HOMEWORK #2（10分）

一、数据预处理和可视化（8分）

1) 新闻数据读入与建立数据框对象(data.frame)。每篇新闻都有多个属性，不必全部保留，但是要求数据框对象中的属性至少包括全文、类别、时间，缺失的用 NA 填充。（hint: library(xml)）
2) 对新闻全文进行预处理，包括去除标点符号、停用词、数字、空白字符，将大写字母都转化为小写，以及词干化处理。（hint: library(tm)）
3) 将每一篇新闻的全文表示成 BagOfWords 向量。
4) 考虑单词在所有新闻中的出现次数。给出出现次数最多的 100 个词并对这些词画出“云图”。（hint: library(wordcloud)）
5) 给出单词长度的分布情况并画出柱状图，每个单词只算一次。(hint: library(ggplot2)）
6) 考虑新闻全文的单词数，分别使用等深分箱和等宽分箱将所有新闻分成 10 个箱，并画出每个箱包含的新闻数量的直方图。
7) 给出每一个类别下的新闻数量的分布情况并画出柱状图。
8) 给出每个月的新闻数量的分布情况并画出柱状图。

（编程语言不限，建议使用 R 或 Python）

新闻数据说明：500 篇新闻数据在 nyt_corpus/samples_500/目录下。每一篇新闻都以 xml 格式存储，部分属性说明如下：

- **全文属性：**在<block class="full_text">节点下
- **时间属性：**在某些 meta 节点下，时间包括年、月、日三部分，用 meta 节点的 name 属性来标识。如出版年份为 <meta content="1987" name="publication_year"/>，出版月份为 <meta content="1" name="publication_month"/>，可以看到 name 属性标识了年和月，content 属性标识了具体的年份和月份。三部分都要求保留到数据框对象中。
- **类别属性：**在某些 classifier 节点下，当 classifier 节点的文本内容为 Top/News/xxx/...或 Top/Features/xxx/...时，取出 xxx 作为该新闻的类别之一。

例如：<classifier class="online_producer" type="taxonomic_classifier">
Top/Features/Travel/Guides/Destinations/North America/United States
</classifier>是某篇新闻的一个节点内容，则将 Travel 取出作为它的一个类别。注意一篇新闻可能有多个类别。

提交说明：提交源代码、报告和 README。报告中说明每一步的结果，要求画图的题目要把生成的图贴出来。README 中说明如何运行你的代码。

二、高维向量可视化（2分）

当我们希望对高维数据进行可视化时，常常会先对数据进行降维处理，主成分分析（Principal components analysis，PCA）是一种经典的降维方法。t-SNE 是一种近年来很受欢迎的降维方法，在很多情况下它比 PCA 表现得更好（对算法本身感兴趣的同学可以参考论文：http://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf）。它的缺点在于参数较多，当参数的设定不适合数据集时，可能会得到较差的结果。

在附件 100_word_vector.txt 中，包含了 100 个英文单词及通过 word2vec 得到的向量表示，向量长度为 100，每个单词占一行，格式如下（不同维度之间用空格分开）

<word>\t<dim0> <dim1> ..... <dim99>\n

要求分别使用 PCA 和 t-SNE 两种方法进行降维，将 100 维词向量降到 2 维并在 2 维坐标系中画出所有单词对应的点，简单对比两种方法的降维效果，根据词义进行分析。（建议使用 Python，可以直接使用 sklearn 库进行降维，使用 matplotlib 库进行可视化。也可以使用 R 及相应的包。）

提交说明：提交源代码和报告。报告中要贴出生成的图并做简要分析。