上海财经大学

课 程 报 告

题目：互联网诊疗发展特点与影响因素的多维文本分析——政策导向、用户需求与医方格局的综合研究

课程名称 文本挖掘

任科教师 史海波

姓　　名 马靖淳

院　　系 统计与管理学院

专　　业 数据科学与大数据技术

学　　号 2020111235

目录

一、项目选题

(一) 选题背景

党的十八大以来，中国坚持把健康摆在优先发展的战略地位。二十大报告提出，推进健康中国建设；促进优质医疗资源扩容和区域均衡布局，坚持预防为主，加强重大慢性病健康管理，提高基层防病治病和健康管理能力。新兴的互联网医疗基于数字技术优势，可以优化医疗供给侧，提升医疗行业整体工作效率，让真正有需求的医患对接，提高医疗资源可及性，降低就医成本，实现医疗服务更便民惠民。

2020 年初爆发的新冠肺炎疫情全球大流行客观上催化了互联网医疗进入全方面加速发展阶段。传统医疗服务行业面临着前所未有的挑战：一方面，医疗资源分配出现严重不足与不均问题，患者产生的就诊需求与医院提供的就诊服务不匹配情况加重；另一方面，医院作为高风险区，许多患者为避免感染不愿也不宜前往医院线下就诊。在此背景下，互联网医疗产业发挥出互联网跨时空共享的优势，迅速发展起来。其中以在线诊疗类服务平台为首，为广大群众提供了线上的医疗支持与咨询服务，不仅满足了患者基本的医疗服务需求，同时也提高了医护人员的工作效率，极大程度上帮助缓解了疫情期间的医疗挤兑问题。以中国头部互联网医疗服务平台好大夫在线为例，截至 2020 年 4 月 12 日，平台日均问诊量同比增长 203%，日均处方量同比增长 430%。在抗疫战场上，互联网医疗在传播抗疫医学知识和信息、协调和配置医疗资源方面发挥了重要作用。

中国互联网医疗的概念最早产生于 2000 年以丁香园和好大夫为代表的网络医疗信息咨询，但其真正的发展还是从“十二五”时期开始，春雨医生、平安好医生等互联网医疗企业创立。时至今日，互联网医疗的概念已拓展为所有新型医疗服务形式的总称，主要基于在线问诊服务，利用现代信息技术与传统医疗服务相结合，其主要的商业模式如下图 1-1所示。其中，在线诊疗平台是互联网医疗领域最重要的商业模式之一，也是本项目的研究重点。

图 1-1: 互联网医疗主要商业模式

(二) 数据集

如前文所述，后疫情时代，互联网医疗行业备受供方、需方和政策方关注。因此对于整个互联网诊疗平台相关文本进行了多渠道的获得，数据丰富且渠道多样化是本项目一大亮点。

对于**政策方**，互联网医疗行业受政策影响极大，为了更好地研究后疫情时代互联网医疗的发展特点，收集获得我国 2014 年至 2022 年互联网医疗共 238 条相关政策文本进行分析。通过对疫情前后政策文本进行政策工具二维分析及词云图绘制，了解国家对互联网医疗行业的政策导向及发展重点。

目前互联网诊疗市场主要参与者如表 1-1，有好大夫、微医、平安好医生、春雨医生等。其中，好大夫在线平台成立最早且活跃订阅者较多，是中国领先的互联网医疗平台之一。并且，截至 2022 年 7 月，平台收录了国内 10000 余家正规医院的 89 万名医生信息，其中 26 万名实名注册医生，累计为全国 8200 万名患者提供在线咨询、预约挂号、远程医疗、健康管理等服务，为用户提供高效、安全、可靠的全方位健康服务。再者，根据上海市卫健委的公开数据可知，截至 2021 年末，上海市所有医院一共有约 8.7 万位医生，其中好大夫在线网站注册医生有约 1.9 万位，占比约为 21.84%。

表 1-1: 互联网诊疗市场参与者

平台	好大夫	微医	春雨医生	平安医生
总部	北京	杭州	北京	上海
创建年份	2006	2010	2011	2014
活跃订阅者	152218	157612	23695	32259
提供服务	在线问诊 在线购药	在线问诊 在线购药 保险	在线问诊 在线购药	在线问诊 在线购药 慢病管理

根据以上分析结果，为了深入了解行业现状，选择好大夫平台作为在线诊疗平台的代表，爬取其重点城市（上海、南京、杭州）的全部数据，包含了 2023 年初上海、南京和杭州三地所有医院和医生 46712 条结构化数据，以及好评墙和意见墙中共计约 270 万字文本评论数据。

对于**需求方**，对上述获得的非结构化用户文本进行语义网络建模分析，提取用户关心的核心问题，另外根据好大夫平台与瑞金医院合作案例，对瑞金医院主页全部评论文本进行进一步爬取，分析合作前后是否对用户有益，分析情感变化；而对于**供给方**，对获得的结构化文本进行聚类分析，并针对获得的四类医生分别选择代表进行访谈，后续再对访谈文本进行主题建模，分析四类医生的不同特征。

(三) 所用的库

在数据爬取过程中，采用 requests 包发送 HTTP 请求，使用 re 包对后续网页链接进行合成，使用 BeautifulSoup 帮助查找、提取和修改数据，使用 time 包对获得数据进行处理。在后续爬取瑞金医院评论过程中，因需要获得评论日期，使用 lxml 库中的 etree 模块解析 html 文档，使用 xpath 函数获得日期信息。

对于词云图绘制部分，采用了 pandas 包读取数据，使用 re 包提取中文字符，使用 jieba.lcut 全模式语法进行分词，其中 cut_all=True 表示采用全模型进行分词。全模式会把文章中有可能的词语都扫描出来，有冗余，即在文本中从不同的角度分词，变成不同的词语。然后使用 numpy 包进行词频统计，wordcloud 包、matplotlib 包以及 PIL 包进行词云图绘制及图片保存等操作。

对于语义网络图绘制部分，采用与上述类似的读取数据、分词操作，使用 tkinter 包中的 _flatten 功能协助分词预处理，使用 numpy 包进行词频统计及共现语义的初始矩阵创建。在矩阵构建完成后，因使用 network 包绘制结果过于粗糙且连边较多不够清晰，最后在 Gephi 软件（https://gephi.org/）中导入前述共现矩阵计算结果并绘制语义网络图。

对于情感分析部分，首先在可视化部分使用 pandas 包读取数据，然后使用 dataframe 相关操作筛选处理数据，最后将获得的数据结果使用 matplotlib 包进行绘制，因数据绘制结果较为粗糙所以将筛选数据结果导入镝数（www.dycharts.com/）进行图片绘制。将爬取获得的好评文本导入 snownlp 包中进行情感分析，输出好评情感得分，并对好评得分进行统计。

对于主题模型部分，首先使用 R 语言对获得的结构化文本进行聚类，使用 readxl 包读取爬取获得的 excel 文件（使用 csv 文件一直报错），然后使用 dplyr 使用 pipe(% > %) 操作符来实现数据的转换和筛选，使用 cluster 包进行后续分析，并结合 factoextra 包进行聚类相关可视化。然后使用 python 进行主题模型建模部分，采用与上述类似的读取数据、分词操作。使用 gensim 包中 corpora 模块用于创建字典（Dictionary）和文档-词袋（Document-Term Matrix），而 models 模块中 LdaModel 类用于训练和使用 LDA 模型。CoherenceModel 类用于计算主题模型的一致性（coherence）分数，通过这个分数计算确定最佳主题数，最后使用 pyLDAvis 包进行可视化操作。

二、方法与模型

(一) 模型理论背景

1. 词云图理论介绍

词云，又称文字云，英文名：Word Cloud，是文本数据的视觉表示形式。词云图是一种图形化展示文本数据中关键词频率的方式。制作词云图需要对文本进行预处理，包括分词、去除停用词（如“的”、“是”等常见词语）、统计词频等。然后，根据词频将关键词按照大小不同进行排列，生成词云图。简而言之就是对指定范围文本中出现频率较高的“关键词”予以视觉上的突出表现，从而过滤掉大量的文本信息，形成“关键词云层”或“关键词渲染”，使浏览者只要一眼扫过图片就可以获得文本的主题宗旨。

2. 语义网络理论介绍

语义网络（Semantic Network）是一种用于表示语义关系和知识结构的图形模型。这种网络结构可以用于在语义空间中建模实体（entities）和它们之间的关系，以及表示各种概念和事实之间的语义联系。

其中，语义网络中的节点代表实体或概念，而边表示它们之间的语义关系。节点和边的组合形成一个图形结构。另外，节点可以通过多个关系连接到其他节点，使得网络更加灵活和丰富，良好的结构性也使其可以把事物的属性以及事物间的各种语义联系直观简洁地表现出来。

3. 情感分析理论介绍

Snownlp 是一个用于中文文本情感分析的 Python 包，它具有对中文文本进行情感分析、文本分类、情感词提取等功能。Snownlp 中的 sentiments 属性是用于获取文本情感得分的属性。情感得分是一个范围在 0 到 1 之间的值，表示文本的情感极性，越接近 1 表示积极的情感，越接近 0 表示消极的情感。这个得分是通过情感分析算法计算得出的。

Snownlp 的情感分析算法基于情感词典和规则，而非深度学习模型。它使用一些启发式规则来判断文本中的情感，并根据情感词在文本中的分布和权重来计算最终的情感得分。

具体而言，Snownlp 的情感分析过程包括以下步骤：

（1）情感词典：Snownlp 使用了内置的中文情感词典，其中包含了一些带有情感极性的词汇，如积极词和消极词。

（2）程度副词和否定词：程度副词（如“很”、“非常”等）和否定词（如“不”、“没有”等）会影响情感的强度和极性，Snownlp 考虑了这些因素。

（3）情感得分计算：根据情感词的权重和分布，以及程度副词和否定词的影响，计算文本的情感得分。这个得分越接近 1 表示积极情感，越接近 0 表示消极情感。

4. 聚类理论及 LDA 模型理论介绍

（1）聚类理论

K-means 聚类分析的目的是将 n 个对象分为 k 类，类内相似度较高，而类间相似度较低。每一簇由其成员对象和其中心或质心定义。对于不同的簇，K-means 聚类分析通过迭代，使得簇内对象到质心的距离和最小。

对于 K-means 聚类，首先确定 k 个聚类中心 $\{C_1, C_2, ..., C_k\}$, 计算 n 个对象中每个对象到聚类中心的欧式距离：

$$d(X_i, C_j) = \sqrt{\sum_{t=1}^{m}(X_{it} - C_{jt})^2}$$

其中 X_i 表示第 i 个对象，X_{it} 表示第 i 个对象的第 t 个属性，其中 C_j 表示第 j 个聚类中心，C_{jt} 表示第 j 个聚类中心的第 t 个属性。依次比较每一个对象到每一个聚类中心的聚类，将对象分配到距离最近的聚类中心的类中，得到 k 个类 $\{S_1, S_2, ..., S_k\}$，并计算类中心：

$$C_t = \frac{\sum_{X_i \in S_1} X_i}{|S_1|}$$

如此重复计算并比较每一个对象到每一类中心的距离，将对象分配到距离最近的类中心的类中，如此反复迭代，实现聚类。

在确定最佳聚类个数过程中，使用了簇内平方和（inertia，也叫误差平方和）评估聚类质量，计算公式为：

$$Inertia = \sum_{i=1}^{n}\sum_{j=1}^{k}||x_i - \mu_j||^2$$

（2）LDA 理论介绍

gensim 包中包含了用于处理大规模文本语料库的工具，例如实现了词嵌入技术（Word Embeddings）的 Word2Vec 模型，以及 Latent Dirichlet Allocation (LDA) 等主题模型。Latent Dirichlet Allocation（LDA）是一种生成式概率模型，用于解释一个文档集合中的主题结构。LDA 假设每个文档都是由多个主题混合而成的，而每个主题则由多个单词组成。这种混合的过程是通过概率分布来描述的，其中 Dirichlet 分布是一种常用的分布。

对每个文档 d，会从 Dirichlet 分布中抽取主题分布 θ_d，然后对于文档中的每个单词 w，从主题分布 θ_d 中抽取主题 z，从主题 z 的多项分布中抽取单词 w。文档 d 中第 n 个单词 $w_{d,n}$ 的生成过程，下述中 α, β 默认状态下会设置为'auto'，在 gensim 包中会自动学习到最佳的参数值：

$$P(z_{d,n} = k|\theta_d) = \theta_{d,k}$$

$$P(w_{d,n} = j|z_{d,n} = k, \beta) = \beta_{k,j}$$

LDA 的似然函数可以表示为对文档集合中所有文档的单词的联合概率：

$$P(\text{ 文档集合 }|\alpha, \beta) = \prod_{D}^{d=1}\int_{\theta_d}(\prod_{n=1}^{N}\sum_{k=1}^{K}P(z_{d,n} = k|\theta_d) \cdot P(w_{d,n}|z_{d,n} = k, \beta)) \cdot P(\theta_d|\alpha)d\theta_d$$

在最佳主题数确定过程中，计算了 Coherence 指标用以评估主题模型质量。计算公式为：

$$Coherence = \frac{2}{|W|\cdot(|W|-1)}\sum_{i=1}^{|W|}\sum_{j=i+1}^{|W|} log\frac{D(w_i,w_j)+1}{D(w_i)\cdot D(w_j)}$$

(二) 整体框架

1. 绘制词云图

对于政策文本词云图绘制整体框架为：

（1）文本读取

采用了 pandas 包读取数据，选择政策标题列以及政策内容列作为词云分析文本。

（2）中文分词及停用词去除

政策文本基本为中文，但部分中含有数字，对数字部分进行去除。另外，导入四个常用停用词表（百度停用词、中文停用词、哈工大停用词、四川大学停用词），另外根据后续绘制情况，对互联网诊疗相关政策中特有的无用词汇进行去除，如“互联网”、“医疗”、“加强”、“推进”等等，自行建立停用词表，进一步去除，从而提取重要信息。

（3）词频统计

在词云图绘制过程中对政策文本中词汇出现次数进行统计，如上述提及的“互联网”、“医疗”、“加强”、“推进”等，通过统计实际出现数量辅助肉眼进行筛选。

（4）图片美化

下载与互联网、医疗相关的元素，进行图片轮廓的美化。另外，报告总体采用蓝色元素，对词云图上文本颜色进行了统一颜色规定。

2. 绘制语义网络图

对于用户评论文本语义网络图绘制整体框架为：

（1）文本预处理

采用了 pandas 包读取数据，以及上述类似的中文分词及停用词去除操作。

（2）词频统计

对上述用户评论文本政策文本中词汇出现次数进行统计，对分词进行词频统计并选取前 20 个高频词作为构建共现矩阵的关键词。

（3）构建共现语义矩阵

使用 numpy 设置共现语义的初始矩阵。计算关键词之间的共现次数，共现规则为关键词之间的位置距离不超过 1，保存为 csv 文件。

（4）图片美化

采用 network 包进行共现网络的初步构建，发现连边较多且过于杂乱，并且中文字符不清晰，最后在 Gephi 软件中导入共现矩阵文件并绘制语义网络图。

3. 情感分析流程

首先，将爬取获得的文本数据导入 pandas 中。

其次，导入 snownlp 中的 sentiments 函数对文本进行情感打分。

最终，输出情感打分结果并根据时间进行统计并获得可视化结果。

4. 聚类及主题模型流程

(1) 聚类流程

首先，对医生热度方面的 5 个指标计算 Pearson 相关系数，发现各指标间均存在较强的相关性，故热度方面从中只选取总访问量作为代表。

然后，综合考虑指标的直观意义、数据的完整性、数据的解释性和变量类型等方面，筛选获得聚类指标。

接着，计算组内平方和随聚类个数变化的情况选择合适的类目数。

最后，对聚类结果进行分析和展示。

(2) LDA 建模流程

首先，将四类医生访谈文本分别导入 pandas 中，然后使用 cancat 操作连接成一个 dataframe，然后进行分词等操作。

然后，构建文本数据的词典 dictionary 以及文档-词袋模型 corpus，它将文本数据表示为一个稀疏矩阵，其中每一行对应一个文档，每一列对应一个词汇表中的单词，而矩阵中的值表示对应单词在文档中出现的次数。

接着，进行 coherence 计算和绘图，从而确定最佳主题数。然后导出最佳主题数的训练结果。

最后，将训练结果导入 pyLDAvis 包中进行可视化，并保存为 html 文件。

(三) 编程实现

这一部分因代码数量过多，只展示了部分数据处理或模型建构等操作，具体全部代码可见最终提交文件，已经按模块分类好便于查看。

1. 爬虫获得数据

对部分爬虫过程的部分代码进行展示，主函数部分太复杂放在报告中占据篇幅太大故不放入：

```
1 def SaveHtml(detailpage_list, date_list): # 保存每个日期中的100个问诊页面的
      html文件
2     print('SaveHtml开始')
3     path_p = r'E:\0、大四上\2、文本挖掘 项目\项目\0、数据爬取\SaveHtml'
4     if not os.path.exists(path_p):
```

```
        os.mkdir(path_p)
    for i in range(len(detailpage_list)):
        path_c = path_p + r'\{}'.format(date_list[i])
        if not os.path.exists(path_c):
            os.mkdir(path_c)
        os.chdir(path_c)
        temp_list = detailpage_list[i][0][:]
        num = 0
        for page in temp_list:
            try:
                r = requests.get(page, headers=UAPool())
                r.raise_for_status()
                r.encoding = 'gbk'
            except:
                time.sleep(2)
                continue
            try:
                soup = BeautifulSoup(r.text, parser='html.parser', features=
                    'lxml')
                #  “ ? *: " < > \ / | ”  敏感字符不可出现在文件名中
                post_title = soup.find_all('div', attrs={'class': 'fl-title
                    ellps'})[0].text.replace('?', ',').replace('*', ',').
                    replace(':', ',').replace('"', ',').replace('<', ',').
                    replace('>', ',').replace('\\', ',').replace('/', ',').
                    replace('|', ',')
                with open(post_title+'.html', 'w', encoding='utf-8') as f:
                    f.write(r.text.replace('<meta charset="gbk">', '<meta
                        charset="utf-8">'))
                num += 1
            except:
                continue
            if num >= 100:
                break
            time.sleep(2)
```

2. 绘制政策文本词云图

对词云图绘制部分核心函数进行展示：

```
def draw_wc(datalist):
    # 导入中文停用词
    name_list = ['stopword/baidu_stopwords.txt', 'stopword/cn_stopwords.txt
        ', 'stopword/hit_stopwords.txt','stopword/scu_stopwords.txt','
        stopword/select.txt']
    stop_word = []
    for x in name_list:
        stop_word.extend(sw(x))
    stop_word = list(set(stop_word))
    # 中英文分词
    sentence = datalist['政策名称'].tolist()
    sentence.extend(datalist['政策内容'].tolist())
    txt_c = ""
    counts = {}
    background_image= np.array(Image.open(r'E:\0、大四上\2、文本挖掘 项目\项
        目\词云图绘制\医院.png'))
    color_list=["#5B9BD5","#4472C4"]
    colormap=colors.ListedColormap(color_list)
    for i in range(0,len(sentence)):
        line = ''.join(re.findall('[\u4e00-\u9fa5]',str(sentence[i])))
        if line:
            c = jieba.lcut(line)
            result = [word for word in c if word not in stop_word]
            c = [word for word in result if len(word)>1]
            txt_c += " ".join(c)
            # 词频统计
            for c_i in c:
                if len(c_i) == 1:
                    continue
                else:
                    counts[c_i] = counts.get(c_i,0) + 1
    # 词频统计处理
```

```
    items = list(counts.items())
    items.sort(key=lambda x:x[1], reverse=True)
    # 绘制中文词云图
    font = r'C:\Windows\Fonts\simkai.ttf'
    wc = WordCloud(font_path=font,
                   background_color='white',
                   width=1000,
                   height=800,
                  mask=background_image, # 指定词云的形状
                  colormap=colormap
                  ).generate(txt_c)
    wc.to_file(get_varname(datalist) + '.png')
    return items
```

3. 绘制用户评论文本语义网络图

对语义网络图中共现矩阵构建部分代码进行展示：

```
matrix = np.zeros((len(keywords)+1)*(len(keywords)+1)).reshape(len(
    keywords)+1, len(keywords)+1).astype(str)
matrix[0][0] = np.NaN
matrix[1:, 0] = matrix[0, 1:] = keywords

cont_list = [cont.split() for cont in cut_word_list]
for i, w1 in enumerate(word_fre[:20].index):
    for j, w2 in enumerate(word_fre[:20].index):
        count = 0
        for cont in cont_list:
            if w1 in cont and w2 in cont:
                if abs(cont.index(w1)-cont.index(w2)) == 0 or abs(cont.index
                    (w1)-cont.index(w2)) == 1:
                    count += 1
        matrix[i+1][j+1] = count
```

4. 对合作前后用户评论进行情感分析

```
for i in range(0,len(data)):
  # 情感得分
  try:
    emo = SnowNLP(data[10][i].strip()).sentiments
    data[11][i] = emo
  except:
    print('异常')
print('write down.')
```

5. 对各类医生访谈文本进行主题分析

(1) 对平台医生进行聚类

对聚类前数据清洗部分进行展示:

```
df <- df %>%
  mutate("职称级别" = case_when(
    str_detect(df$职级, "主治") ~ 2,
    str_detect(df$职级, "副主任") ~ 3,
    str_detect(df$职级, "主任") ~ 4,
    str_detect(df$职级, "主管") ~ 2,
    TRUE ~ 1 # 默认值, 可以根据需要修改
  ))

df <- df %>%
  mutate("问诊价格" = as.numeric(str_extract_all(df$在线问诊, "\\d+")))

df <- df %>%
  mutate("开通年数" = year(Sys.Date()) - year(ymd(df$开通时间)))
```

计算簇内平方和选择最佳聚类个数:

```
# Compute and plot wss for k = 1 to k = 15
k.values <- 1:15
# extract wss for 2-15 clusters
wss_values <- map_dbl(k.values, wss)
```

```
\begin{figure}
    \centering
    \includegraphics[width=0.5\linewidth]{coherence.png}
    \caption{Enter Caption}
    \label{fig:enter-label}
\end{figure}
plot(k.values, wss_values,
     type="b", pch = 19, frame = FALSE,
     xlab="Number of clusters K",
     ylab="Total within-clusters sum of squares")
```

聚类过程：

```
num_clusters <- 4
kmeans_result <- kmeans(df, centers = num_clusters)
kmeans_result$cluster
kmeans_result$centers
```

（2）各类医生访谈文本进行主题分析

构建文本数据的词典 dictionary 以及文档-词袋模型 corpus：

```
data_set = df['分词'].tolist()
dictionary = corpora.Dictionary(data_set) # 构建词典
corpus = [dictionary.doc2bow(text) for text in data_set] #表示为第几个单词
    出现了几次
```

计算 coherence 选择最佳主题数：

```
#计算coherence
def coherence(num_topics):
    ldamodel = LdaModel(corpus, num_topics=num_topics, id2word = dictionary
        , passes=50, random_state = 1)
    # print(ldamodel.print_topics(num_topics=num_topics, num_words=5))
    ldacm = CoherenceModel(model=ldamodel, texts=data_set, dictionary=
        dictionary, coherence='c_v')
    print(ldacm.get_coherence())
    return ldacm.get_coherence()

```

```
x = range(1,8)
y = [coherence(i) for i in x]
plt.plot(x, y)
plt.xlabel('主题数目')
plt.ylabel('coherence大小')
plt.rcParams['font.sans-serif']=['SimHei']
matplotlib.rcParams['axes.unicode_minus']=False
plt.title('主题-coherence变化情况')
plt.show()
```

LDA 模型训练:

```
lda = LdaModel(corpus=corpus, id2word=dictionary, num_topics=4, passes =
    50,random_state=1)

for topic in lda.print_topics(num_words = 5):
    termNumber = topic[0]
    print(topic[0], ':', sep='')
    listOfTerms = topic[1].split('+')
    for term in listOfTerms:
        listItems = term.split('*')
        print('  ', listItems[1], '(', listItems[0], ')', sep='')
```

pyLDAvis 包中进行可视化，并保存为 html 文件

```
pyLDAvis.enable_notebook()
d = gensimvis.prepare(lda, corpus, dictionary)
pyLDAvis.save_html(d, '医生.html')
```

三、实验结果

(一) 政策文本词云图展示：从认可、规范诊疗到纳入医保

本文将我国互联网医疗发展划分为三个阶段如图 3-1。一是政策发展阶段（2014-2017 年），互联网医疗政策在国家层面正式提出，但仍处于不稳定状态；二是疫情突发阶段（2018-2020 年），2018 年开始，在线诊疗逐渐进入市场理性期，新冠疫情黑天鹅推动行业实现突破性进展；三是疫情常态化阶段（2020 年末至今），随着突发阶段线上医疗助力国内医疗资源调配，政府出台了一系列政策保障互联网医疗有序发展，政策落实方面更具时代性、针对性也更强。

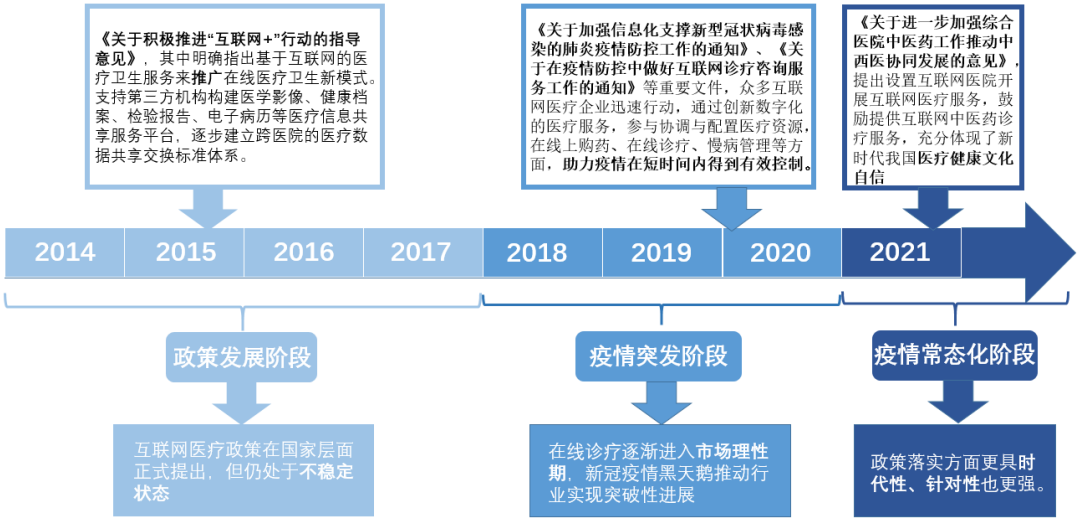

图 3-1: 我国互联网医疗政策发展阶段

表 3-1: 互联网医疗政策工具二维分析

政策工具	发展阶段	规范阶段	政策工具	发展阶段	规范阶段	政策工具	发展阶段	规范阶段
供给型			**需求型**			**环境型**		
基础建设	16	41	医保支付	0	18	目标规划	5	0
资金投入	1	4	国际合作	1	0	组织实现	2	8
人才培养	3	2	对口支援	11	5	法规管制	6	21
公共服务	18	14	公私合作	4	4	技术标准	3	10
技术支持	11	15	试点\示范	3	4	宣传推广	3	5
合计	49	76	合计	19	31	合计	19	44

对疫情前后相关政策进行政策工具二维分析，汇总后如表 3-1所示。在政策发展阶段，在互联网医疗政策工具的运用具有明显的差异，在所运用的 87 条政策工具中，供给型政策工具有 49 条，运用最多，环境型与需求型政策工具均为 19 条。在疫情爆发以后，政策工具共 151 条，供给型政策工具中基础设施建设工具的使用仍为最多，共计 76 条；需求型 31 条政策工具中，医保支付政策工

具显著增加，为 18 条。而在 44 条环境型政策工具中，为保证人民群众线上就医的安全与质量，法规管制政策工具的使用显著增加。

对疫情突发阶段进行词云图绘制如图 3-2，可以看出这一阶段中政策中提及“管理”、“完善”、“建立”两词的次数相对较多，反映出国家在疫情爆发阶段比较重视互联网医疗的建立工作，提出相应的完善和管理要求，整个行业处于快速发展阶段。

对疫情常态化阶段政策进行词云图绘制如图 3-3，政策中提及“制度”“机制”较多，国家在此阶段比较重视行业的规范化，并保障使用者的权益。另外，在这一阶段国家也出台一系列医保政策，希望能够打通消费者使用在线诊疗的关键环节。

图 3-2: 疫情爆发阶段政策词云图

图 3-3: 疫情常态化阶段政策词云图

总而言之，我国互联网诊疗平台政策也经历了从认可、规范诊疗到纳入医保的演变。在疫情爆发阶段，国家重视互联网医疗的建立工作，政策方面倾向于推动整个行业快速发展；在疫情常态化阶段，因行业发展较快，国家更重视整个行业的规范化，从而保障使用者的权益。

(二) 语义网络图展示：用户的好评与差评特征

对于平台上爬取到的好评墙和意见墙的文本数据，本项目希望从中挖掘出评价的内在驱动因素，因此考虑进行语义网络分析。

1. 好评墙：“康复”、“顺利”、“耐心”

在好评墙的语义网络图 3-4中，病情、看病、手术、医生等词语处于网络的核心地位，它反映了求医治病仍是在线问诊用户所关注的核心内容。此外，语义网络中正常、良好、好转等正面的形容词指向康复，而耐心、详细、顺利等指向过程，说明良好的疗效，以及诊疗过程中医生的耐心细致容易促使用户诊后做出好评。

图 3-4: 好评墙语义网络

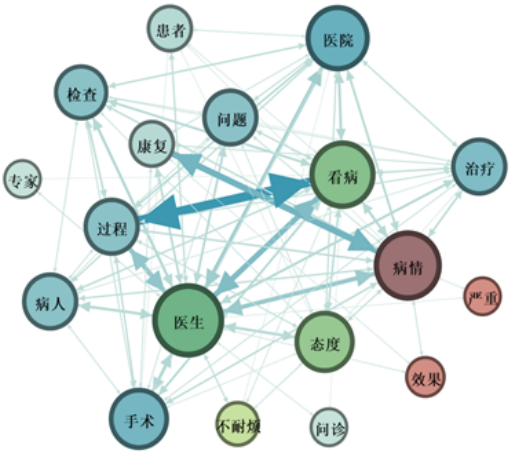

图 3-5: 意见墙语义网络

2. 意见墙：“严重”、“不耐烦”

在意见墙图 3-5的语义网络中，病情、看病、手术、医生等词仍处于网络的核心地位，与好评墙的语义网络相比，意见墙的语义网络中态度占据了相对更重要的位置，并且医生、看病、态度的箭头指向了不耐烦，反映出不耐烦的态度容易促使就诊者产生负面意见。此外，网络中还出现了严重和效果两词与病情的连接，说明疗效的欠佳、病情的严重同样将会促使就诊者产生负面意见。

综合好评墙和意见墙的语义网络，可见疗效和态度的好坏是影响在线问诊用户评价好坏的两个关键要素，这反映了在线问诊用户的核心需求是获得良好的疗效，而在此过程中医生态度是否耐心细致也至关重要。

(三) 情感得分展示：医企合作提升患者使用感受

为探究实体医院对互联网医疗平台的供给情况，针对瑞金医院与好大夫平台的合作案例展开分析。自 2021 年 4 月起，上海交通大学附属瑞金医院与好大夫在线诊疗平台正式开展战略合作。尽管在此之前，好大夫平台上已有许多瑞金医院的医生在此服务患者，但在合作后仍发生了新的变化。通过对好大夫平台瑞金医院数据进行爬取，对比分析合作前后得到如下结论：

1. 合作前后可视化展示

（1）瑞金医院已开通预约挂号比例在全国领先

据研究，上海交通大学附属瑞金医院与好大夫平台于 2021 年 4 月开展合作主要内容是精准对接预约通道，鼓励实体医院医生积极在好大夫平台开通无偿加班门诊，从而精准对接患者。通过将瑞金医院与前面部分爬取全部医院的数据进行对比，可以发现瑞金医院医生在好大夫平台开放预约通道的已达 33.66%，显著高于全部医院的平均水平 26.42%。预约通道的开通便捷了患者的就医流程，让患者可以提前在好大夫平台进行预约，直接前往瑞金医院一楼的护士台领取相关就诊凭证，省

去了线下排队的困难。

（2）合作后新增 101 位医生覆盖 31 个科室，且有新科室出现

爬虫数据显示，在 2021 年四月以前，瑞金医院在好大夫平台注册的医生总人数为 838 人，分布在 56 个科室。在瑞金医院与好大夫平台展开合作后，瑞金医院在好大夫平台新增医生 101 名，涨幅高达 12%，新增人数多；新增医生覆盖 31 个科室，覆盖范围广。如上图 3-6展示新增医生数量最多的 15 个科室，尽管普外科因相较其他科室使用在线平台判断疾病更加便捷，总人数和新增人数相较其他科室遥遥领先，但许多科室如神经内科、乳腺疾病诊疗中心等，涨幅也超过 30%，并且有新科室出现，如护理部。可见在合作后，在线诊疗服务朝更精细化、纵深化的方向发展。

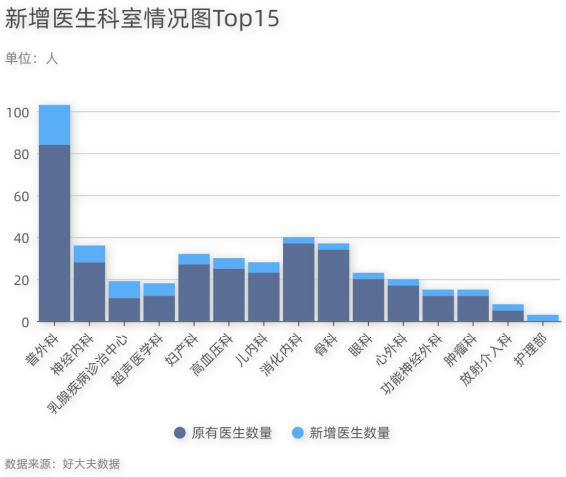

图 3-6: 新增医生科室情况图

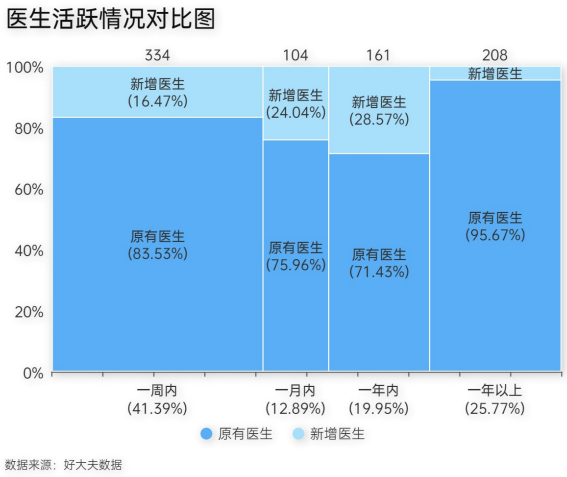

图 3-7: 医生活跃情况对比图

(3) 合作后平台新增医生活跃度显著提升

爬虫数据统计结果表明，在 2021 年 4 月后，瑞金医院医生登录平台数量占全部医生数量的 80%。由上图 3-7也可以清晰地看出，对比合作时间（2021 年 4 月）前后注册的医生可见，在距今一年内（2022 年 4 月后）登录好大夫平台的医生总人数占 75% 左右，其中新增医生占比 21%；一年以上未登录好大夫平台的医生相对不活跃，其中新增医生仅有 9 人，占比 4%。由此看出，在瑞金医院与好大夫平台合作后，新增医生活跃度显著提升。

2. 合作前后用户情感分析

根据上述可视化结果，在合作后瑞金医院的医生活跃度以及科室丰富度都有了很大的提升，那是否真正便捷了患者就医，下面将对瑞金医院患者评论文本进行情感分析。

通过爬取数据，发现自 2021 年 4 月起，瑞金医院在好大夫平台上展示的全部好评文本达到了 7019 条，差评文本只有 79 条样本太少训练效果较差，所以主要对好评文本进行分析。进一步对这些好评文本进行了情感分析，通过 snownlp 为每条评论获得了情感得分。这些情感得分反映了患者在评价中表达的情感强度，高得分代表积极情绪越强烈。

对这些数据进行年度分析，发现了一个有趣的趋势如图 3-8：随着时间推移，患者好评的积极程

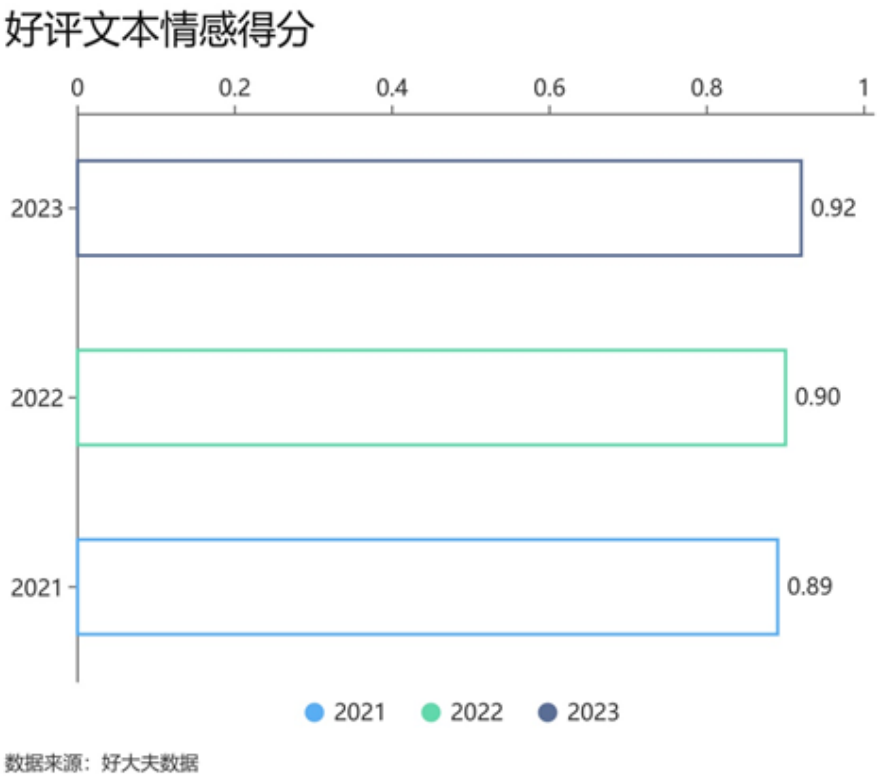

图 3-8: 好评文本情感得分图

度逐年增强。具体而言，2021 年的平均情感得分为 0.89，2022 年为 0.9，而 2023 年更是达到了 0.92。根据 snownlp 情感打分理论得知，越接近 1 表示积极的情感，越接近 0 表示消极的情感。这说明患者对好大夫平台服务的满意度逐年提升，患者们在评价中表达的积极情感不断增强。这一趋势反映了瑞金医院在好大夫平台上提供的医疗服务的不断改进和优化。患者的满意度提升可能与医疗技术的进步、医护团队的专业水平以及服务质量的提高等因素密切相关。通过这些数据，可以更全面地了解患者对医院服务的整体评价，并为医院提供进一步改进的方向和建议。

由上述结果可知，在好大夫平台与瑞金医院合作后，一方面，医生活跃度显著提高，预约挂号实现精准对接，在线诊疗服务朝更精细化、纵深化的方向发展。另一方面，患者对平台的满意度逐年提高，预约挂号服务的有效推进也便捷了患者。由此可见，线下医院应积极与诊疗平台展开相关合作，拓宽互联网诊疗的供给，为患者提供便利。

(四) 主题模型展示：平台医生分化状况明显

1. 聚类分析：平台医生分为四类，流量集中效应明显

本节将进行 K-means 聚类分析。综合考虑指标的直观意义、数据的完整性、数据的解释性和变量类型等方面，筛选后聚类指标如下表 3-2：

表 3-2: 聚类指标及含义

聚类指标	含义
总访问量	医生主页被访问的总次数，单位：次
职称级别	分类变量，1 为初级，2 为中级，3 为副高级，4 为正高级
在线问诊	在线问诊的价格，单位：元
病友推荐度	平台为每个医生给出的推荐度，取值范围 0-5
总文章数	医生在平台上发表的科普文章等数量，单位：篇
开通年数	用最新年份-开通年份计算所得，单位：年
昨日访问	医生主页昨日被访问的次数，单位：次
上次在线时间	取值范围 0-31，单位：天前

图 3-9反映组内平方和随聚类个数变化的情况，从中可见组内平方和在类别数由 1 变化到 4 时下降很快，而之后下降逐渐变缓，故选取类别个数 k=4，然后利用 R 软件得到聚类结果如下：

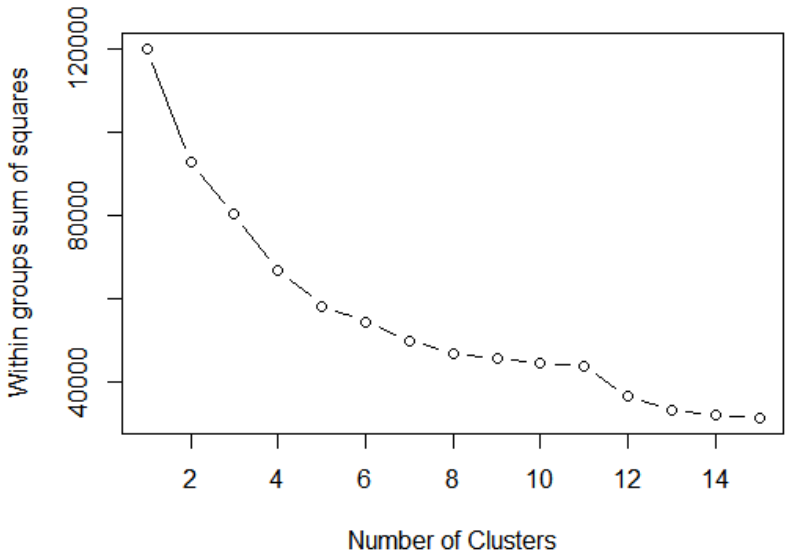

图 3-9: 组内平方和随类别数变化图

数据分析结果如下表 3-3：

表 3-3: 聚类指标及含义

cluster	类别占比	总访问量	职称级别	病友推荐度	总文章数	上次在线	开通年数
1	33.6%	917464.5	3.3	3.8	20.6	2.9	8.6
2	27.1%	308635.8	3.4	3.4	5.2	29.2	9.5
3	38.4%	53885.2	2.0	3.3	2.9	20.9	4.1
4	0.9%	18519884.4	3.7	4.1	348.1	2.9	12.9

综合数据分析四类医生的特征，可将聚类结果归纳如下图 3-10，注意这里顶流医生、优质医生、普通医生和尾部医生分别对应 K-means 聚类结果的类别 4、类别 1、类别 2 和类别 3，更高的活跃度指更频繁的在线和更多的科普文章。

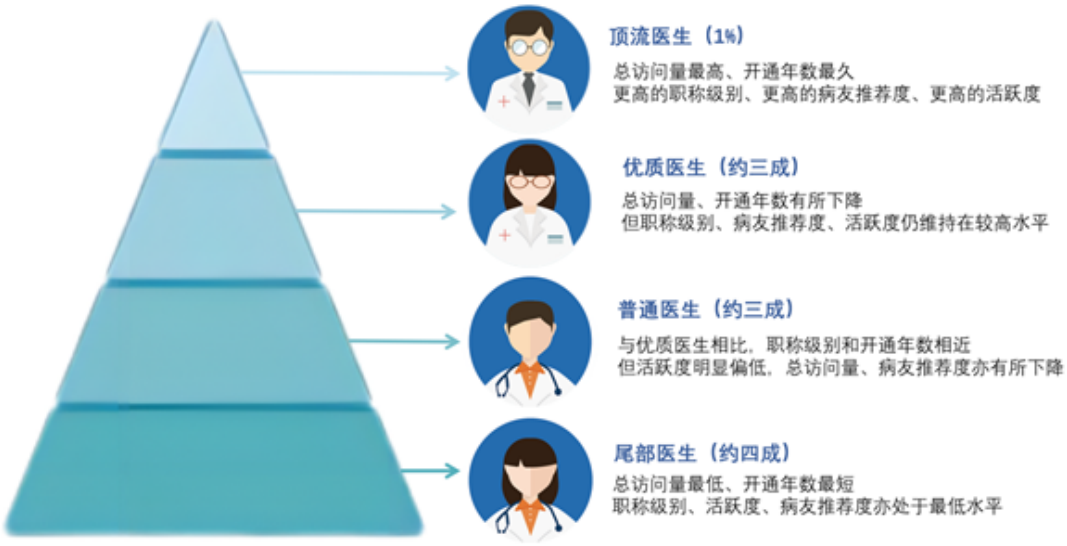

图 3-10: 医生分类结果

由上述分析可见，平台上的医生存在一定的流量集中效应：极少数医生能够获得很高的流量，而占比最多的为流量较低的尾部医生，并且医生活跃度、职称级别、病友推荐度的明显差异在一定程度上促进了这种分化。

2. 主题模型分析：不同类别医生使用感受不同

根据上述分类结果，为从医生角度深度细致地了解医生对在线问诊的认知及意愿、培训及规范、工作情况及感受，问题及展望，对于上海医院就职医生就在线问诊相关话题进行访谈。进行访谈的医生均使用过好大夫平台，任职科室涉及外科、内科，并覆盖各职级。依据聚类结果，四位医生分别属于顶流医生、普通医生、优质医生、尾部医生。医生基本情况如下表 3-4：

表 3-4: 医生基本情况表

	医生 1	医生 2	医生 3	医生 4
任职医院	上海肿瘤医院	复旦大学附属华山医院	上海第一人民医院	大桥社区医院
任职科室	外科	感染科	心内科	社区医院医生
所属类别	顶流医生	优质医生	普通医生	尾部医生

对上述四位医生分别进行互联网诊疗相关访谈获得的文本去除一些互联网诊疗常见词比如“互联网”、“诊疗”，“医疗”等词汇后一起投入 LDA 模型进行训练，图 3-11反映 coherence 随主题个数变化的情况，从中可见组内平方和在类别数由 1 变化到 4 时上升很快，而之后下降逐渐变缓，故主题个数 k=4。

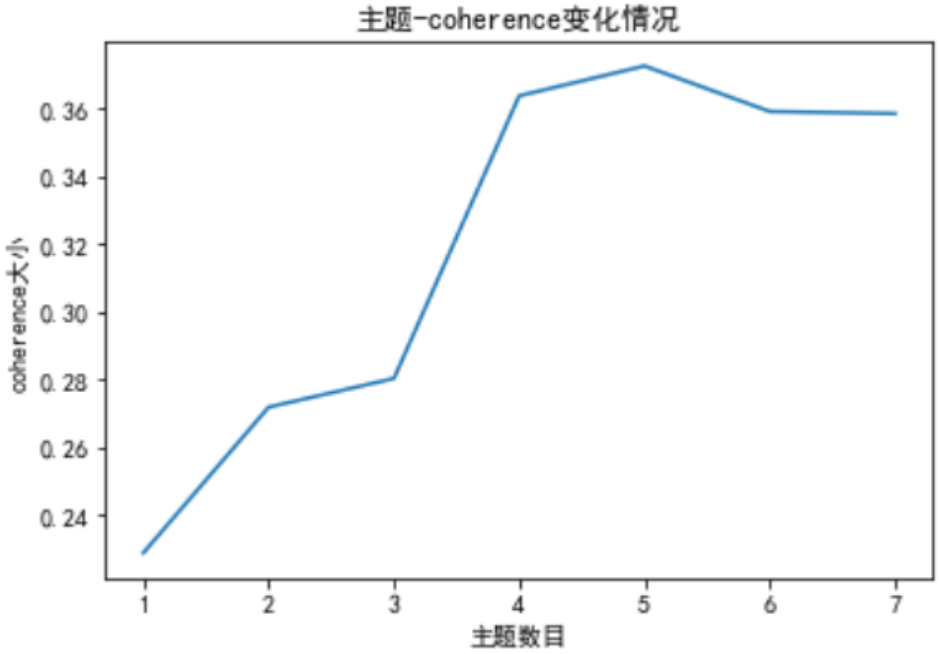

图 3-11: coherence 随主题数变化图

对于 pyLDAvis 可视化结果 html 文件进行分析发现，首先根据气泡图分布可以看出四类主题几乎没有交叉，可见上述四类医生对于互联网诊疗的认知和看法，以及未来的发展测重点各有不同。具体分析每一个主题发现：

第一类主题如图 3-12，关键词为“学科”、“影响”，“AI”等，可以对应于优质医生，其认为目前业态发展较快，新技术比较发达，并且从医生角度能发挥所处医疗学科（比如某种罕见肿瘤）优势，增加业务量，因此优质医生目前积极性较高。

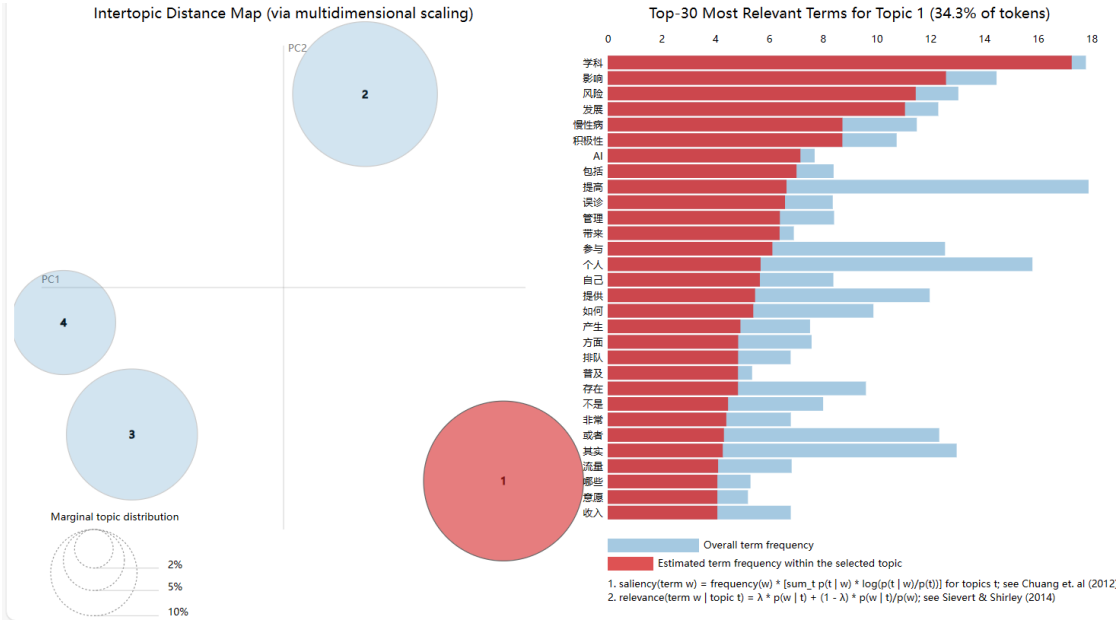

图 3-12: 主题 1 结果图

第二类主题如图 3-13，关键词为“个人”，“IP”，“薪酬”等，可以对应于顶流医生，其认为未来随着流量增加可能会出现医生个人 IP 等，可以提高薪酬，结合前述聚类结果发现目前平台上的医生存在一定的流量集中效应，因此顶流医生目前积极性也很高。

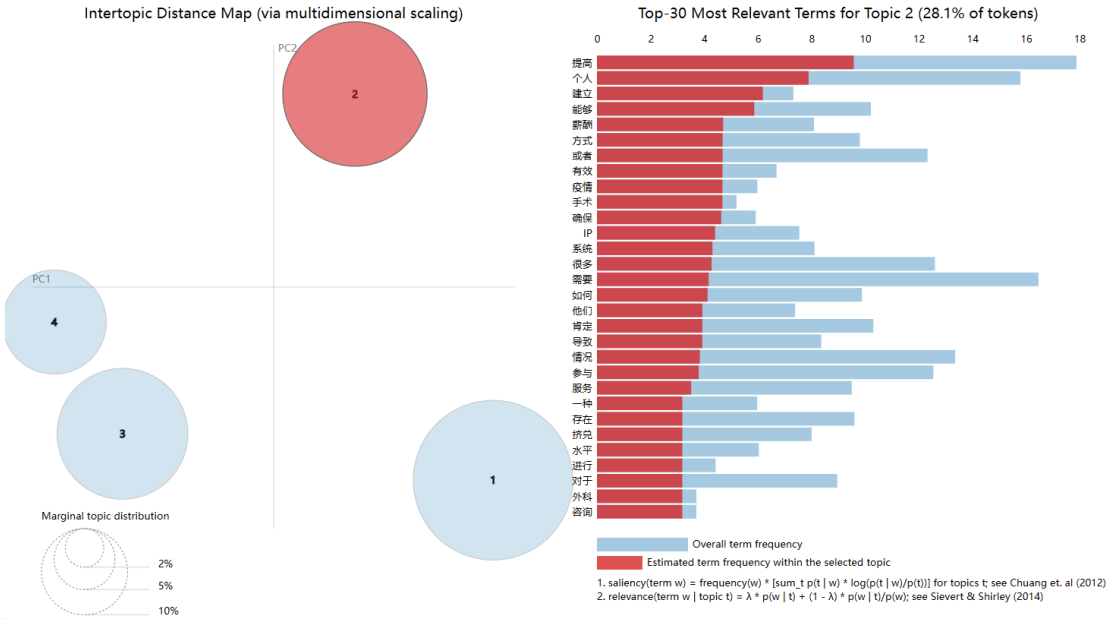

图 3-13: 主题 2 结果图

第三类主题如图 3-14，关键词为“预约”，“资源”，“挤兑”等，可以对应于普通医生，其观点主要为线上预约导致线下挂号更加混乱，其实加剧了医疗资源的挤兑问题，其认为没有很好的帮助缓解看病难，挂号难等问题。

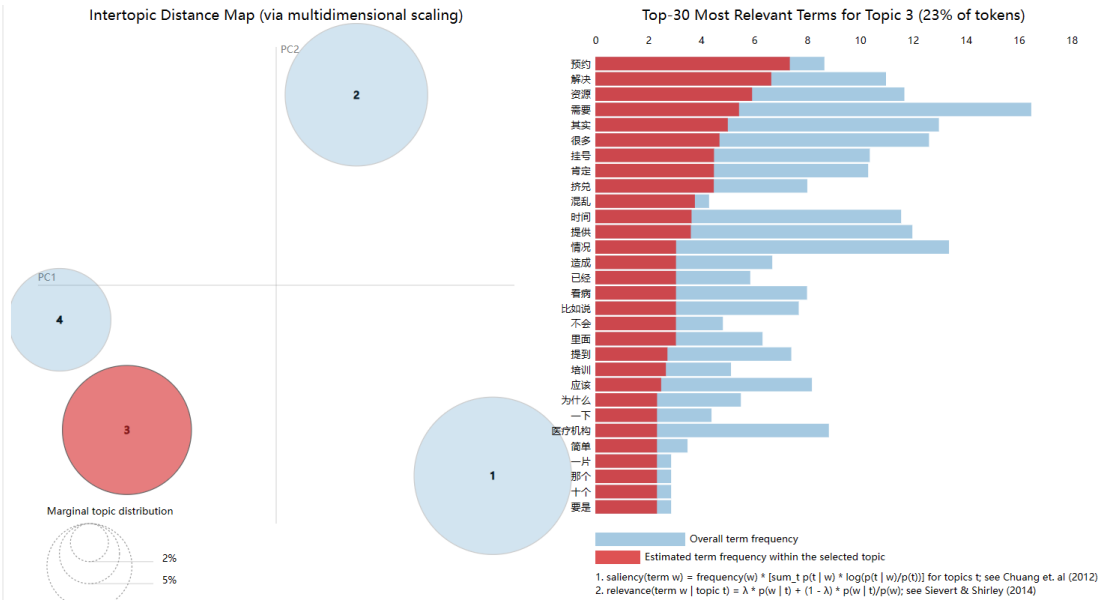

图 3-14: 主题 3 结果图

第四类主题如图 3-15，关键词为“效率”，“时间”，“浪费”等，可以对应于尾部医生，其观点主要为线上诊疗无法替代传统医疗模式中的面诊小效率较低，不适用于首次病发的患者；并且占用医生个人时间，因此使用互联网诊疗的积极性不足。

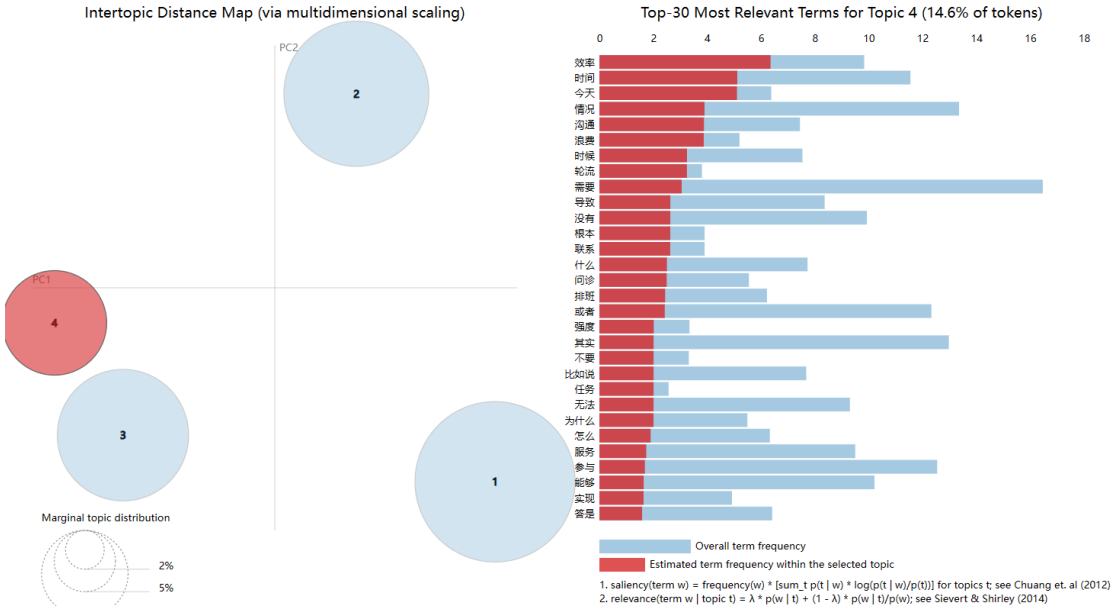

图 3-15: 主题 4 结果图

四、总结与讨论

(一) 方法与结果总结

通过对词云图的绘制过程，对课上所授方法进行了实践，但在实际文本处理过程中要根据实际情况进行调整，比如根据绘制结果与词频统计结果对获得的词云图进行调整，去除不需要的信息，保留有针对性的信息。

通过对 LDA 模型的使用，对课上所授的方法进一步理解，更深刻的理解了主题模型的整体建模过程，从而进行调参等操作，最终获得了不错的主题分类结果。

在 LDA 模型可视化过程中，gensim 包和 pyLDAvis 包下载过程中需要注意版本的配合和函数名称的变化，出现 LDA 进行 pyldavis 可视化报错：TypeError: Object of type complex is not JSON serializable 时，可以怼 pyLDAvis/utils.py 文件进行修改。

```
class NumPyEncoder(json.JSONEncoder):
    def default(self, obj):
        if isinstance(obj, np.int64) or isinstance(obj, np.int32):
            return int(obj)
        if isinstance(obj, np.float64) or isinstance(obj, np.float32):
            return float(obj)
        if np.iscomplexobj(obj):
            return abs(obj)
        return json.JSONEncoder.default(self, obj)
```

通过多种多样数据的获取，并针对数据特点使用不同的模型，锻炼了我对文本分析的实践过程，收获了很多。

(二) 课程外技术运用

使用爬虫技术，从而获得丰富的结构化数据和非结构化数据，通过大量结构化数据可视化等操作，可以帮助后续非结构化数据的信息挖掘。

受本学期修读的一门社交大数据选修课的启发，除了社交关系外，是否可以根据同一条文本中各分词同时出现的频率从而构建出类似社交网络的语义网络模型，通过查找发现确实可以进行类似的操作。使用 python 中相关包进行数据预处理和统计，再与 Gephi 软件配合，获得了较为清晰的语义网络构建，并获得了不错的结论。

使用 R 语言对平台医生进行聚类分析，将医生分为四类，从而有针对性的获取访谈文本，再对不同类别医生访谈结果进行主题分析，从而挖掘不同医生的关注点和看法。

另外，在使用 python 或 r 等编程工具筛选数据后，可初步尝试使用 matplotlib 或 ggplot2 绘制

相关图表，再保存数据分析结果导入一些可视化工具中（如 Gephi、熵数等），会相较于先前绘图结果更清晰，更适合在报告中进行展示。

(三) 不足与改进

对于访谈文本的获取，目前可访谈到的医生大多是通过私交，因为很多医生因为本身工作较多也较难约到，并且我个人时间也相对有限，因此获得的样本量相对较小，希望未来能拓展途径扩大这一部分的访谈文本量。

五、项目分工

马靖淳：项目策划，项目数据收集与爬取，数据预处理，模型构建，结果分析与可视化，报告撰写。