任务定义

Text classification:

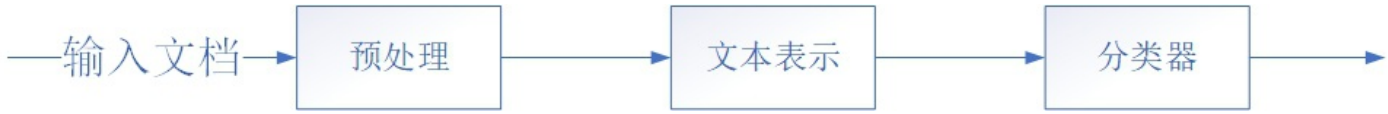
- This data set contains 1000 text articles posted to each of 20 online newgroups, for a total of 20,000 articles. For documentation and download, see http://www-2.cs.cmu.edu/afs/cs/project/theo-11/www/naive-bayes.html.
- The "label" of each article is which of the 20 newsgroups it belongs to. The newsgroups (labels) are hierarchically organized (e.g., "sports", "hockey").
- You should provide model evaluation results and discuss the reasons of the results.

输入输出

输入：用于训练的文本文件、用于测试的文本文件
输出：用于测试的文本文件的分类

方法描述

一个文本分类系统可以简略地用下图表示：



文本表示：向量空间模型（VSM）

- 每个词作为向量空间模型的特征项。一个文档的内容看成是它含有的特征项所组成的集和，表示为：$Document = D(t_1,t_2,\cdots,t_n)$，其中$t_n$是特征项，$1 \le k \le n$。
- 采用了绝对词频（TF）的权重计算方法。即统计特征项在文档中出现的频数。然后所有特征项的权重分别除以文档的全部特征项的总数做归一化。
- 每个特征项赋予一个权重，表示为：$D = D(t_1,w_1;t_2,w_2;\cdots;t_n,w_n)$，其中$w_k$就是$t_k n$的特征项,$1 \le k \le n$。

分类器：朴素贝叶斯分类器

朴素贝叶斯分类器的思想是例用特征项和类别的联合概率来估计给定文档的类别概率。假设文本是基于词的一元模型，即文本中当前词的出现依赖于文本类别，但不依赖于其他词及文本的长度，也就是说词与词之间是独立的。

因为本方案采用了TF向量表示法，即文档向量V的分量为响应特征在该文档中出现的频度，则文档Doc属于c_i类文档的概率为：

$$P(C_i \mid Doc) = \frac{P(C_i)\prod_{t_j \in V} P(t_j \mid C_i)^{TF(t_j, Doc)}}{\sum_j \left[P(C_j)\prod_{t_i \in V} P(t_i \mid C_j)^{TF(t_i, Doc)} \right]}$$

其中，$TF(t_i,Doc)$是文档Doc中特征t_i出现的频度，$P(t_i|C_j)$是对C_i类文档中特征t_i出现的条件概率的拉普拉斯概率估计：

$$P(t_i \mid C_i) = \frac{1 + TF(t_i, C_i)}{|V| + \sum_j TF(t_j, C_i)}$$

这里，$TF(t_i,C_i)$是C_i类文档中特征t_i出现的频度，$|V|$为特征集的大小，即文档表示中所包含的不同特征的总数目。

方案优化

- 方案中 $\prod_{tj \in V} P(t_j|C_i)^{TF(t_j,Doc)}$ 的计算包含了对概率的连乘与指数计算，造成了数值无限逼近于0。使用$\log_2$函数对其进行单调性的等价。
- 向量的相似性过小时，使用$\log_2$函数求和也会产生Nan数值，在结果中判断该数值的产生并跳过。

结果分析

最终成功率达到了85.55%
Success Rate：0.8555
alt.atheism:0.95
comp.graphics:0.835
comp.os.ms-windows.misc:0.82
comp.sys.ibm.pc.hardware:0.735
comp.sys.mac.hardware:0.79
comp.windows.x:0.835
misc.forsale:0.965
rec.autos:0.855
rec.motorcycles:0.895

rec.sport.baseball:0.905
rec.sport.hockey:0.935
sci.crypt:0.88
sci.electronics:0.815
sci.med:0.88
sci.space:0.835
soc.religion.christian:0.97
talk.politics.guns:0.88
talk.politics.mideast:0.865
talk.politics.misc:0.675
talk.religion.misc:0.79

- 其中talk.religion.misc表现较差，查看talk.religion.misc类文档进行分析，发现文档中对语料的标注有多个如Newsgroups: talk.abortion,alt.atheism,talk.religion.misc。再查看分类结果中alt.atheism也占了很大的比例，我想这是由于talk.religion.misc与alt.atheism类文本特征重叠导致的。

源码运行环境

OS: Windows 10 Version 1709
Python SDK Version: 3.6.3
IDE: PyCharm 2017.3.2