

# 数据挖掘 Project1 报告

---

高童 2014011357

陈雅正 2014011423

[github.com/tonygaosh/DM-Project1](https://github.com/tonygaosh/DM-Project1)

## 一、数据预处理和可视化

---

### 1. 从XML中读取数据到data.frame

在解析每个XML新闻文件的时候，使用了R语言的 `XML` 包。在运行 `xmlParse()` 函数后，依次调用 `getNodeSet()` 和 `xmlValue()` 或者 `xmlGetAttr()` 函数获取全文、年份、月份、日期、类别五个属性。

在获取全文过程中，如果判断第一段的前四个字母为"LEAD"，则确认此为重复的LEAD段，将这—个 `<p>` 标签内包含的内容删去。

最后生成的data.frame包含文件名、全文、日期、分类四个属性。

### 2. 对全文进行预处理

采用 `tm` 包，分别用6条语句对文本进行将大写字母都转化为小写，去除停用词、标点符号、数字、空白字符，以及词干化处理。

### 3. 将文本转为 Bag of Words 向量

直接采用 `qdap` 包的 `word_list()` 函数，直接得到每篇全文的单词列表和频数，即为稀疏形式表示的所需BoW向量。

### 4. 作单词云图

对整个语料库采用 `tm` 包的 `DocumentTermMatrix()` 函数，得到了整个语料库的单词-频数列表。对之排序后取前100个，随后用 `wordcloud` 包作图即可。

作出的图位于 `result/wordcloud.pdf`。

### 5. 作单词长度直方图

对上一问得到的 `DocumentTermMatrix`，取其单词部分，用 `nchar` 函数后，再用 `ggplot2` 包作直方图即可。

作出的图位于 `result/word_length.pdf`。

### 6. 作新闻类别直方图

直接将data.frame中的“类别”属性合成一个list后，用 `ggplot2` 作直方图即可。

作出的图位于 `result/category.pdf`。

## 7. 作月份分布直方图

将data.frame中的“日期”属性的月份取出后合成一个list，用 `ggplot2` 作直方图即可。

作出的图位于 `result/monthly.pdf`。

## 二、新闻相似度计算

---

### 1. 计算余弦相似度矩阵

由于相似度矩阵为对称阵，因此只计算一半即可。对传入的两个稀疏形式的BoW向量，取他们的单词的交集，分别相乘再求和，即得到两个向量内积。随后分别除以两个向量的模，即得到余弦相似度。

图中有几个元素的行、列为空白，是因为这些文章中没有“全文”属性。

作出的图位于 `result/similarity.pdf`。

### 2. 计算每个类内平均相似度

首先得到所有类的列表，然后分别找到每个类的所有元素的index。在相似度矩阵中找到那个类部分对应的子方阵，去掉对角线后求均值即可。

作出的图位于 `result/incat_similarity.pdf`。

### 3. 计算类间平均相似度

选择类内相似度最高的两类Education(0.229)和Theater(0.208)。

取得这两类所有元素的index，去除交集后，分别按行、列在相似度矩阵中求出子矩阵，随后直接取平均即可。

计算得出两类的类间平均相似度为0.074，明显低于类内相似度。

## 三、实验总结、好用的包推荐

---

本次实验分工为：陈雅正负责1.3，1.4，2.1部分内容，高童负责框架搭建和其余部分内容。陈雅正将在Project2中承担框架搭建等主要工作。

本次实验使用了R语言包 `qdap`，其CRAN链接[在此](#)。`qdap` 包是一个自然语言处理的包，有很多很有用的函数。如在求BoW向量的过程中，直接用 `word_list()` 函数得到文中的单词频数统计，一步到位。