中文文本情感分析

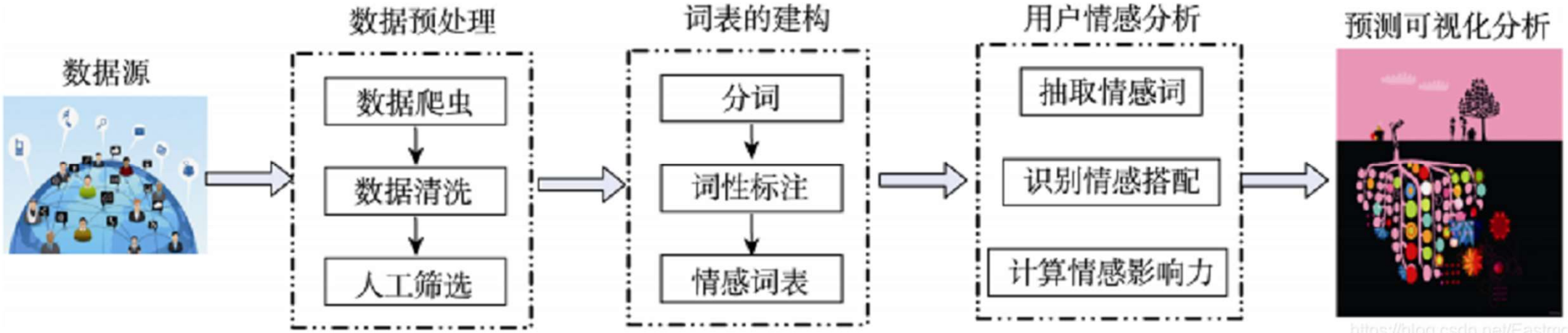
2025.04.10

研讨会目标：达到能够独立使用LDA与情感分析的能力

국제대학원

情感分析（Sentiment Analysis）是自然语言处理（NLP）的一个重要分支，也被称为观点挖掘（Opinion Mining）。它是一种通过计算机技术自动识别和提取文本中主观情感倾向的过程。简单来说，情感分析就是判断一段文字表达的是正面、负面还是中性的情绪或态度。

"这家餐厅的食物很美味(+1)，但服务态度差劲(-1)"
→ 情感得分：0（中性）



前期准备（上节课已经做过了）

1.安装PyCharm

1. 如果尚未安装PyCharm，请下载并安装，见群内教程

2.将文件复制到桌面

1. 在开始之前，确保您的Python文件(.py)和相关数据源已复制到桌面上
2. 请复制 **数据.CSV** 和 **情感分析代码.py** 以及**情感词汇本体.xlsx**到桌面

1. 数据.CSV

这个文件包含了需要进行情感分析的原始文本数据:
它是一个表格格式的文件，通常包含多列数据

2. 情感分析代码.py

这是执行情感分析的Python脚本文件:
包含完整的情感分析流程，从数据读取到结果输出
主要功能包括:

- 自动从桌面读取CSV文件并处理中文编码
- 进行中文分词（使用jieba库）
- 结合SnowNLP和情感词典进行情感打分
- 生成可视化图表（散点图和直方图）
- 输出情感分析结果到新的CSV文件

3. 情感词汇本体.xlsx

这是增强情感分析精确度的关键资源文件:
包含大量已标注情感极性的中文词汇(可以理解成词典)

基础科学计算和数据处理库
pip install **numpy pandas matplotlib**
中文分词库
pip install **jieba**
中文情感分析库
pip install **snownlp**
机器学习库（可选）
pip install **scikit-learn**
Excel处理库（用于读取情感词汇本体）
pip install **openpyxl**

SnowNLP，中文语言处理的必备工具 - 知乎

什么是 SnowNLP?

SnowNLP 是一个专为中文文本处理设计的 Python 库，由中国开发者创建，旨在提供简单易用的中文自然语言处理工具。其名称中的"Snow"（雪）象征着纯净和简洁，反映了该库设计理念——为中文 NLP 任务提供轻量级、高效的解决方案。

应用场景

1.社交媒体分析：监测微博、社交平台上的用户情感

2.产品评论分析：自动处理电商平台的产品评价

3.新闻情感监测：分析新闻报道的情感倾向性

4.客户反馈处理：快速分类和提取客户反馈的关键信息

5.教育研究：作为中文NLP入门学习的工具

1. 文件读取和准备

•自动从桌面读取CSV文件

•自动尝试不同编码（UTF-8, GBK）以确保正确读取中文文本

•准备分析环境


```
9
10  warnings.filterwarnings('ignore')
11
12  # 获取桌面路径
13  desktop = str(Path.home() / "Desktop")
14
15  # 在桌面上寻找所有CSV文件并获取第一个
16  csv_files = [f for f in os.listdir(desktop) if f.endswith('.csv') and f != '情感分析结果.csv']
17  if not csv_files:
18      print("桌面上没有找到CSV文件（除了情感分析结果.csv）")
19      exit(1)
20
21  file_path = os.path.join(desktop, csv_files[0])
22  print(f"正在读取文件: {file_path}")
23
```

2. 情感词典建立

•读取情感词汇本体文件（如果存在）

•构建包含正面词汇、负面词汇和程度词的情感词典

•处理否定词，用于识别情感反转


```
# 情感词典初始化
sentiment_words = {
    'positive': set(),
    'negative': set(),
    'degrees': {}  # 程度词及其权重
}

# 否定词，用于情感反转
negation_words = set([
    '不'  '没'  '无'  '非'  '莫'  '弗'  '毋'  '勿'  '未'  '否'
    '别'  '反'  '难'  '禁'  '还是'  '尚未'  '并未'  '并没'  '不曾'  '不太'
    '不怎么'  '不很'  '从不'  '绝不'  '几乎不'  '从未'  '尚无'  '并非'  '决非'
])
# 寻找情感词汇本体
print("寻找情感词汇本体...")
sentiment_lexicon_path = os.path.join(desktop, "情感词汇本体.xlsx")
if not os.path.exists(sentiment_lexicon_path):
```

3. 中文文本分词与处理

•使用jieba对文本进行分词

•预处理文本以便进行情感分析


```
207  # 执行中文分词
208  text_column = None
209  for col in df.columns:
210      if df[col].dtype == object:  # 找第一个文本列
211          text_column = col
212          break
213
```

4. 情感分析方法

•使用SnowNLP进行基础情感打分
•结合情感词汇本体进行自定义分析
•考虑否定词对情感的反转作用
•考虑程度副词对情感强度的影响

국제대학원

```
223
224  def enhanced_sentiment_analysis(text):  1个用法
225      """
226      结合SnowNLP和情感词汇本体的增强情感分析函数
227      返回0-1之间的得分，0表示极度负面，1表示极度正面
228      """
229      try:
230          # 基本SnowNLP情感分析
231          snow_score = SnowNLP(str(text)).sentiments
232
233          # 使用情感词汇本体进行自定义分析
234          text = str(text)
235          words = jieba.lcut(text)
236
237          # 初始化
238          sentiment_score = 0
239          negation_count = 0  # 否定词计数
```

5. 数据可视化

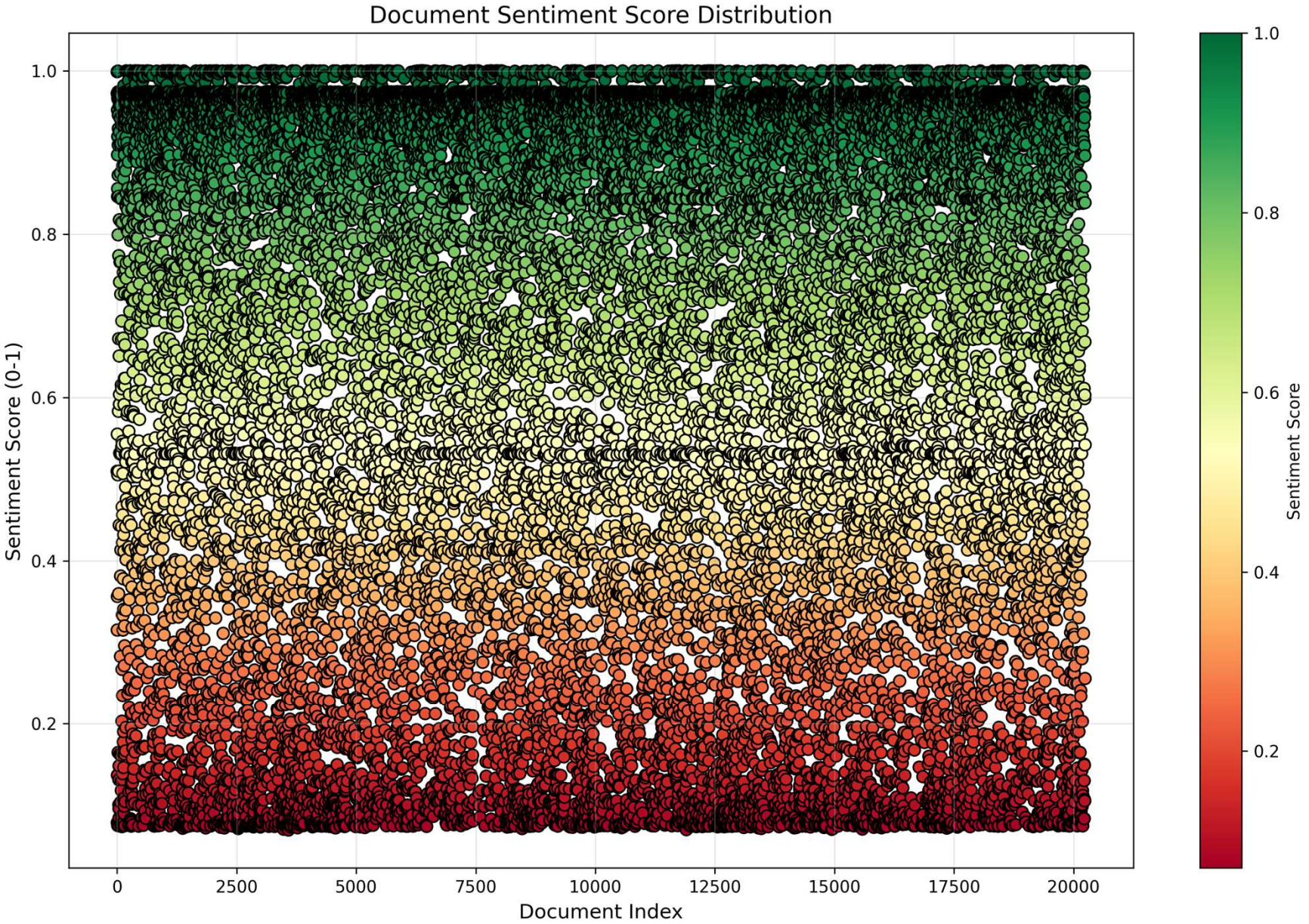
•创建情感得分散点图，使用从红到绿的颜色映射
•生成情感得分分布直方图
•保存可视化结果到桌面

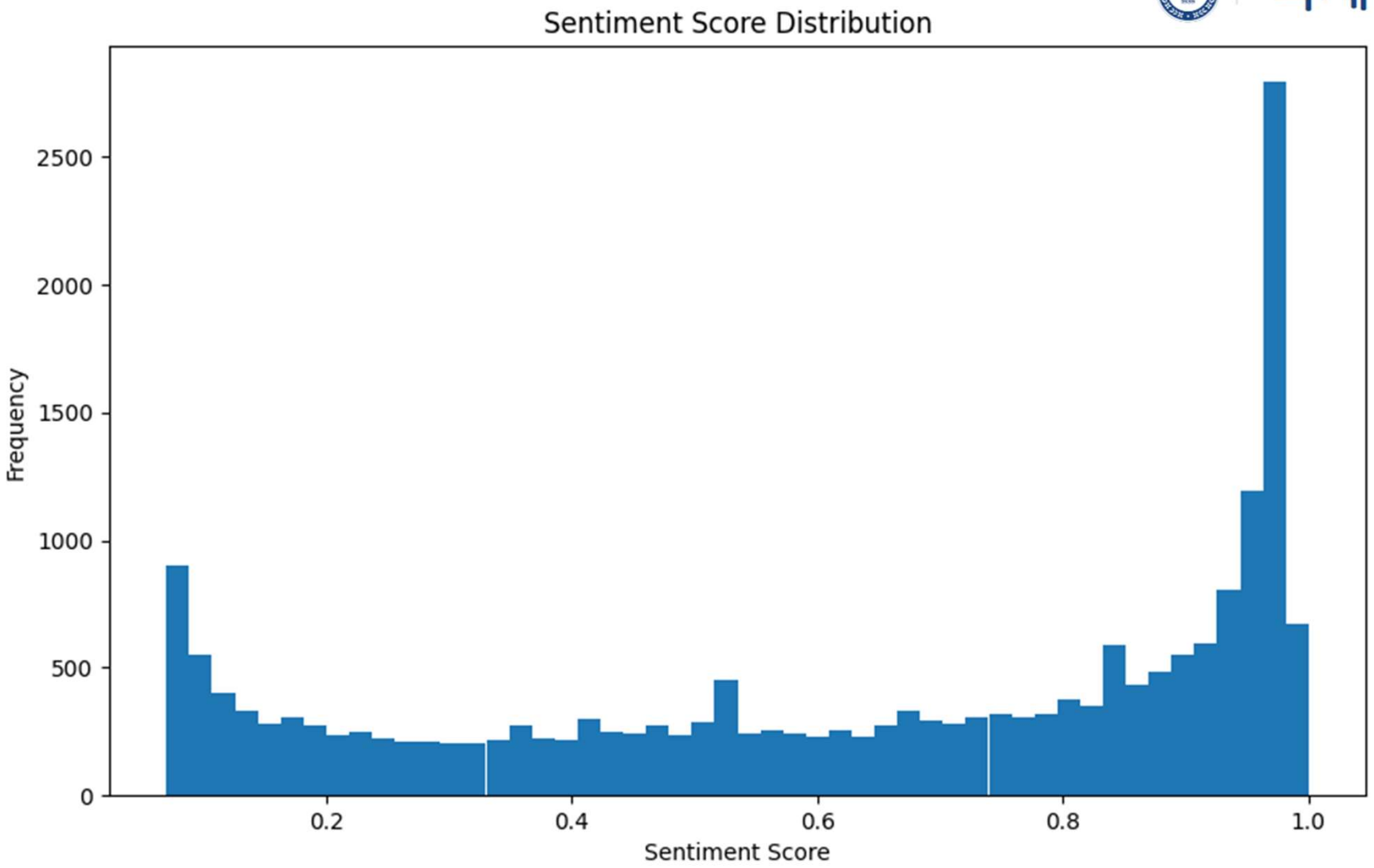
```
22
23  plt.xlabel( xlabel: 'Document Index', fontsize=12)
24  plt.ylabel( ylabel: 'Sentiment Score (0-1)', fontsize=12)
25  plt.title( label: 'Document Sentiment Score Distribution', fontsize=14)
26  plt.grid( visible: True, alpha=0.3)
27  plt.colorbar(scatter, label='Sentiment Score')
28  plt.tight_layout()
29  plt.savefig( *args: os.path.join(desktop, 'sentiment_score_scatter.png'), dpi=300)
30  plt.close()
31  print(f"情感得分散点图已保存至: {os.path.join(desktop, 'sentiment_score_scatter.png')}")
32  plt.figure(figsize=(10, 6))
33  plt.hist(df['sentiment_score'], bins=50)
34  plt.xlabel('Sentiment Score')
35  plt.ylabel('Frequency')
36  plt.title('Sentiment Score Distribution')
37  plt.savefig(os.path.join(desktop, 'score_distribution.png'))
38  # 创建包含评分的新CSV文件
```

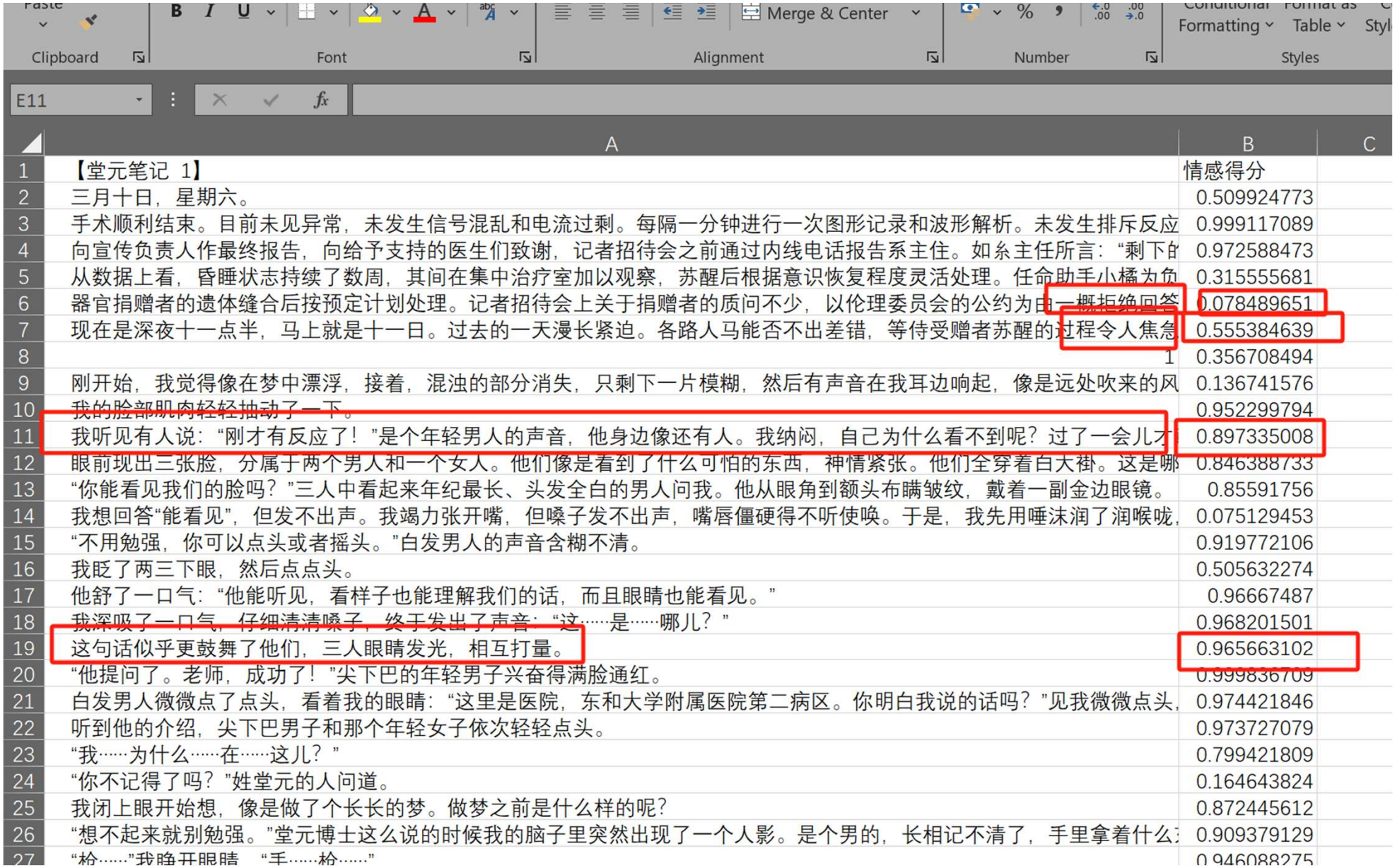
请复制 **数据.CSV** 和 **情感分析代码.py** 以及**情感词汇本体.xlsx**到桌面

执行代码

```
Prefix dict has been built successfully.
准备情感分析...
执行情感分析...
情感得分散点图已保存至: C:\Users\misty\Desktop\sentiment_score_scatter.png
生成包含情感评分的CSV文件...
包含情感分析结果的CSV文件已保存至: C:\Users\misty\Desktop\情感分析结果.csv
中文文本情感分析已完成!
```


Document Sentiment Score Distribution
Sentiment Score (0-1)
1.0
0.8
0.6
0.4
0.2
0
2500
5000
7500
10000
12500
15000
17500
20000
Document Index
Sentiment Score
1.0
0.8
0.6
0.4
0.2


Sentiment Score Distribution
Frequency
2500
2000
1500
1000
500
0
0.2
0.4
0.6
0.8
1.0
Sentiment Score



	A	B	C
1	【堂元笔记 1】	情感得分	
2	三月十日，星期六。	0.509924773	
3	手术顺利结束。目前未见异常，未发生信号混乱和电流过剩。每隔一分钟进行一次图形记录和波形解析。未发生排斥反应	0.999117089	
4	向宣传负责人作最终报告，向给予支持的医生们致谢，记者招待会之前通过内线电话报告系主住。如糸主任所言：“剩下的	0.972588473	
5	从数据上看，昏睡状志持续了数周，其间在集中治疗室加以观察，苏醒后根据意识恢复程度灵活处理。任命助手小橘为负	0.315555681	
6	器官捐赠者的遗体缝合后按预定计划处理。记者招待会上关于捐赠者的质问不少，以伦理委员会的公约为由一概拒绝回答	0.078489651	
7	现在是深夜十一点半，马上就是十一日。过去的一天漫长紧迫。各路人马能否不出差错，等待受赠者苏醒的过程令人焦急	0.555384639	
8	1	0.356708494	
9	刚开始，我觉得像在梦中漂浮，接着，混浊的部分消失，只剩下一片模糊，然后有声音在我耳边响起，像是远处吹来的风	0.136741576	
10	我的脸部肌肉轻轻抽动了一下。	0.952299794	
11	我听见有人说：“刚才有反应了！”是个年轻男人的声音，他身边像还有人。我纳闷，自己为什么看不到呢？过了一会儿才	0.897335008	
12	眼前现出三张脸，分属于两个男人和一个女人。他们像是看到了什么可怕的东西，神情紧张。他们全穿着白大褂。这是哪	0.846388733	
13	“你能看见我们的脸吗？”三人中看起来年纪最长、头发全白的男人问我。他从眼角到额头布瞒皱纹，戴着一副金边眼镜。	0.85591756	
14	我想回答“能看见”，但发不出声。我竭力张开嘴，但嗓子发不出声，嘴唇僵硬得不听使唤。于是，我先用唾沫润了润喉咙，	0.075129453	
15	“不用勉强，你可以点头或者摇头。”白发男人的声音含糊不清。	0.919772106	
16	我眨了两三下眼，然后点点头。	0.505632274	
17	他舒了一口气：“他能听见，看样子也能理解我们的话，而且眼睛也能看见。”	0.96667487	
18	我深吸了一口气，仔细清清嗓子，终于发出了声音：“这……是……哪儿？”	0.968201501	
19	这句话似乎更鼓舞了他们，三人眼睛发光，相互打量。	0.965663102	
20	“他提问了。老师，成功了！”尖下巴的年轻男子兴奋得满脸通红。	0.999836709	
21	白发男人微微点了点头，看着我的眼睛：“这里是医院，东和大学附属医院第二病区。你明白我说的话吗？”见我微微点头，	0.974421846	
22	听到他的介绍，尖下巴男子和那个年轻女子依次轻轻点头。	0.973727079	
23	“我……为什么……在……这儿？”	0.799421809	
24	“你不记得了吗？”姓堂元的人问道。	0.164643824	
25	我闭上眼开始想，像是做了个长长的梦。做梦之前是什么样的呢？	0.872445612	
26	“想不起来就别勉强。”堂元博士这么说的时候我的脑子里突然出现了一个人影。是个男的，长相记不清了，手里拿着什么	0.909379129	
27	“枪……”我睁开眼睛 “手……枪……”	0.946088275	

政治人物媒体报道中的性别偏见

Article

Social Science Computer Review
2025, Vol. 43(2) 279–294

DOI: 10.1177/08944393241269097
journals.sagepub.com/home/ssc
Sage

Combining Natural Language Processing and Statistical Methods to Assess Gender Gaps in the Mediated Personalization of Politics

Emanuele Brugnoli[1,2], Rosaria Simone[3], and Marco Delmastro[2,4]

数据收集与范围

研究对象包括意大利担任部长、副国务卿、州长和6万以上人口城市市长的政治人物，共**270位**（57位女性，213位男性），时间跨度为2017-2020年。

媒体来源覆盖了254家在线新闻媒体（覆盖93%的意大利互联网用户）和83家主要意大利报纸（覆盖43%的成年人口）。研究者总计分析了约180万篇提及这些政治人物的新闻文章。

个人化词汇表的构建

研究者创建了**一个包含3,303个词的词汇表**，用于识别政治人物报道中的个人化元素，包括2,125个形容词、1,084个名词和94个动词。

这些词被分为三个主要类别：道德和行为态度（如"强硬"、"诚实"）、身体外表和个人风格（如"高大"、"时尚"）以及社会和经济背景（如"富有"、"贵族"）。

每个词由5位评注者**赋予情感值**：-1（负面）、0（中性）、1（正面）。词语的总体情感取五位评注者评分的平均值，最终分为五类：强烈负面、轻微负面、中性、轻微正面和强烈正面。评注者间一致性达到0.712（Krippendorff's α系数），表明评分相当可靠。

考虑性别不平衡的覆盖指数

为解决政治界女性代表性不足的问题，研究者开发了覆盖偏见指数 I(w):

$I(w) = [\tilde{t}F(w) - \tilde{t}M(w)] / [\tilde{t}F(w) + \tilde{t}M(w)]$

其中$\tilde{t}F(w)$是调整后的词语w与女性政治人物关联的出现率，$\tilde{t}M(w)$是调整后的词语w与男性政治人物关联的出现率。

该指数取值范围为[-1, 1]：I(w) = 1表示该词仅用于女性政治人物，I(w) = -1表示仅用于男性政治人物，I(w) = 0表示两性平衡。

1. 不同性别的关注焦点

•**男性政治人物**：媒体报道更多关注道德和行为特征

•**女性政治人物**：媒体不成比例地关注外表和社会经济背景

2. 性别特定的刻板印象

•**男性**：通常被描述为强大、积极和具有攻击性

•**女性**：经常被描述为不适合担任公职，更多关注她们的吸引力和身体部位

3. 父母表现差距

•提及女性政治人物的父母更为常见，暗示女性被认为是通过家族关系达到政治职位

•男性政治人物更可能被描述为父亲角色

4. 情感分析揭示的偏见

•与男性相比，关于女性政治人物的个人细节报道带有更负面的语调

•这种负面性在所有类别中（道德、身体、社会经济）都很一致

5. 媒体类型差异

•在线新闻来源比印刷报纸表现出更强的性别偏见

•在线媒体更倾向于使用负面个人化描述，特别是关于女性政治人物的身体特征

意义和影响

这项研究表明，政治报道中的性别偏见远不止是简单的代表性不足。即使女性政治人物得到报道，报道的性质也基本不同 ——更个人化，更关注外表，语调更负面。

这些发现特别重要，因为他们分析了几乎整个意大利媒体环境（印刷和在线），为政治报道中结构性性别偏见提供了全面证据。

研究人员表示，这种偏见可能有助于解释为什么女性在政治中仍然代表性不足，因为媒体描述可能会抑制女性的政治抱负，并强化关于女性适合领导角色的刻板印象。