**农业大数据应用课程报告**

（20 年春（）学期）

课程名称： 农业大数据应用 任课教师：

学 号： 姓 名：

**一、目标及意义**

本项目以“五一”假期文旅相关文本作为分析对象，运用Python语言及词云可视化技术。探索文本挖掘在农业大数据领域的实操方法，提升对数据预处理、分词、停用词过滤、词频统计、可视化输出等核心流程的理解与掌握。

此项目技术路径与农业数据挖掘具有高度一致性，通过项目的实施，不仅加深了我对数据清洗与词频分析等内容的理解，也为今后在农业新闻分析、农产品电商评论挖掘、以及各类相关产业中技术学习打下基础。

**二、过程分析**

项目主要包括文本读取、分词与同义词替换、专业词频增强、词云生成与展示等几个步骤。整体流程如下：

1. **文本准备**

选取了一篇有关“五一”假期文旅活动及AI模型发展方向的新闻稿，用作后续处理与词云可视化。

**2.词汇优化**

为了增强分词效果，提前通过 jieba.add\_word() 方法将如“文旅大模型”“AI智能体”等专业名词添加至分词词典，避免误切分。同时整理构建了自定义同义词表，如将“文化旅游”与“文旅”等进行归一，避免词云中出现重复词义的高频词。

1. **停用词处理**

加载停用词列表，过滤掉如“的”“了”“和”等对主题贡献较低的词汇。进一步设置长度筛选（过滤掉单个字），避免无意义信息干扰分析结果。

1. **词频统计与加权**

利用 collections.Counter 统计词频，并对重要专业术语进行额外加权处理（每个术语+20），确保其在词云中呈现主导地位。最终筛选前100个高频词汇参与词云绘制。

1. **词云生成与美化**

使用WordCloud库，指定字体路径（微软雅黑），设置背景颜色为“天蓝色（skyblue）”，生成词云图像。并设置 collocations=False，避免词云中自动组合词影响结果表达。

**三、项目代码及分析**

**1.添加专业词汇**

terms = [

'五一假期', '旅游热度', '智慧旅游', '文化和旅游消费周', '演艺潮流游',

    '文旅大模型', 'AI智能体', '智慧景区', 'AI+文旅', 'CBG模型'

]

for term in terms:

jieba.add\_word(term)

1. **读取停用词表**

stopwords = set()

with open(r'Z:/study/xuexi/stopwords.txt', 'r', encoding='utf-8') as f:

for line in f:

stopwords.add(line.strip())

1. **词云设置**

wc = WordCloud(

    font\_path='C:/Windows/Fonts/msyh.ttc',  #设置文字

    background\_color='skyblue',  #设置背景

   # mask=mask,

    max\_words=100,   #设置频率

    collocations=False   #拒绝自动组合

四、运行结果



**五、遇到的问题及解决方案**

**中文字体无法显示**

初次运行时词云中中文乱码，后通过指定 font\_path='C:/Windows/Fonts/msyh.ttc' 使用微软雅黑字体解决。

**专业词被错误切分**

例如“文旅大模型”被切为“文旅”“大”“模型”，影响统计准确性。通过 jieba.add\_word() 方法添加至词典，显著改善分词质量。

**背景颜色效果不佳**

默认深色背景与高频词颜色冲突，调整为 background\_color='skyblue' 天蓝色，整体更清爽协调。

**六、结论及总结**

本次实验以“五一”假期文旅热点文章为基础，经历文本分词、词频统计、同义词归一、图像生成等数据处理环节，最终成功生成一张主题突出、清晰美观的词云图。

通过这个项目实践，我不仅巩固了对Python基础操作，更掌握了农业大数据应用中最常见的词频分析技术。未来，可将类似方法拓展应用至农业政策文本、农户调研数据、农村新闻报道等领域，挖掘农业发展的潜在趋势、热点话题及舆情动态。