**书生·浦语项目词云可视化小赛**

1. **小赛描述**

**任务描述：**将书生·浦语大模型实战营 500+ 项目的名称可视化为词云

**激励：**投票排名前三获得 50% A100 15 天使用权限

截止时间：2024年 5 月 23 日 18:00

**提交方式：**直接在本文档中第三部分编辑提交

**投票：**将编辑为公众号文章进行投票安排



**参考效果**

背景一：浦语小助手头像。

背景二：书生·浦语四个字。

词云素材：书生·浦语实战营所有项目，可以从[InternLM 项目登记表](https://aicarrier.feishu.cn/wiki/DoKWwqslwiyjjKkHhqJcvXAZnwd?table=tblyxy5MZV7gJ7yS&view=vew0rj0WuN)表格中获取。



2. **小赛交流群**



3. **作品提交**

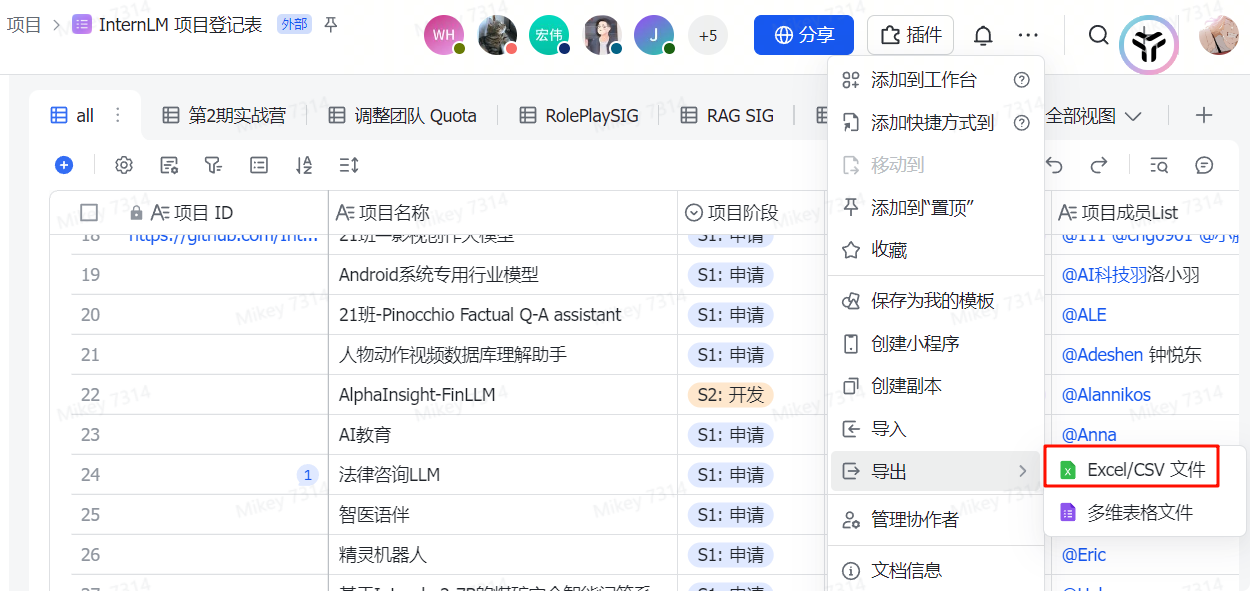
需要提交作品图片及复现方案，复现方案可先写在不可访问的文档。

3.1 **作品1：**

创作人：@Mikey

3.1.1 **导出数据**

首先打开[InternLM 项目登记表](https://aicarrier.feishu.cn/wiki/DoKWwqslwiyjjKkHhqJcvXAZnwd?table=tblyxy5MZV7gJ7yS&view=vew0rj0WuN)，将所有项目信息导出为csv格式，或者只复制项目名称，将内容保存为csv格式。



3.1.2 **代码实现**

我们需要用到pandas、jieba、matplotlib和wordcloud库，其中pandas用来读取csv文件中的数据，jieba用来拆分字符串，matplotlib用来绘图，wordcloud用来生成词云图样式。

|  |
| --- |
| Python import pandas as pd import jieba import matplotlib.pyplot as plt from wordcloud import WordCloud |

定义一个类用来对项目名称的词频进行处理，准备停用词用与去除无效字符，同时将一些额外的字符（如方括号、反斜杠、逗号）添加到 stop\_list 中，作为补充的停用词。

|  |
| --- |
| Python class *Action*:  def *\_\_init\_\_*(*self*, *file\_name*):  *self*.file\_name = *file\_name*  def *analyze*(*self*):  data = pd.*DataFrame*()  for file in file\_name:  data = data.*\_append*(pd.*read\_csv*(file))  # 对项目名称下的值进行处理，将其转化为str类型  data['项目名称'] = data['项目名称'].*astype*(*str*)  data\_cut = data['项目名称'].*apply*(jieba.lcut) *# 对项目名称进行分词*  with *open*('stoplist.txt', 'r', *encoding*='utf-8') as f: *# 停用词去除* stop\_list = f.*read*()  stop\_list = stop\_list + "[\]\," *# 补充停用词* data\_after\_stop = data\_cut.*apply*(lambda *x*: [i for i in x if i not in stop\_list])  data\_after\_stop.*to\_csv*('data\_after\_stop.csv', *index*=None, *encoding*='utf-8-sig')   data\_after\_stop\_flaten = [] *# 将数据展平* a = *list*(data\_after\_stop)  for i in a:  data\_after\_stop\_flaten.*extend*(i)  word\_freq = pd.*Series*(data\_after\_stop\_flaten).*value\_counts*()  return word\_freq |

定义App类用于绘制词云图，其中zhongguo.jpg为词云图的样式蒙版，simkai.ttf为词云图中的字体样式，在代码中设置背景颜色，字体大小等属性，使用双三次插值用来放大图像或图像的某个部分时减少失真和模糊效果。figsize可以调整图片大小用来确保图片清晰度。ImageColorGenerator用来控制字体颜色。

**[simkai.ttf]**

|  |
| --- |
| 注：由于中文下默认字体会显示乱码，所以需要将字体更换。 |

|  |
| --- |
| Python class *App*:  def *\_\_init\_\_*(*self*, *word\_freq*):  *self*.word\_freq = *word\_freq*  def *draw*(*self*):  *# 绘制词云图* mask = plt.*imread*('InternLM.jpg')  *# 创建词云图对象* wc = *WordCloud*(*font\_path*='./simkai.ttf', *mask*=mask, *background\_color*='white',  *max\_words*=500,  *max\_font\_size*=150, *# 减小最大字体大小*  *relative\_scaling*=0.6, *# 设置字体大小与词频的关联程度为0.4*  *random\_state*=50,  *scale*=2 *# 增加 scale 值* ) *# font\_path的相对路径*  *# 加载词频* wc.*fit\_words*(*self*.word\_freq)  image\_color = *ImageColorGenerator*(mask) *# 设置生成词云的颜色，如去掉这两行则字体为默认颜色* wc.*recolor*(*color\_func*=image\_color)   *# plt 绘出词云图* plt.*figure*(*figsize*=(20, 20)) *# 调整图片大小* plt.*imshow*(wc, *interpolation*='bicubic') *# 使用双三次插值* plt.*axis*('off')  plt.*show*() |

最后定义主函数，绘制词云图，效果如下：

|  |
| --- |
| Python if \_\_name\_\_ == '\_\_main\_\_':  file\_name = ['InternLM\_all.csv'] *# 文件名*  *# 实例化Action对象* action = *Action*(file\_name)  word\_freq = action.*analyze*()  *# 实例化APP对象* app = *App*(word\_freq)  app.*draw*() |



3.1.3 **gradio界面**

|  |
| --- |
| Python *# -\*- coding: utf-8 -\*-* *# @Time : 2024/5/19 13:35* *# @Author : 郭志航* *# @Site : https://blog.lingkongstudy.com.cn* *# @File : web\_ui.py* *# @Software: PyCharm*  *# @Comment : acwwt*  import gradio as gr import pandas as pd import jieba from wordcloud import WordCloud, ImageColorGenerator import matplotlib.pyplot as plt import numpy as np from PIL import Image   *# 定义一个函数来处理文件并生成词云* *# 定义一个函数来处理文件并生成词云* def *process\_files\_and\_create\_wordcloud*(*file*, *mask\_file*, max\_font\_size, *x*, *y*, *custom\_field*):  *# 读取 CSV 文件* data = pd.*read\_csv*(*file*.name)   *# 确保自定义字段存在* if *custom\_field* not in data.columns:  return "错误：上传的文件中没有找到字段 '{}'。".*format*(*custom\_field*)   *# 对指定字段进行分词* data[*custom\_field*] = data[*custom\_field*].*astype*(*str*)  data\_cut = data[*custom\_field*].*apply*(jieba.lcut)   *# 停用词去除（这里需要提供停用词文件的路径或内容）* with *open*('stoplist.txt', 'r', *encoding*='utf-8') as f:  stop\_list = f.*read*()  data\_after\_stop = data\_cut.*apply*(lambda *x*: [i for i in *x* if i not in stop\_list])   *# 数据展平* data\_after\_stop\_flaten = [word for sublist in data\_after\_stop for word in sublist]  word\_freq = pd.*Series*(data\_after\_stop\_flaten).*value\_counts*()   mask = *mask\_file*   *# 创建词云图对象* wc = *WordCloud*(*font\_path*='./simkai.ttf', *mask*=mask, *background\_color*='white',  *max\_words*=500,  *max\_font\_size*=150, *# 减小最大字体大小*  *relative\_scaling*=0.6, *# 设置字体大小与词频的关联程度为0.4*  *random\_state*=50,  *scale*=2 *# 增加 scale 值* ) *# font\_path的相对路径*  *# 加载词频* wc.*fit\_words*(word\_freq)  image\_color = *ImageColorGenerator*(mask) *# 设置生成词云的颜色，如去掉这两行则字体为默认颜色* wc.*recolor*(*color\_func*=image\_color)   *# 绘制词云图* plt.*figure*(*figsize*=(*x*, *y*))  plt.*imshow*(wc, *interpolation*='bicubic')  plt.*axis*('off')  plt.*show*()  plt.*close*() *# 关闭绘图以避免内存泄漏* return np.*array*(wc.*to\_image*())   html\_code = ''' <div> <h1 style="text-align: center;">🔎 书生·浦语项目词云可视化</h1>  <p align="center"> <a href="https://github.com/InternLM"> <img  src="https://intern-ai.org.cn/assets/headerLogo-4ea34f23.svg" alt="Logo" width="20%"> </a> </p>   <h2  style="text-align: center;">💪 Power by <a href="https://github.com/SmartFlowAI/EmoLLM/">InternLM</a>, If you like, please click a  little⭐ .</h2>  </div>'''  with gr.*Blocks*(*theme*=gr.themes.*Soft*()) as demo:  gr.*HTML*(html\_code)  input\_image = gr.*Image*(*format*="jpg")  output\_image = gr.*Image*(*format*="png")  gr.*Interface*(  process\_files\_and\_create\_wordcloud,  *inputs*=['file', input\_image, 'number', 'number', 'number', 'text'],  *outputs*=output\_image,  *examples*=[['InternLM\_all.csv', 'InternLM.jpg', 150, 20, 20, '项目名称']],  )  *# 启动应用* demo.*launch*() |

3.1.4 **在线体验地址**

**[This type of content is unavailable for download at the moment.]**

3.2 **作品 2**

作者：舒克

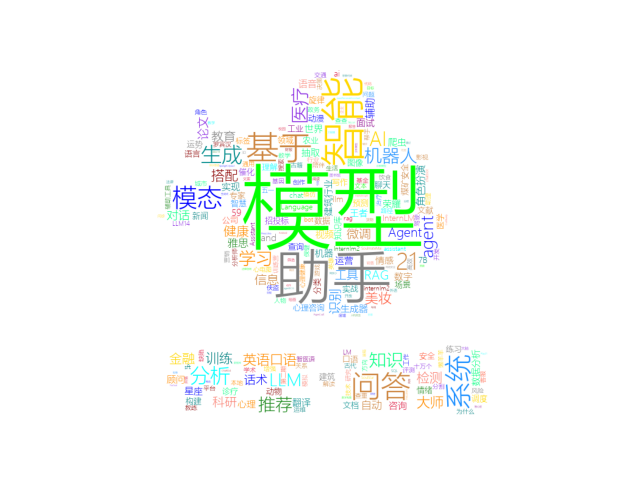


3.2.1 **代码实现**

稍后补充，先优化优化

3.3 **作品 3**

@axyzdong



3.4 **作品 4**



|  |
| --- |
| 作者：@张子杰  Code及方案文档：[书生浦语-词云活动-代码方案](https://bddg4y8wii.feishu.cn/docx/P2fWdeKMNoS12BxDITEcm9yVnBh?from=from_copylink)（打不开请申请权限） |

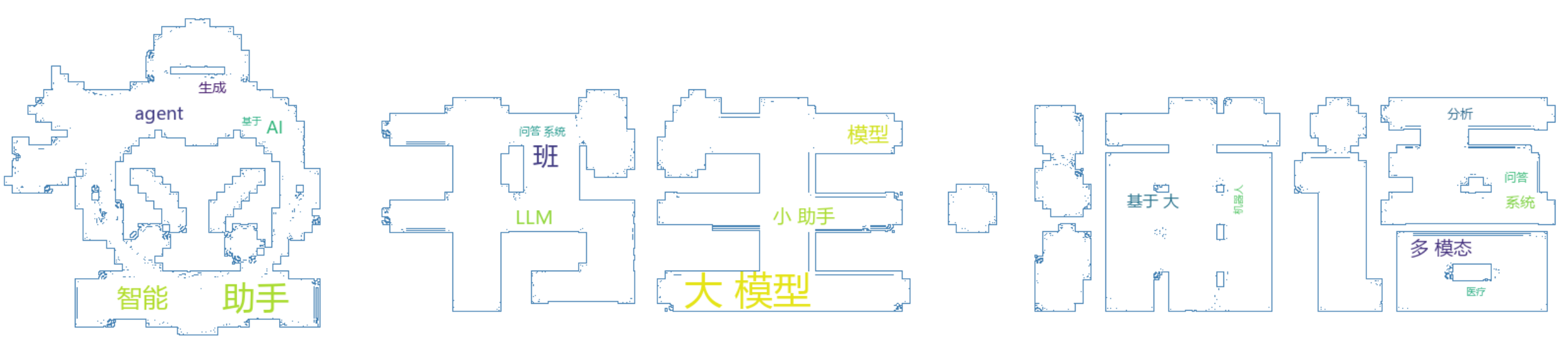
3.5 **作品 5**

|  |
| --- |
| 作者：@伍鲜 |

稍后...

3.6 **作品 6**

作者 @zky



3.7 **作品 7**

3.8 **作品 8**

|  |
| --- |
| 作者：@虞渊 |





3.9 **作品 9**

稍后上传

作者：@红亲王、(王泓清-Jeffrey)

3.10 **作品 10**

**3.11 作品11**

**3.12 作品 12**

作者@张进叶

|  |  |
| --- | --- |
|  |  |
|  |  |

4. **QA**

15 天权限 50% A100 为比赛后固定时间开启权限。

备注：书生·浦语运营团队对本次小赛具有最终解释权，最终根据实际的完成情况发放算力点。