需求分析

旅游业是一个综合性很强的信息依赖产业，旅游信息的获取，加工，传播和利用对旅游业的发展起着举足轻重的作用，从旅游者和服务提供方面的需求出发建立旅游信息数据库并进行可视化展示，不仅可以使旅游者能够直观的参考利用已经得出的数据统计结果，还可以进行快速，准确的查找和检索自己需要的旅游信息，促进旅游信息的共享，打破旅游信息的封锁。

在爬取携程网站上的游记以及旅游线路数据后，本小组计划在本次实训项目中实现如下数据分析和功能：

1. 基础功能

1.1 不同城市的平均评分

在爬取下来的数据里面，有如下一项数值值得分析：一座城市的平均评分，对于游客来说，选择出游的很重要一点就是基于以往旅游者的先验经验。本小组对数据进行处理，得到一座城市的平均评分，因为选取的数据量比较大，一定程度上可以去除个体的差异性，得到一个比较公正可信的结果。

将每座城市的平均评分按照由高到低的顺序进行排序，在前端以柱状图的形式进行展示，横坐标为每个城市对应的分数，纵坐标为城市名称，此种受欢迎度信息可以在一定程度上给使用者在选择出游的时候提供参考

1.2 对于不同城市的行程安排，导游服务，酒店体验评分分析

由于顾客的个体差异性，每个顾客在选择出游时，都有自己的偏好性和口味，对于不同方面的要求不尽相同。所以本小组根据爬取的携程数据，得到每个城市所有的旅行线路中关于评分部分的不同方面，有行程安排，导游服务，酒店体验这三项。这样，就将评分进行了细化，游客在选择出游时参考的因素也会更加全面。

在前端对于每个城市的行程安排，导游服务，酒店体验这三项的平均成绩，进行可视化展示，展现形式为柱状图，横坐标为城市名称，纵坐标为上述三项的平均成绩。

1.3 不同城市旅游价位分析

对于游客来说，选择城市旅游的因素除了城市受欢迎指数，服务好坏之外，性价比也是值得考虑的一点。本小组得到每个城市的旅游线路消费平均水平，对其进行高到低的排序，在前端进行展示，展示形式为柱状图，横坐标为城市名称，纵坐标为城市评分。同样地，该分析会为游客出行提供一定的参考

1.4 不同城市酒店偏好

去每个城市旅游的重要一点还有城市酒店的选择。本小组计划统计每个城市的旅游景点中比较多被顾客选择的酒店，在前端以饼图的形式展示出来。以每个城市为分组，可以浏览每个城市的比较优质的酒店。

1.5 旅行线路统计分析

根据爬取的旅行线路，统计所有从北京出发到各个热点城市旅游线路数量，进行可视化展示。在展示形式上，采取了绘制地图的方式，地图的展示方式较为直观，观者可以较为清晰地感受到热门城市的分布。除此之外，本小组绘制了柱状图，展示从北京出发到各个城市线路的数量。

1.6 出游伙伴的分析

旅游业主要是凭借旅游资源和设施，专门或者主要从事招徕、接待游客、为其提供交通、游览、住宿、餐饮、购物、文娱等环节的综合性行业。而游客选择的出游伙伴不同，势必面临着在服务提供方面的差异性。因此对于商家或者服务方来说，分析游客出游伙伴的类型是很有必要的。根据出游伙伴地不同，商家或者服务方可以更加有针对性地提供服务，提高该城市的竞争力。

本小组将爬取的不同的旅游路线作为整体分析，对出游伙伴类型计次，在前端以饼图的形式展示。

2. 拓展功能

2.1 基于游记的旅游线路推荐功能

筛选出代表性的游记，方便用户获取更具参考性的评价信息；以游记文本作为主体，通过相关算法获取相应热门景点，并基于此生成推荐的旅游线路，帮助用户在有限的时间内规划更能展现城市魅力、更有意义的行程。

通过爬虫获取游记文本，对文本进行预处理，包括设定自定义词表、分词和去停用词等，将文本向量化，然后构建LDA主体模型，得到游记数据集中的主体概率分布，并对所识别的高频特征词进行分析描述，形成文本主题的相关标签；最后通过对每篇游记于用户根据需要设定的标签的描述度、忠诚度和代表度的计算，得到相应标签的代表性游记及相关内容。依据篇频次对热门的旅游景点进行提取，对热门旅游景点之间的路线关系进行计算并聚类，然后进行中心性分析，依据结果生成推荐的旅游线路。

2.2 自动生成旅游文案功能

经历一天充实的旅行后，许多人都迫不及待地想将自己的所见所想分享给身边的人。但你一定遇到过这样的情况：一天的奔波使得对许多旅游细节的记忆已经模糊，无法细致入微地写出当时的所见所想；抑或别人的朋友圈文案都是“落霞与孤鹜齐飞，秋水共长天一色”，而自己却感受到“只能意会，无法言传”的词穷。

本小组的旅游文案生成功能就是瞄准了这一痛点，爬取大量旅行游记数据，通过分词、去除无用词句等数据处理操作，最终将其作为长短期记忆神经网络（LSTM）的数据集进行训练，生成模型。用户可以通过输入一些语句来得到一段自动生成的旅游文案，为记忆模糊、词穷等燃眉之急提供灵感。

2.3 旅游线路聚类分析

现如今，各大旅游网站都有海量的旅游线路供用户选择，但过多的线路也对用户造成了困扰：信息量过大，常常混淆不同线路，无从下手选择，为旅行增添了不必要的麻烦。本小组通过对大量数据的观察、比对，发现导致这一通点的主要原因是存在大量的相似线路，它们分布零散，不易记忆。

本小组通过对旅游线路的行程信息分词处理，再与爬取的景点名称匹配，得到具体旅游线路的景点，将其聚类分析，将相似的旅游线路归为一类，帮助用户解决困扰。