论文选题 & 前段时间总结

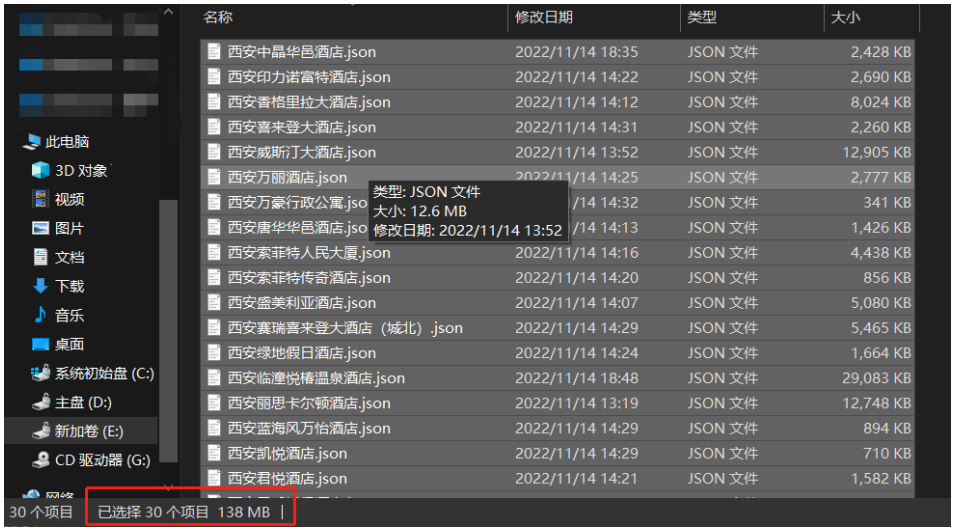
张锦羽

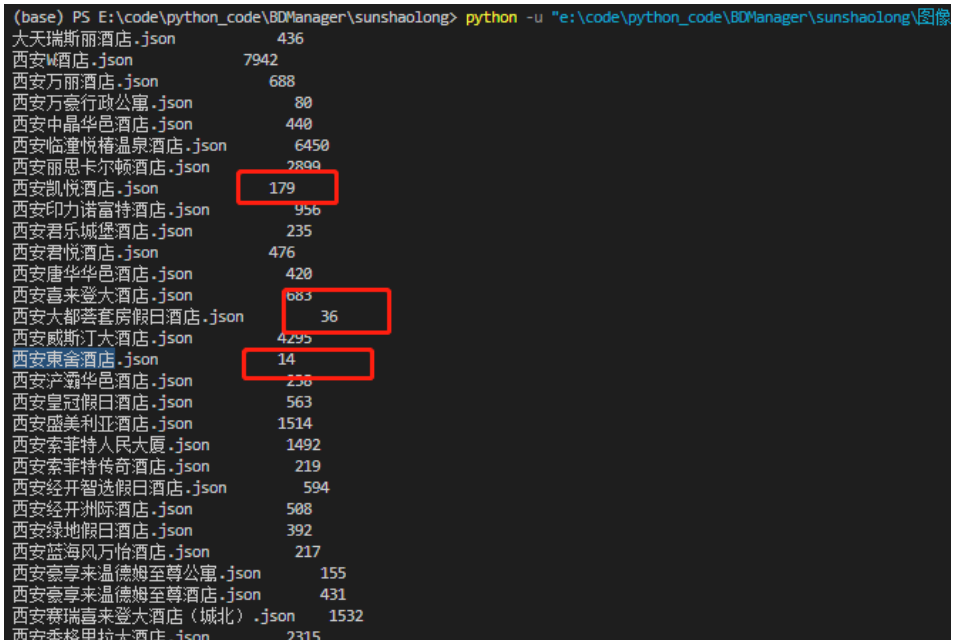
西安交通大学管理学院

2022 年 12 月 9 日

目录

猫途鹰网站爬虫





我基本上把能获取到的信息全爬下来了，而且猫途鹰网获取数据非常方便，我找到了接口，可以直接下载 json 数据，根据论文需要我还可以获取更多的数据。

1. 用户回复文本内容、用户回复时间、用户注册时间、用户附的照片、用户的评价星级（对酒店的打分）、用户出行方式、用户的入住时间、用户评论发布时间....
2. 酒店经纬度、酒店回复内容、酒店照片、酒店回复时间
3. 景点的经纬度、评论还有美食的经纬度、评论

发现几个有意思的地方：用户评论中都是包含有一个标题属性的：

标题可能是下面评论的概括，也可以是评论信息中用户最想突出的部分。能否通过评论标题预测用户的打分倾向？

A	B	C	D
	name	lat	lng
0	西安丽思卡尔顿酒	.2308587098764	8.900503933175
1	安临潼悦椿温泉酒	4.3680076256560	9.185996193375
2	西安威斯汀大酒店	4.221519761109	8.968729735858
3	西安皇冠假日酒店	.2385740159870	8.941762929466
4	西安W酒店	4.2028069906490	8.994968643622
5	西安经开洲际酒店	4.3457100612998	8.951594826661
6	西安盛美利亚酒店	4.2141549379263	8.982925996211
7	西安香格里拉大酒	4.2392133670790	8.904420773617
8	豪享来温德姆至尊	4.2120982822358	8.972293909070
9	西安浐灞华邑酒店	4.3503459485906	9.049206960355
10	西安唐华华邑酒店	4.2237348966594	8.975044730573
11	西安索菲特人民大	.2711833495250	8.966181410397
12	安经开智选假日酒	4.3440645925200	8.950608756319

酒店，景点，美食，三者的经纬度和评论都可以获取。是否可以作出一些东西呢？（以前打数模的时候做过西安市房价预测的题目，当时就是爬取了一些二手房信息利用百度地图 api 可视化出来）

- 由于一部分用户可能只是偶尔使用该平台，但是可能有一些用户是该平台的深度用户。使用全部用户评论对酒店进行评价时，需要不需要考虑不同类别用户之间的权重？
- 有些用户注册比较早，有些用户注册晚，注册时间会不会对两种用户有影响（比如同一用户分为后疫情和前疫情去研究？但是如何控制变量得到疫情影响的因果效应？）

我想做的方向是偏向统计学 + 机器学习方法 + 实证方法做出一篇文章，对标的期刊应该是**运筹与管理、管理工程学报**。
该文章的大概流程应该是：引言，综述，实验部分（数据获取、数据探索、特征工程：**机器学习方法**），实证部分（探究政策，时间，地理等因素，某变量对某变量的影响），管理学意义。

1	基于机器学习的Airbnb房源价格预测及影响因素研究——以北京市为例	毕文杰; 扶春娟	运筹与管理	2022-09-25	期刊		396

	题名	作者	来源	发表时间	数据库	被引	下载	操作
1	管理者回复对在线评论与有用性关系的调节效应:基于TripAdvisor的实证研究	陈远高; 应梦茜; 毕然; 杨水清	管理工程学报	2021-06-22 14:29	期刊	2	1371	

- 做评论中文本和图片相关的实证分析。但是我题目没想好，第一篇论文感觉自己把握不好，选题太大或者太小没概念。而且太小了感觉又太简单没什么意义...
- 我发现可以爬到一些空间的数据，而且可以爬到西安市 top 排名的酒店和 top 排名的景点。我觉得很有意思，可以和老师讨论一下。如果能做运筹与优化方面的我也挺感兴趣的。

中国旅游大数据研究：二十年回顾与展望

本文是类似于综述的一篇文章，回顾了中国旅游大数据发展的 20 年来，知网上核心期刊的发刊数量以及研究领域等。文章根据旅游大数据中数据类型的不同，分为了不同的研究方法，我在阅读这篇文章的过程中联系了我的研究问题，将注意力重点关注到了 UGC 数据的研究方法中。

中国旅游大数据研究：二十年回顾与展望

本文中将旅游大数据的数据类型分为了三类：UGC(User Generated Content) 数据、设备数据（GPS、气象、智能穿戴设备）、事务性数据（网页搜索、网页浏览、在线预定数据）。
根据获取难度，UGC 数据是最容易获取的，设备数据和事务性数据一般是私密的，难以获取的。

旅游大数据 UGC 研究数据：

数据来源 Data source	数据类型 Data type	研究主题 Research topic	数据特点 Data characteristics	分析技术 Analytical methods
UGC 数据 UGC data	在线文本数据	游客满意度、旅游目的地形象、旅游意象、旅游体验、旅游情感分析、旅游流、旅游推荐	数据来源多样、低成本、适用范围广；表达旅游者对旅游产品的态度和体验	确定数据源，数据收集，数据预处理（数据清理、转化、分词、词库），数据挖掘（内容分析、扎根理论、IPA 及深度学习、机器学习等）
	在线图片数据	旅游目的地形象、旅游体验、旅游者时空行为、旅游兴趣点挖掘、旅游流、旅游推荐	低成本、包含多元数据、数据来源多样	内容：内容分析法、隐喻抽取技术或符号学相关分析方法、社会网络分析法、深度学习算法中的深度卷积神经网络和残差神经网络等。 地理标签：核密度估计和空间聚类分析法、GIS 空间分析方法
	微博签到数据	旅游者时空行为、旅游流及结构演化、旅游情感体验	低成本、信息量大、互动性强、结构复杂；适合较大范围研究	数据清理、时间分层法、核密度分析方法

我获取到的主要是 UGC(User Generated Content) 数据（用户在线评论、用户配图、用户评分等）。此外还有一些事务型数据：如开源的百度指数 api 可以调用。

- **文本数据的处理流程：**
  数据清洗，去重去空 $\Longleftrightarrow$ 数据转化，同义词 $\Longleftrightarrow$ 分词，词向量化 $\Longleftrightarrow$ 评论分类，情感分析等
- **图片数据的处理流程：**
  内容分析法、cv 和图像处理辅助图片内容识别及分类。
  图片地理空间分析（拍摄图片的位置、时间等主要是一些图片分享的网站有相关的接口，tripadvisor 无此类接口）。

科研任务总结

总结

1. 确定了猫途鹰爬虫的难度不大，数据获取非常容易。探索了一些能够获取到的数据类型。基本上平台上有的都能获取到，非常方便。
2. 选择了一些重点的数据进行思考：评论中包含标题信息和图片信息，也有用户评星。研究影响用户评价的因素？个体异质性？是否可以找到一些明星用户，比较他们对于不同酒店的评价，缓解个体异质性。给酒店管理决策？
3. 用户 id 与酒店 id？我可以获取到一个用户所有的评论数据。可以根据评论的时间和地点知道用户的活动轨迹。是否可以依据一些明星用户的轨迹做研究。（定义明星用户，半年内评论超过 xx 次的）
4. 图片数据不知道怎么处理.... 我对 cv 还是小白，但简单操作或者调包跑模型应该不在话下的，这方面我学的很快，只要有明确的方向。
5. ....