**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

通过学习python课程，学会编写词云程序。用数据可视化来分析旅游地区的前景和规划，用词汇云的方式来展现研究文档中的核心观点。为之后的发展方向给出基础建议。结合琼海市旅游产业的研究数据，和相关部门的旅游政策等。尽量在程序编写过程中用到所学知识，停用词表的分析生成，文档清洗，以及词汇云的频率颜色背景设定等。

学会词云相关技术对于大数据可视化分析以及清洗都有一定的帮助。而大数据经过这些年的发展，在我们的生活中开始扮演越来越重要的角色。对旅游，农业，食品，各行各业都意义深远。

# 二、过程分析

首先从网上下载到琼海市旅游相关的专业文档，删除数据表格和参考文档等，完成文档清洗。下载一张琼海市的轮廓地图，删除背景像素使其变成透明，以达到想要的词云形状。编写停用此表，类似‘一些’‘的’‘这样’‘那样’的词汇，过滤词云显示。添加一些专属名词，防止词云分析过程中自动切割专业文档中的部分专业词汇。

然后设置词云的背景，颜色，大小，以及过滤频率等。生成大小适宜的画布，生成词云，并且保存在本地。程序运行时也会将生成的词云打印在屏幕上。

三、项目代码及分析

import jieba

import numpy as np

from wordcloud import WordCloud

import matplotlib.pyplot as plt

from PIL import Image

from collections import Counter

from PIL import ImageOps

# 1. 开始预处理文本

# 读取专业论文内容

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

    text = f.read()      #打开该路径下的文本文件。

# 添加专业词汇

terms = [

    '体育+旅游+康养', '多元文化', '文旅体验', '博鳌亚洲论坛品牌', '前店后厂',

    '乐城康养、琼海度假', '乐城医疗健康', '全域旅游', '万泉河计划', '健康岛·快乐城'

]

for term in terms:

    jieba.add\_word(term)   #添加这些词，在后续分词过程中专业词汇不被切分

# 读取停用词表

stopwords = set()

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

    for line in f:

        stopwords.add(line.strip())  #读取该路径下的停用此表，一些数字，连接词都被收录停用切分统计。

# 分词 & 过滤停用词

words = jieba.lcut(text)

filtered\_words = [word for word in words if len(word) > 1 and word not in stopwords]

# 同义词替换

synonyms = {

    '体育+旅游+康养': '康养+旅游',

}         #将部分意思相近的词汇做一次同义词替换，以免词云中出现多频同义词。

final\_words = [synonyms.get(word, word) for word in filtered\_words]

# 统计词频

word\_counts = Counter(final\_words)

top\_100 = word\_counts.most\_common(100)

#2. 生成词云

# 加载琼海轮廓图

mask = np.array(Image.open(r'Z:/study/qionghai/mask.png'))

# 设置词云对象

wc = WordCloud(

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

    background\_color='white',  #设置背景

    mask=mask,

    max\_words=100,   #设置频率

    collocations=False   #拒绝自动组合

)

# 根据频率生成词云

wc.generate\_from\_frequencies(dict(top\_100))

# 3. 输出词云

plt.figure(figsize=(12,8))#创建一块画布，设置大小

plt.imshow(wc, interpolation='bilinear')

plt.axis('off') #隐藏坐标

plt.savefig('qwionghaiciyun.png', dpi=300, bbox\_inches='tight')  #保存图片

plt.show()

# 四、运行结果



# 五、遇到的问题及解决方案

在编写程序的过程中，主要遇到了三个问题。

一是停用此表词汇不完善，导致词云刚生成时，出现很多无意义的连接词和数词。解决方式是先生成词云，然后根据上面的无意义词汇去改动停用词表。

二是轮廓图刚载入时词云无法按照想要的形状生成，最后使用ps扣掉轮廓图背景像素，词云正常显示。

三是代码编写过程中，对于部分功能的不熟悉，一遍遍调试，最终成功。

六、结论及总结

通过项目的开发实践，我进一步掌握了文字处理、数据清洗、词汇云可视化生成等技能，并能够在实际问题中灵活应用。  
 词云作为一种直观、生动的文本可视化手段，在农业大数据应用中也具有重要的参考价值，例如用于分析农产品销售热点、农业政策舆情等。

在开发过程中，面对格式处理、图像处理、编码兼容等一系列问题，我学会了系统性地排查和调试程序，提高了独立开发和解决问题的能力。  
 本项目最终成功生成符合琼海市旅游规划相关词云图像，不仅提升了文档的可视化表现力，也为后续相关研究提供部分技术支撑。