数据预处理：

数据清洗：1.删除重复评论2.删除评论字数过少的评论（字数多少看分布）不需要说太多

数据分布观测：评分的分布，是否存在数据偏斜问题（检验和解决）

**a:**

数据：使用的数据是三个数据集（不区分产品品牌）

模型1：可信度使用三个指标（四个字段help\_votes/total\_votes/vine/verify\_purchase）选择具有较高可信度对评论进行筛选，选出的评论作为训练集的自变量，选出的评论对应的评级(star\_rating)作为因变量，进行嵌入词向量模型训练，得到每个词对应的向量，然后根据嵌入词向量的模型的向量计算所有评论的分数（这里为求解），从而将产品的对应的评论转化为分数。

选择指标：使用不同降维方法(PCA,kernelPCA等方法)将评分和评论对应的评分二维数据降维为一维，然后根据生成的一维数据表示二维数据的百分比高低，选择一个最能表达的一维数据作为指标。然后通过加权的方式将数据转化为1个数。

**b:**

**先总体看三种产品中的每个产品，然后对于不同产品中的不同类进行分析（）**

数据：使用的数据是三个数据集

输入：时间、综合性指标(a)（这里的是一个区间的综合成为的一个值） 输出 时间、综合性指标(a得出指标这里的是一个区间的综合成为的一个值)

模型2：ARIMA和LSTM时间序列模型

时间窗口的大小：按季度或按年

作图：横坐标是时间（季度或年数）纵坐标是a问得出选择指标的大小

X季度，y一个评价指标

**c:**

关键词提取（模型）（情感类词还是功能类词），提取最为关键的词（可以分为时间和产品类型讨论），找关键词占根据第一问得到指标分等级后的评论的占比，关键词在分等级后评论最多，打分（0-1）区间，然后在对产品的品牌的评论数目（热度），加上关键词在产品评论的评分。

**d:**

Y轴所有评论的数目

X轴时间

做图求所有拐点(可以通过斜率做箱线图选择拐点) 然后根据数据分析然后归以下四类

1.低星级-》多写好评论

2.低星级-》多写坏评论

3.好星级-》多写好评论

4.好星级-》多写坏评论

**e:**

对c题中的找到的关键词，然后更加a题的嵌入词向量对关键词进行聚类，找到聚类中心的词，然后根据聚类中心的词，然后根据聚类中心的词，得到里面含有的该词的评论，然后根据评论得到几条评论的平均打分（star\_rating）值，然后根据平均打分和打分区间的可信度。