "情感": ""
 }
]
}
```

```
},
{
 "部位": "胸部",
 "症状": "胸闷",
 "情感": "沉重"
},
{
 "部位": "背部",
 "症状": "沉重如压石头",
 "情感": "压抑"
},
{
 "部位": "感官",
 "症状": "对光线和声音极度敏感",
 "情感": "惊恐"
},
{
 "部位": "行为",
 "症状": "不愿出门，不与人交流，关在屋里拉紧窗帘",
 "情感": "生活毫无意义"
}
]
```

## 提问：

- 1. 请问你觉得大语言模型在识别患者的身体、心理、情感的工作中，表现如何？
- 2. 除了健康领域的情感世界理解与识别，你还能够想到哪些其他重要的领域，可以开展类似的工作呢？

解答：. 请问你觉得大语言模型在识别患者的身体、心理、情感的工作中，表现如何？

大语言模型在处理和识别患者的身体、心理和情感症状方面表现出色，主要体现在以下几个方面：

**自然语言理解**：大语言模型能够理解和解析自然语言文本，从而识别出文本中描述的症状和情感状态。

**上下文关联**：通过分析上下文，模型能够更准确地理解症状和情感的表达，从而提供更相关的信息。

**多模态处理能力**：一些先进的模型甚至可以处理多模态数据（如文本、声音、图像等），这有助于更全面地理解患者的健康状况。

**持续学习和适应**：随着时间的推移，大语言模型可以通过机器学习不断优化，以更好地适应新的数据和情况。

然而，也存在一些挑战和限制：

**数据质量和偏差**：模型的表现很大程度上依赖于训练数据的质量和多样性。如果训练数据存在偏差，模型的预测也可能存在偏差。

**解释性和透明度**：大语言模型通常被认为是“黑箱”模型，其决策过程可能不够透明，这对于医疗领域尤为重要。

**隐私和安全**：处理敏感的医疗数据时，需要确保数据的隐私和安全。

## 2. 除了健康领域的情感世界理解与识别，你还能想到哪些其他重要的领域，可以开展类似的工作呢？

大语言模型在情感分析和理解方面的应用非常广泛，以下是一些重要的领域：

**客户服务**：通过分析客户的反馈和评论，企业可以更好地理解客户需求和情感，从而改进产品和服务。

**社交媒体监控**：政府和企业可以监控社交媒体上的公众情绪，以应对危机、了解公众对政策或事件的反应。

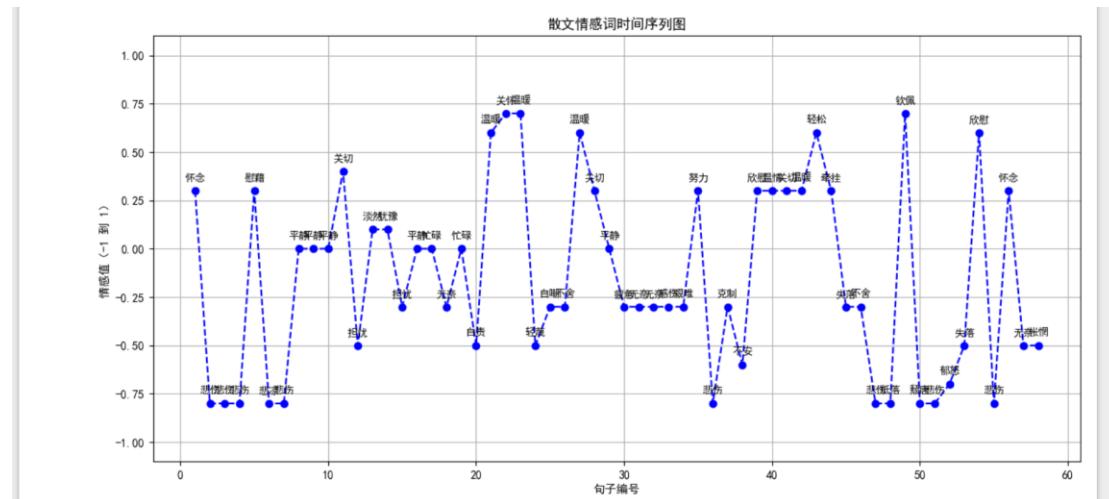
**教育**：在教育领域，通过分析学生的作业、考试答案或在线互动，教师可以更好地理解学生的学习状态和情绪变化。

**人力资源管理**：企业可以通过分析员工的反馈和沟通记录，了解员工的满意度和工作环境，从而改进管理策略。

**市场研究**：通过分析消费者对产品或服务的在线评论，企业可以获取消费者的情感倾向，指导产品开发和市场策略。

**法律领域**：在法律领域，情感分析可以帮助分析证词、法律文件等文本中的情感倾向，为案件提供额外的见解。

代码 4：



句子 6: 情感词: 悲凉, 情感值: -0.8

句子 7: 情感词: 悲伤, 情感值: -0.8

句子 8: 情感词: 平静, 情感值: 0.0

句子 9: 情感词: 平静, 情感值: 0.0

句子 10: 情感词: 平静, 情感值: 0.0

句子 11: 情感词: 关切, 情感值: 0.4

句子 12: 情感词: 担忧, 情感值: -0.5

句子 13: 情感词: 淡然, 情感值: 0.1

句子 14: 情感词: 犹豫, 情感值: 0.1

句子 15: 情感词: 担忧, 情感值: -0.3

句子 16: 情感词: 平静, 情感值: 0.0

句子 17: 情感词: 忙碌, 情感值: 0.0

句子 18: 情感词: 无奈, 情感值: -0.3

句子 19: 情感词: 忙碌, 情感值: 0.0

句子 20: 情感词: 自责, 情感值: -0.5

句子 21: 情感词: 温暖, 情感值: 0.6

句子 22: 情感词: 关怀, 情感值: 0.7

句子 23: 情感词: 温暖, 情感值: 0.7

句子 24: 情感词: 轻蔑, 情感值: -0.5

句子 25: 情感词: 自嘲, 情感值: -0.3

句子 26: 情感词: 不舍, 情感值: -0.3

句子 27: 情感词: 温暖, 情感值: 0.6

句子 28: 情感词: 关切, 情感值: 0.3

句子 29: 情感词: 平静, 情感值: 0.0

句子 30: 情感词: 疲惫, 情感值: -0.3

句子 31: 情感词: 无奈, 情感值: -0.3

句子 32: 情感词: 无奈, 情感值: -0.3

句子 33: 情感词: 感伤, 情感值: -0.3

句子 34: 情感词: 艰难, 情感值: -0.3

句子 35: 情感词: 努力, 情感值: 0.3

句子 36: 情感词: 悲伤, 情感值: -0.8

句子 37: 情感词: 克制, 情感值: -0.3

句子 38: 情感词: 不安, 情感值: -0.6

句子 39: 情感词: 欣慰, 情感值: 0.3

句子 40: 情感词: 温情, 情感值: 0.3

句子 41: 情感词: 关切, 情感值: 0.3

句子 42: 情感词: 温暖, 情感值: 0.3

句子 43: 情感词: 轻松, 情感值: 0.6

句子 44: 情感词: 牵挂, 情感值: 0.3

句子 45: 情感词: 失落, 情感值: -0.3

句子 46: 情感词: 不舍, 情感值: -0.3

句子 47: 情感词: 悲伤, 情感值: -0.8

句子 48: 情感词: 低落, 情感值: -0.8

句子 49: 情感词: 钦佩, 情感值: 0.7

句子 50: 情感词: 颓唐, 情感值: -0.8

句子 51: 情感词: 悲伤, 情感值: -0.8

句子 52: 情感词: 郁怒, 情感值: -0.7

句子 53: 情感词: 失落, 情感值: -0.5

句子 54: 情感词: 欣慰, 情感值: 0.6

句子 55: 情感词: 悲伤, 情感值: -0.8

句子 56: 情感词: 怀念, 情感值: 0.3

句子 57: 情感词: 无奈, 情感值: -0.5

句子 58: 情感词: 怅惘, 情感值: -0.5

## 第六讲:知识图谱理念

### 作业 1:

The screenshot shows a search results page on Taobao. The search bar at the top contains the query "去海南要干什么". Below the search bar, there are several product cards displayed in a grid format. The products include:

- 野生溪黄草**: 野生嫩叶 当季新货 (Price: ¥59.90)
- 野生溪黄草**: 自然晾晒 连州特产 (Price: ¥11.60)
- 新货海麟牌坚果美饼干**: 独立小包装 (Price: ¥35.68)
- 三和茶包邮 不怕肝火**: 三和茶 溪黄草 鸡骨草 罗汉果 花茶 (Price: ¥11.64)
- 压缩海带丝**: 一片就是一盘菜 (Price: ¥33.80)
- 压缩海带丝**: 农科院霞浦压缩海带丝 (Price: ¥37.80)

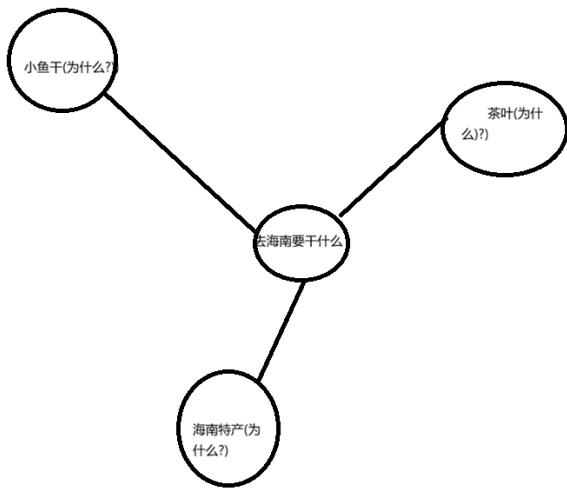
Below these, there are two rows of more products:

- 包邮**: 回头客9千 小雪精选 广西道地药材 (Price: ¥22.90)
- 包邮**: 回头客9千 善圣味旗舰店 (Price: ¥29.80)
- 退货宝 包邮**: 回头客1千 维禾倍达为健专卖店 (Price: ¥21.10)
- 退货宝 包邮**: 回头客3万 海食3盟旗舰店 (Price: ¥12.47)
- 退货宝 破壳包退 包邮**: 回头客3万 海食3盟旗舰店 (Price: ¥10.60)

Finally, there are two rows of products from the "劲仔" brand:

- 劲仔**: 劲仔小鱼仔60包香辣 麻辣小鱼干 毛毛鱼湖南特产... (Price: ¥22.90)
- 劲仔**: 烤鱿鱼丝 (Price: ¥29.80)
- 劲仔**: 深海小鱼仔40包小鱼干 湖南特产办公室辣味休... (Price: ¥21.10)
- 劲仔**: 整箱100包 (Price: ¥13.5)
- 劲仔**: 劲仔小鱼仔麻辣香辣 小鱼干 毛毛鱼小包装湖南特... (Price: ¥12.47)
- 劲仔**: 波力渔趣40克\*5袋 (Price: ¥9.90)

The page also features a banner at the top right encouraging users to search for items on Taobao.

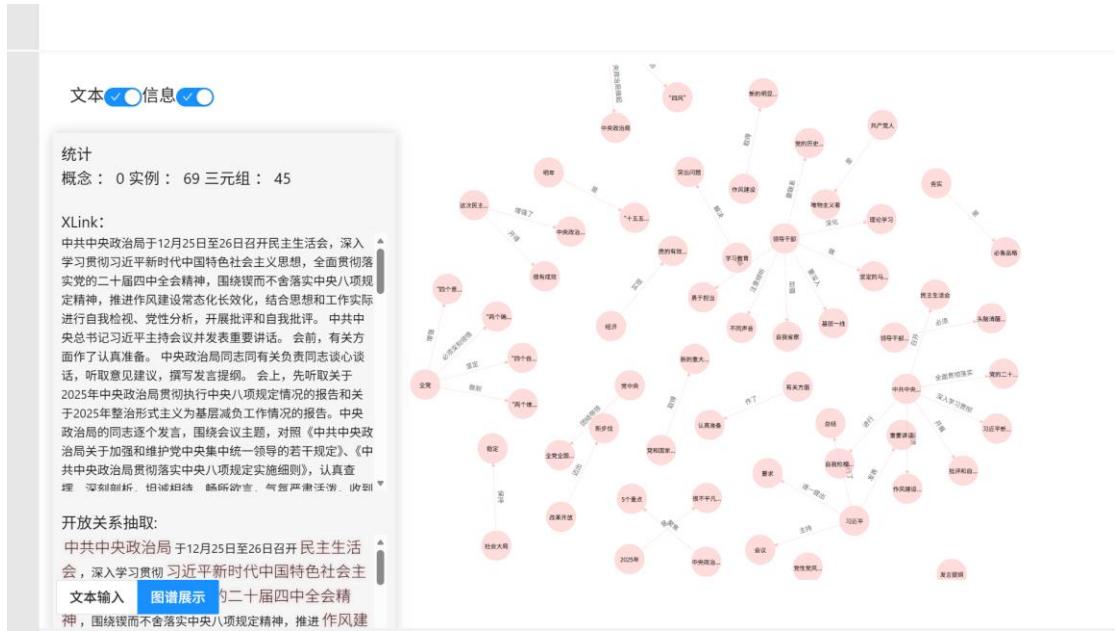


小结:当输入“我要去海南干什么时”,阿里商品大脑的智能导购会想到我去海南的天气是闷热的,需要茶叶进行去火降温.在去海南的路上,阿里商品大脑的智能导购会想你可能会想要吃小零食,所以推荐你小鱼干.当你准备离开海南时,阿里商品大脑的智能导购会想,你需要买海南特产回家

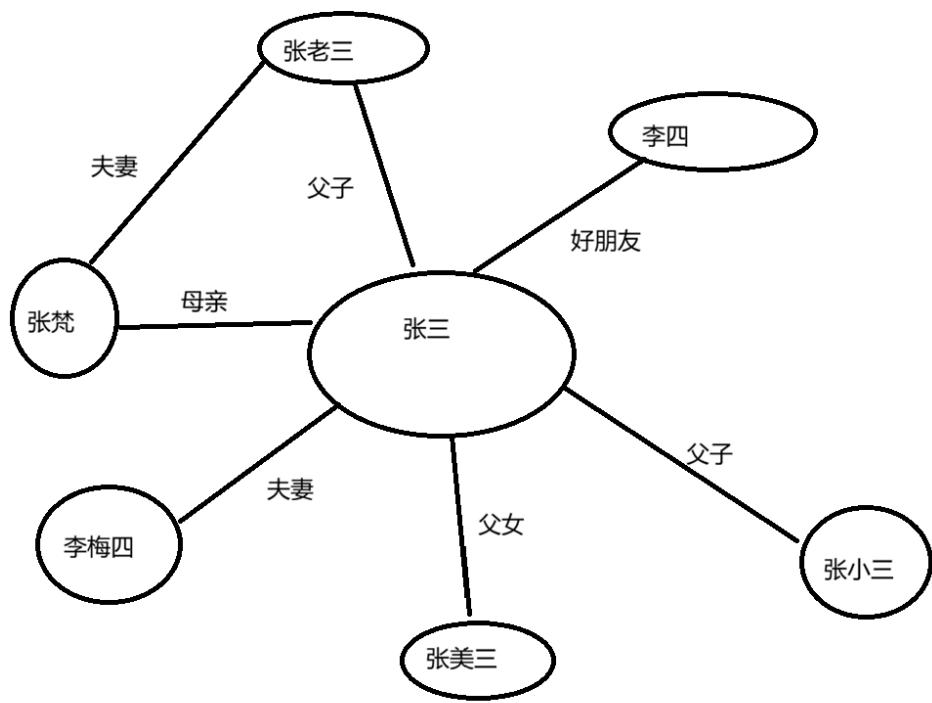
1. 阿里商品大脑的智能导购考虑你从上飞机到下飞机再到离开海南整个时间过程,加入了时间变化的考虑因素
2. 阿里商品大脑的智能导购考虑了海南当季的变化,或者饮食因素,你会需要茶叶去降火减温
3. 阿里商品大脑的智能导购虽然考虑了海南的地理环境,但是没有考虑现在处于冬季,没有实时的考虑事件变化,这是它现在的缺点

## 第六讲(2):知识图谱理念:

作业 1:



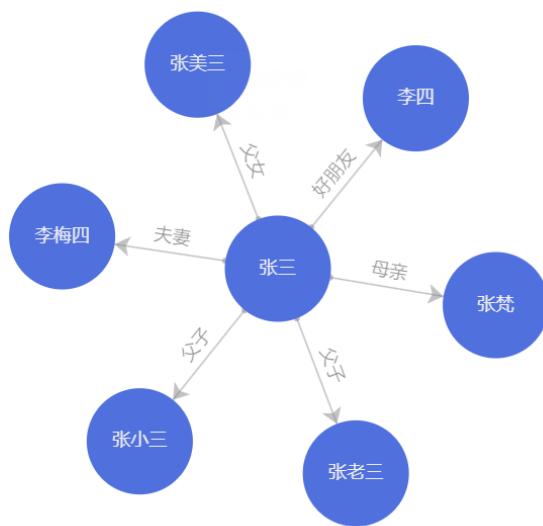
## 作业 2:



## 作业 3:



人物关系图谱



作业 4:

