**区别2：英文单词有多种形态**

英文单词存在丰富的变形变换。为了应对这些复杂的变换，英文NLP相比中文存在一些独特的处理步骤，我们称为词形还原（Lemmatization）和词干提取（Stemming）。中文则不需要

词性还原：does，done，doing，did 需要通过词性还原恢复成 do。

词干提取：cities，children，teeth 这些词，需要转换为 city，child，tooth”这些基本形态

C题是个综合题，多元统计分析、聚类、时间序列，优化，数据挖掘都可能涉及。大家可看看当年我的汽车售后数据处理以及高速公路路面评价那篇论文。

这是个总数据，首先第一步应把具体型号的数据拎出来（product\_parent）。

（1）论文的第一部分应该有个数据的初步分析。每种产品，以及参数间的时间分布特性（有无趋势性，周期性），是否是平稳的，能否用ARMA或B样条长期预测？产品间，指标间有无相关性（主成分分析法），参数间能否提取特征用聚类算法描述（K均值聚类）

（2）问题d是因子分析的内容，可参看我当年的高速公路评价那篇建模。

（3）问题e是事件挖掘（语义挖掘）的内容，可参看我师兄赵城利的《事件挖掘综述》（还有他的博士论文），从语义链评价产品的好坏。根据评论也可大致分为几个等级，一般来说应和评分具有相关性。

用相关系数衡量。