上海对外经贸大学

Python 数据采集与挖掘

课程设计报告

学院: 统计与信息学院
专业: 人工智能
学号: 23089056
姓名: 孙琉屿

目录

1 引言 **2**

2 数据来源 **2**

3 爬虫设计 **2**

3.1 代码中使用的库: 2

3.2 爬虫流程: 2

3.2.1 函数 GetPresidentList(): 2

3.2.2 函数 GetPresidentImage(): 3

3.2.3 函数 GetPresidentDescription(): 3

3.2.4 函数 GetWordCloud(): 3

4 总结与展望 **4**

4.1 项目总结 4

4.2 未来展望 4

1 引言

本学期通过 Python 数据采集与挖掘这门课程学习了爬虫技术。本项目使用所学的知识，从白宫官方网站上爬取了关于美国历任总统的介绍，结合自然语言处理（NLP）技术对文本进行分词，并使用爬取到的总统图片生成词云图。本报告将简要介绍该项目。

2 数据来源

数据来源为白宫官方网站上的“Presidents”网页。
(`https://www.whitehouse.gov/about-the-white-house/presidents/`)

3 爬虫设计

3.1 代码中使用的库:

1. **requests 库:** 用于访问网页；
2. **lxml 库:**
   - 用于解析 HTML 文件；
   - 用于解析 XPath；
3. **os 库:** 用于创建文件或目录;
4. **re 库:** 用于执行正则表达式匹配和替换操作;
5. **urljoin 库:** 用于处理和拼接 URL 地址;
6. **rembg 库:** 用于从图像中去除背景;
7. **PIL 库:** 用于图像处理;
8. **io 库:** 用于处理内存中的文件；
9. **wordcloud 库:** 用于生成词云图；
10. **matplotlib 库:** 用于绘制图表和图像显示；
11. **imageio 库:** 用于读取和写入图像文件；
12. **nltk 库:** 用于文本处理和分析；
13. **nltk.tokenize 库:** 用于英文文本分词。

3.2 爬虫流程:

3.2.1 函数 GetPresidentList():

1. 从网站`https://www.whitehouse.gov/about-the-white-house/presidents/`上获取所有总统的姓名，并保存到列表 presidentList；

2. 将 presidentList 中的总统姓名的空格替换为'_'，并保存至列表 presidentList2, 用于后续爬取网页上各前总统的图像。调用函数 GetPresidentImage()；
3. 将 presidentList 中的总统姓名的空格替换为'-'，并保存至列表 presidentList3, 用于后续爬取网页上各总统的文字介绍。调用函数 GetPresidentDescription()；
4. 在图片和介绍都爬取完成后，调用函数 GetWordCloud()。

3.2.2 函数 GetPresidentImage():

1. 通过函数 GetPresidentList() 获取到的总统姓名列表 presidentList2, 构造总统图像的网页地址；
2. 将爬取到的总统图像并保存到文件夹 president；
3. 使用 rembg 库中的 remove 模块对总统图像进行抠图；
4. 使用 PIL 库中的 Image 模块将抠出的图像粘贴到一个纯白色背景, 并保存至原图所在目录，用于后续制作词云图。

3.2.3 函数 GetPresidentDescription():

1. 通过函数 GetPresidentList() 获取到的总统姓名列表 presidentList3, 构造总统介绍的网页地址；
2. 通过提前从网页中获取到的 XPath 爬取总统介绍；
3. 将爬取到的总统介绍并保存到文件夹 president。

3.2.4 函数 GetWordCloud():

1. 将之前获取到的总统介绍进行分词；
2. 生成词云图。

图 1: 华盛顿的词云图

4 总结与展望

4.1 项目总结

本项目通过 Python 数据采集与挖掘技术，运用多个技术以及库，爬取白宫官方网站上的美国总统信息，并利用自然语言处理技术进行文本分析，最后生成词云图。

4.2 未来展望

部分总统的介绍和图片未能成功下载，原因是白宫网站上的标题（总统姓名）与总统介绍地址以及总统图片地址上使用的总统姓名地址格式不相同，并且这种改变毫无规律。如特朗普总统的介绍页面标题是”Donald Trump”, 但介绍页面的网页地址为”*/Donald-J-Trump”；又如克林顿总统的介绍页面标题是”William J. Clinton”，但他的图片网页地址却是”*/42_bill_Clinton”。因时间有限，并未找到该问题的解决方法，希望未来能够解决该问题。