体裁识别 On-page 特征分析

一、算法介绍

首先对 article 进行过滤，去除文章内容中单词个数小于 100 的文章。对余下的文章进行特征提取。通过一种词性标注柱状图的特征提取方法获得特征向量，步骤如下。

a) 1. 为长度为 l 的词序列打上词性标注。
b) 2. 用 w 长度的窗口在 l 上滑动，获得 1, ... l-w+1 个窗。
c) 3. 统计每个窗中的词性然后获得二维向量[mean, deviation]。
d) 4. 正则化这些向量并作为最终的特征向量。

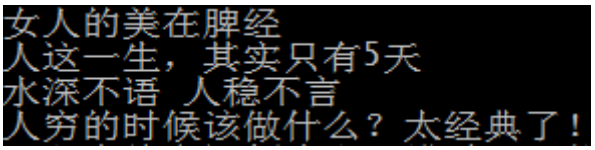
进行特征提取根据文章使用两种方式，一种是直接对文章中词出现的顺序作为词序列，一种是对文章使用 tf-idf 方法提取排名前 l 个关键词，然后将这些关键词排序结果作为词序列。进行实验时，l=100，w=5。

然后使用 SVM 作为分类器，使用的核函数为线性核函数。模型的评估使用准确率，并且使用交叉检验。

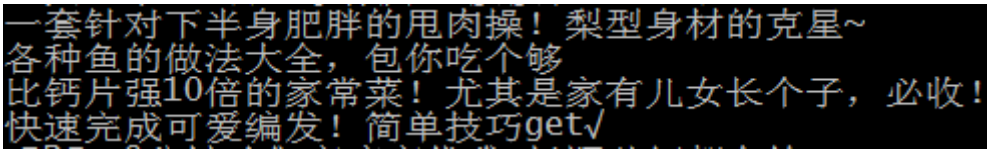
最后使用后向分布的特征选择算法进行特征选择，选择出使得模型效果最好的特征来训练分类器。

二、数据集

选择 200 篇文章，人工标签。类别分为两种，分别是知识性文章（可以理解成优质文章）和道理性文章（可以理解成鸡汤文）。如下图。



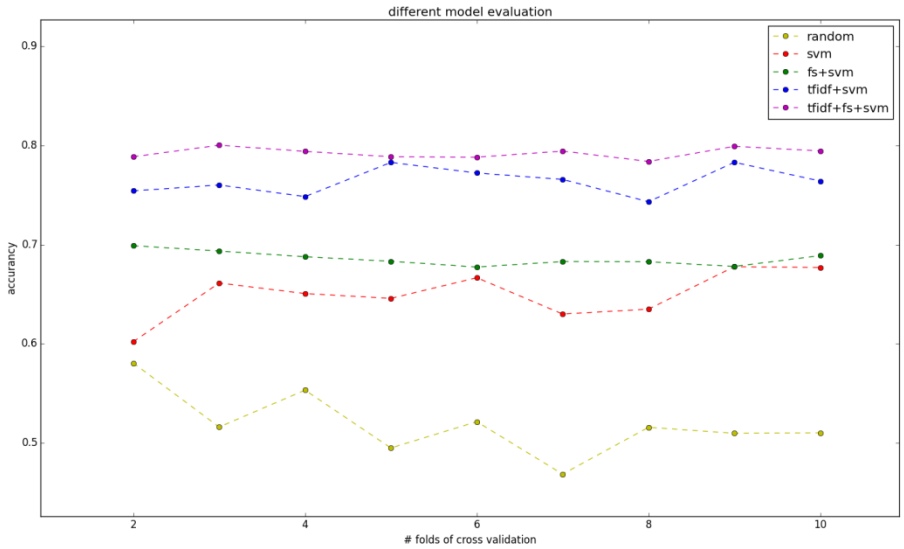
图一、道理性文章标题



图二、知识性文章标题

知识性文章共 120 篇，道理性文章共 80 篇。这个分类体系还不够完善，有待商榷。

三、实验结果



图三、实验结果

图中，横坐标表示交叉检验的划分个数，5 代表将 200 篇文章划分成 5 堆，每次取 4 堆作为训练集，另外 1 堆作为测试集。纵坐标表示准确率，就是分类正确样本占测试样本的比例。各个颜色代表的实验如下。

黄色：随机选择。

红色：直接用文章前 100 个词作为词序列使用 histogram 算法进行特征提取，不使用特征选择，分类器使用 SVM。

绿色：直接用文章前 100 个词作为词序列使用 histogram 算法进行特征提取，使用后向分步算法进行特征选择，分类器使用 SVM。

蓝色：用 tf-idf 排序前 100 个关键词作为词序列使用 histogram 算法进行特征提取，不使用特征选择，分类器使用 SVM。

紫色：用 tf-idf 排序前 100 个关键词作为词序列使用 histogram 算法进行特征提取，使用后向分步算法进行特征选择，分类器使用 SVM。

最终结果可以发现。

1. 线性 SVM 分类器有效。
2. 特征选择会使分类结果更优。
3. tf-idf 改进后分类结果更优。