第1章　R语言简介

1. 请简要描述数据分析的过程。

【参考答案】数据分析是依据明确目标，使用适当的统计分析方法和工具，对收集到的数据进行处理分析，从数据中提取有价值的信息并形成结论的过程。一个典型的数据分析过程包含五个步骤：明确目标、数据收集、数据处理、统计分析、形成报告。

首先需要根据需求明确分析目标，并依据这个目标有组织有针对性地收集各类原始数据，接着需要对收集到的数据进行加工处理(包括清洗、缺失值填补、变换等)，然后分析人员对处理好的数据进行统计分析及建模等，结合业务目标得出结论，最终形成报告，以帮助管理层作出判断并采取适当行动。

2. 请简述数据分析和数据挖掘的联系与区别。

【参考答案】联系：数据分析可以分为广义的数据分析和狭义的数据分析，广义的数据分析就包括狭义的数据分析和数据挖掘，通常意义上的数据分析是指狭义的数据分析。

区别：

(1)作用：数据分析主要实现现状分析、原因分析、预测分析(定量)。数据分析的目标明确，先做假设，然后通过数据分析来验证假设是否正确，从而得到相应的结论。数据挖掘主要侧重解决分类、聚类、关联和预测(定量、定性)等问题。数据挖掘的重点是寻找未知的模式与规律，这些模式和规律通常是事先未知的，但又是非常有价值的信息。

(2)方法：数据分析主要采用对比分析、分组分析、交叉分析、回归分析等常用分析方法。数据挖掘主要采用决策树、神经网络、关联规则、聚类分析等统计学、人工智能、机器学习等方法。

(3)结果：数据分析的结果通常是一些指标的统计量，如均值、方差、中位数等，且指标数据需要结合业务进行解读，才能发挥出数据的价值与作用。数据挖掘的结果通常是输出模型或规则，以及相应的模型得分或标签等。这些模型和规则表示事先未知的且有价值的信息和知识，如流失率、相似度、预测值等；标签如用户类别、信用等级等。

3. 请参照示例1-3，选取合适的中文文本(比如宋词、短篇小说等)，生成词云。

【参考答案】方法提示：可参照代码1-7。先选择中文文本保存为txt文档，然后依次进行分词、统计词频，最后依据词频生成词云。需要注意，为了避免词云效果太差，需要设置适当的词频(示例1-4的条件为词频大于5)。