Project 1

李林翼 * 朱祺 †

April 14, 2017

Contents

*计 43, 2014011361, limyik.li96@gmail.com
†计 43, 2014011336, zhu-q14@mails.tsinghua.edu.cn

1 数据预处理和可视化

1.1 新闻数据读入与建立数据框对象

第一步是数据的读取。此处为了后面便于处理，将读取的数据持久化，保存为 result/data.csv 文件。可以分为以下三步：确定要读取的新闻的属性；读取新闻到 data.frame；保存为.csv 文件。

1.1.1 确定要读取的新闻的属性

根据 proj1 的要求和 new_york_times_annotated_corpus.pdf 文件，选定以下属性读取：

docid 新闻唯一标识符，也是文档的名字

title 新闻的标题

categories 新闻的类别，使用 online_sections 属性。例子："Business; Technology"。

locations 新闻中提到的地点，使用 Locations 和 Online_Locations 属性。例子："NEW YORK, NY"。

day_of_month,month,year 发行日期，使用 publication_* 属性。例子：26; 06; 1995。

publication_date 发行日期，使用 Publication Date 属性。例子：19950627T000000。

body 新闻正文。

1.1.2 读取新闻到 data.frame

这一部分主要的函数 readDoc() 在 readDoc.R 中。使用了 XML 和 stringr 两个库辅助处理。属性不存在时标记为 NA。

1.1.3 数据持久化

主要的函数 readAll() 和 extractAll() 在 readDoc.R 中。读取目录下所有新闻，将 data.frame 写入 data.csv

1.2 对新闻全文进行预处理

tm 库中有很方便的函数可以进行预处理，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。所有的这些处理都可以使用 tm_map() 函数，通过 map 的方式将转化函数应用到每一个文档语料上。主要函数 getCorpus() 在 process.R 中，返回 Corpus。

1.3 将新闻表示成 BagOfWords 向量

利用上一步得到的 Corpus，借助 DocumentTermMatrix 函数，可以得到文档-词条矩阵，每一行即是 BagOfWords 向量。

1.4 筛选出出现次数大于 100 的词并画 wordcloud

DocumentTermMatrix 得到的文档-词条矩阵通过 findFreqTerms 函数找出出现次数大于 100 的词。利用 wordcloud 函数绘制云图。实现在 process.R 的 drawWordCloud() 中。结果见1。出现最多的词是 said，挺符合新闻报道的特点的。其他高频词如 state、compani、school、work、peopl、american 还是很合理的。但也有些没什么实际含义的词如 also、dont、next、get、just。

1.5 画出单词长度的分布直方图

与上类似，findFreqTerms(DocumentTermMatrix(corpus), 0) 得到所有 word，按单词长度统计。利用 qplot 画图。实现在 process.R 的 drawWordLength() 中。结果见2。可以看到单词的长度基本上在 10 以内，主要集中在 3-6 个字母之间。

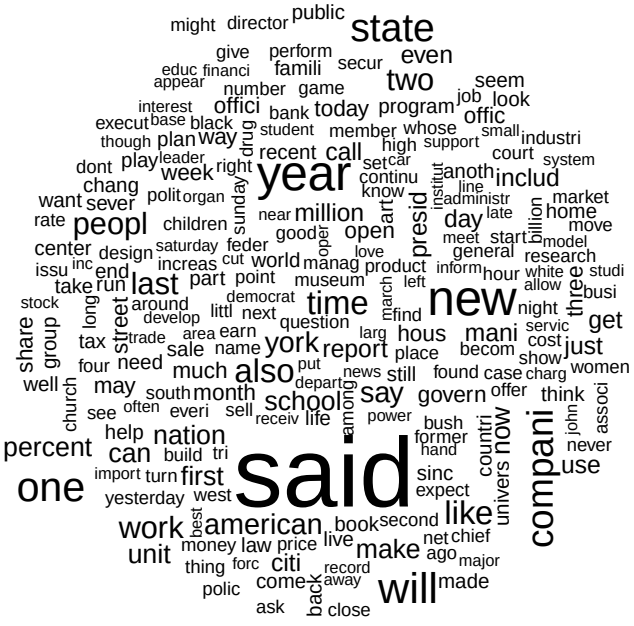

Figure 1: wordCloud

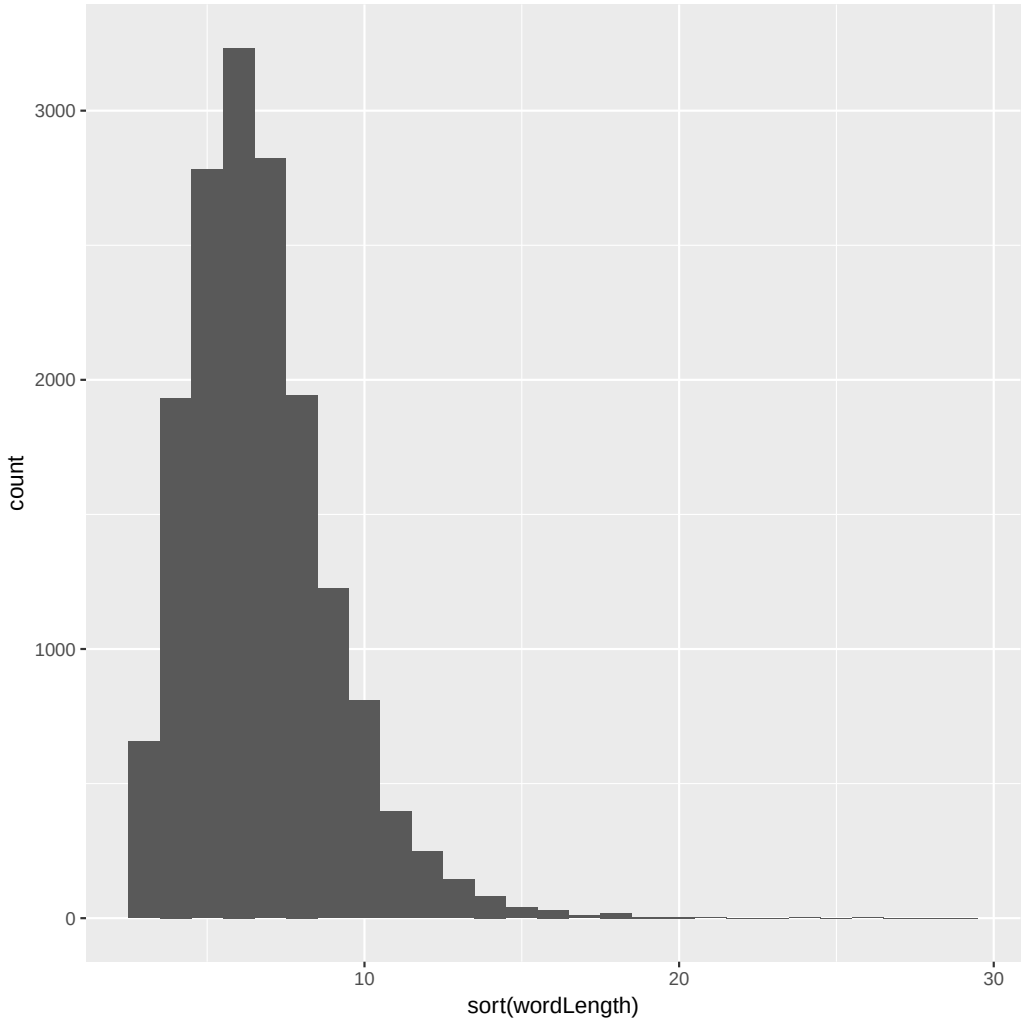

Figure 2: wordLength