Project 1

一. 数据预处理和可视化

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。给出所有出现次数超过 100 的词。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直方图。(hint: library(ggplot2))
6) 给出每一个类别下的新闻数量的分布情况并画出直方图。
7) 给出每个月的新闻数量的分布情况并画出直方图。
8) 加分项：其他你认为有趣的分析；其他有趣的可视化；分享一些好用的包。

二. 新闻相似度计算

1) 利用之前生成的 BagOfWords 向量，计算新闻之间的余弦相似度矩阵。
2) 利用刚刚生成的相似度矩阵，对每个类别，计算该类别内新闻之间的平均相似度。
3) 选取两个类别，计算这两个类别的新闻之间的平均相似度。

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。报告中要求说明每一步的结果，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，确保助教可以根据 README 成功运行你的代码。