**数据说明：**

商品分类，顾名思义就是根据商品的性质、特点将其划分到合适的类别中。在现代商业社会，商品分类是商品流通的基础步骤之一。目前的商品分类普遍存在三大困难：数据大、信息少、人工繁。本案例针对上述商品分类面对的困难，提出了一套基于商品名称的商品自动化分类体系，它可以完全利用商品名称信息，以自动化的分类技术大批量高效率的对海量商品进行分类处理，从而延伸商品价值。数据集“catalogs.xlsx”中共包含四个字段，分别是商品名称、商品一级品类、二级品类和三级品类。此外，还数据中还提供了一个词典“userdict.dat”供分词时使用。

**分析任务：**

1. 读入catalogs.xlsx，并使用head()函数查看数据形式。
2. 使用ggplot函数，分别绘制“生鲜”和“食品饮料、保健食品”两个一级品类下二级品类分布的饼图，并进行解读。
3. 进行商品名称的处理，使用jiebaR包中的分词函数对商品名称进行分词处理。为了更精确的分词结果，我们提供了一个自定义词典，即“userdict.dat”，请再自定义撰写一个包含5个停用词的“stop\_words.dat”，加载到分词worker中，对全部商品名称进行分词。
4. 基于上一题的分词结果，使用apply族的函数分别计算 “生鲜”和“食品饮料、保健食品”两个类别下，每个商品名称分词后单词的数量，并以此使用ggplot函数绘制两个一级品类商品名称分词后单词数量的直方图。
5. 分别绘制“食品饮料、保健食品”类别下三级品类“牛奶乳品”和“成人奶粉”商品名称分词后的top100高频词的词云图，比较两种品类商品名称的语言特点，并做简单点评。
6. 对分词结果进行预处理，预处理包括去掉纯数字或者纯字母的组合，去掉低频词（包含该词的商品数量<10）、高频词（在多于75%的商品名称中出现），在去除低频词和高频词时，可以使用函数pruned\_vocab()。最后，保留高频词中的top1000的词语，作为分类器的特征，并建立文档-词频矩阵。在建立文档词频矩阵时，需要用到text2vec包中的create\_dtm()函数。
7. 挑选出“食品饮料、保健食品”类的样本，构建朴素贝叶斯分类器。固定随机数种子为666，按照7:3将这部分样本划分为训练集与测试集两部分。使用e1071包中的naiveBayes()函数构建朴素贝叶斯分类器，并计算测试集上的预测准确率。
8. 使用混淆矩阵的形式展示上一问在训练集和测试集上的混淆矩阵和对应的准确率，并做简要点评。