Ⅰ. 论坛

1. **目的**：用于挖掘“用户和关注的人”之间的潜在的联系

2. **方法**：

a. 进入汽车之家论坛精选（https://club.autohome.com.cn/jingxuan）在“搜选精选”左侧白框输入“路虎”或“捷豹”。

b. 以“路虎”为例，取前 10“楼盖得最高的帖子”，等同寻找“门槛值找楼高的帖子”尽量确保这 10 楼有不同的“主题”，以暂定 10 个不同的“关键词”。

c. **爬虫内容**：

- 主帖的内容
- 回帖的内容
- 主帖&回帖用户的个人信息：用户名，性别，生日，所在地，关注（车型），身份
- 主帖&回帖用户的其他信息：等级，注册时间，里程值，帖子发布量（精华帖数量&内容，帖子数量&回帖量），‘他的车库’，‘他的口碑’，‘他的油耗’，‘他的车友圈’。

爬虫的目的：

- 挖掘用户的“星级”，探寻有无“意见领袖”的存在（类似微博的大 v）并提取相关特征
- 每栋“楼”划为独立的 group，每个 group 里发帖的用户视为样本，探寻每栋楼的用户之间关系的强弱，也就是影响力的强弱
- 任意点开用户的个人信息，‘他的车库中’如有路虎,捷豹的照片，视为‘已购买者’；若有其他品牌的照片，视为‘未购买者’；若无任何照片，同视为‘未购买者’。

Ⅱ. 车商城 （https://mall.autohome.com.cn/index.html）

1. **爬虫内容**

路虎+捷豹：分别按车型（每个车型分别建立文档），进行数据爬虫（车型，颜色，报价，提车地区，购买用户的评价，评价的星级，追加）并分类。

爬虫的目的：

- 车商城中的用户并不能和论坛、口碑中进行关联，所以要分别进行爬虫
- “车商城”中进行评价的用户，都是已购买该车型的消费者。用户的个人信息并不公开，只显示购买该款车的颜色，时间，购买评价和星级，以及距离已评价的天数。
- 没有“未购买者”的评价

III. **口碑**（https://k.autohome.com.cn/)

1. **爬虫内容：**路虎+捷豹：分别按车型（每个车型分别建立文档），进行数据爬虫并分类。

- 购买车型，购买地点，购车经销商，购买时间，裸车购买价，油耗，目前行驶
- 汽车星级评价：空间，动力，操控，油耗，舒适性，外观，内饰，油价比
- 购车目的
- 该口碑的支持量（多少人支持该口碑）
- 该口碑的阅读量（多少人看过）
- 该口碑下面的回复内容和回复数量
- 口碑用户的个人信息：用户名，性别，生日，所在地，关注（车型），身份
- 口碑用户的其他信息：等级，注册时间，里程值，帖子发布量（精华帖数量&内容，帖子数量&回帖量），‘他的车库’，‘他的口碑’，‘他的油耗’，‘他的车友圈’。

爬虫的目的：

- 车商城中的用户并不能和论坛、口碑中进行关联，所以要分别进行爬虫
- 发布“口碑”中的用户，都是已购买该车型的消费者，视为“已购买者”。回复该则口碑的用户，视为“未购买者”
- 可探寻论坛和口碑中用户的强弱联系，该则口碑对未购买者购买行为影响力的大小。