## Introduction

项目主要包括对Disneyland乐园的评论数据进行了全面的分析，首先通过Pandas加载了名为 **DisneylandReviews.csv** 的CSV文件，并使用 **cp1252** 编码进行解码。去除了包含 'missing' 值的行。去除了重复的评论。使用NLTK的SentimentIntensityAnalyzer对评论进行情感分析。定义了一个函数，根据情感分数分配 'Positive'、'Negative' 或 'Neutral' 标签。创建了一个新列 'sentiment' 存储评论的情感标签。并且使用Seaborn绘制了一个条形图，展示了评论情感的分布，使用Gensim进行了主题建模（Latent Dirichlet Allocation）。创建了一个词典和语料库，并使用LDA模型提取了5个主题。还有趣的使用WordCloud库生成了评论的词云，利用了迪士尼城堡的logo只做了遮罩图。最后使用Plotly Express制作了一个国家/地区评论分布的地图可视化

This is where you will give a detailed background to the project. You should use this section to give a broader context to the project. This includes a technical background on any of the methods you will be using, with citations relating to the techniques used. If there are any examples of related work to your project, it should go in this section (with citations). You can also use this section as an opportunity to give context and go into detail about the subject of the dataset you are using and any social, cultural, or critical perspectives relating to the data domain.

華特迪士尼先生說：「只要你有夢想，你就能做到。別忘了迪士尼的一切，是從一個夢想和一隻老鼠開始的。」作为全球最成功的IP制造巨头，迪士尼承载着几代人的童年与美梦。迪士尼乐园是全球最知名、最受欢迎的主题公园之一，其成功不仅仅体现在财务上的繁荣，更表现在对全球文化的深远影响。这一成功的背后有多个因素相互交织，共同构成了迪士尼乐园独特的魅力。

首先，通过深入了解游客对乐园的看法，我们能够获取对游客体验的直观认识。通过挖掘评论中的关键词和情感，我们可以识别游客最关心和赞赏的方面，这为乐园提供了改进和增强的方向。

其次，项目旨在为改进乐园体验提供有力支持。通过分析评分分布和情感分析，我们能够确定游客对乐园的整体感受，从而确定改进的重点。例如，如果负面评价集中在某一方面，比如服务或设施，乐园管理团队就可以有针对性地进行改善，提升游客满意度。

最后，通过对乐园评论进行深入分析，我们可以为乐园经营提供建议。这包括了解哪些特定方面受到了游客的热烈欢迎，以及哪些方面可能需要更多关注。这样的见解对于制定长期经营战略、推出新的吸引活动或改进特定服务领域都具有指导意义。

This is where you describe the work done. This may be:

* Technical details of how you built or implemented your project (we only need to know about the original code you have made)
* Details of any pre-processing to your dataset, including:
  + Manual data cleaning
  + Code to remove any unwanted text from your data
  + Stemming or lemmatization algorithms used
  + Stop words that were used
* A detailed summary of the experiments you have performed, including:
  + What algorithms were used
  + What parameters were used for your algorithm
  + Has the data been altered between experiments (comparing an algorithm using stemming vs lemmatization)
  + Any other changes made between experiments that you are comparing

#### 1.1. 数据加载和清理：

* 使用 **pandas** 加载 CSV 文件。
* 删除 **Year\_Month** 列值为 'missing' 的行。
* 删除重复的评论文本行。

#### 1.2. 情感分析：

* 使用 **nltk** 库的 **SentimentIntensityAnalyzer** 进行情感分析。
* 定义了一个函数来解释情感分数，并根据情感分数给评论分配标签（积极、消极、中性）。
* 使用 **seaborn** 绘制了情感分布的计数图。

#### 1.3. 主题建模：

* 使用 **gensim** 库进行主题建模。
* 预处理文本数据，包括分词和去除停用词。
* 创建字典和语料库。
* 构建 LDA（Latent Dirichlet Allocation） 主题模型。
* 打印了每个主题的前五个词。

#### 1.4. 词云生成：

* 使用 **WordCloud** 和 **matplotlib** 库生成词云。
* 读取了一张图像，并转换为只包含 RGB 通道的图像。
* 创建了词云对象，设置了背景颜色、遮罩图、颜色等参数。
* 生成词云图像并显示。

#### 1.5. 国家评论分布可视化：

* 使用 **plotly.express** 创建了国家评论分布的 Choropleth 图。

### **2. 数据预处理：**

* 将 **Year\_Month** 拆分为 **Year** 和 **Month** 列。
* 将 **Year** 和 **Month** 列的数据类型转换为整数类型。

### **3. 模型参数和实验细节：**

#### 3.1. 情感分析实验：

* 使用 **SentimentIntensityAnalyzer** 进行情感分析。
* 通过修改 **compound** 情感分数的阈值，将评论分为积极、消极和中性。

#### 3.2. 主题建模实验：

* 使用 LDA 模型，设置了主题数为 5，迭代次数为 15 次。

### **4. 词云生成实验：**

* 使用 **WordCloud** 生成词云，设置了背景颜色、遮罩图、颜色等参数。

### **5. 国家评论分布可视化实验：**

* 使用 **plotly.express** 创建 Choropleth 图，显示了不同国家的评论分布。

This is where you discuss the results of your work. Here you should give context to the results, can you explain why some of your experiments gave poor or unsatisfactory results? Why did one experiment work better than the other? Can you use your knowledge of the data domain to make sense of the results of your experiments. Can you describe what you would do to improve the results given more time?

The results of your work don’t always have to be perfect. Working with new datasets can be unpredictable. The discussion section is your opportunity to reflect on what worked well and what could be improved. This is your opportunity to show your understanding of the algorithms you have used and demonstrate **critical reflection.**

美国迪士尼公司今年迎接创业100周年，从电影、电视节目到主题乐园，百年来持续以神奇魅力吸引着儿童与成人。迪士尼在全球的6间迪士尼乐园持续填补亏损，每年造访人数合计超过1亿人次。迪士尼[主题乐园](https://www.secretchina.com/news/gb/tag/%E4%B8%BB%E9%A2%98%E4%B9%90%E5%9B%AD" \t "/Users/abc/Documents\\x/_blank)在今年4到6月的营业利润约为24亿2500万美元，占总利润的70%。百年间，迪士尼努力地守护着大家的童年，但与此同时，这家“百年老店”也在经历着它自己的“老年危机”。近年来，迪士尼在全球不同区域范围内，出现了不同程度的裁员、亏损、票房惨淡、乐园门票涨价以及Disney+付费会员的流失等问题。所以了解游客对于迪士尼乐园的评价和看法对迪士尼有重要的作用，它不仅能帮助企业跟踪游客的意见，还能帮助企业发现提升乐园活力的方法，我希望通过情感化分析和主题建模等方法，对这些评论进行分类，找到共同的主题，理解游客关注的方面，以及可能的热门话题，帮助我更全面地了解游客的反馈，为迪士尼改进服务、提高游客体验以及制定营销策略提供有价值的见解。希望迪士尼能在下一个一百周年继续为来自全世界的观众提供更美好的童年和生活。

但是很可惜的是，我对该数据集的分析仍然有很多缺憾。

目前收集到的评论只持续到了2019年，而这三年经过疫情等全球不定因素的影响，数据一定会有一些起伏波动，这会使得分析的结果不够真实。同时在情感分析部分使用了随机生成情感标签，我想要使用更复杂的情感分析模型，例如深度学习模型，以提高情感分析的准确性，但由于技术力的问题无法实现。在分析的过程中，可视化部分也不够全面，无法帮助更清晰地传达我的分析结果。