深度学习与自然语言处理第三次作业

李志渊
PT2200081

一．问题描述

从给定的语料库中均匀抽取 200 个段落（每个段落大于 500 个词）， 每个段落的标签就是对应段落所属的小说。利用 LDA 模型对于文本建模，并把每个段落表示为主题分布后进行分类。验证与分析分类结果，（1）在不同数量的主题个数下分类性能的变化；（2）以"词"和以"字"为基本单元下分类结果有什么差异?

二．具体实现方法

首先，读取语料库中的小说，对其进行预处理。以 utf8 编码格式读取文件内容后，删去小说内的非中文字符以及和小说内容无关的片段，得到字符串形式的语料库，然后使用 jieba 分词进行分词，并过滤停用词，最终返回小说的分词列表。

其次，对于 LDA 模型，首先需要定义模型需要使用的一些变量及参数，其核心部分包括模型初始化和 Gibbs Sampling 两部分，在模型初始化是，随机为文档中的每个词分配一个主题，之后统计每个主题下词出现的概率，及每个文档下出现主题的数量，以及每个主题下词的总量。为了逼 main 出现某一词极少等现象导致的除零异常，设置了 alpha 和 beta 参数。之后则迭代运行 Gibbs Sampling。

再经过迭代后，可计算得到各个文档对应主题的分布，同时每个文档的标签为对应的小说，因此可以使用其主题分布作为特征训练分类器，判断随机的文档属于哪个小说。我使用的是 SVM 的 OneVsRestClassifier 作为多分类模型，首先随机将 200 个段落分为训练集和测试集，比例为 4:1，然后使用训练集进行训练，训练集上评估效果，并将模型保存，之后在测试集上进行测试。

三．运行结果

对每个主题下的词根据概率进行降序排列，结果如下：

```
主题0的高频词为：
中:0.000833      还:0.000579     黄药师:0.000544 见:0.000524     听:0.000481     郭:0.000388     一个:0.000372   只见:0.000331   道:0.000327     今日:0.
000305
--------------------------
主题1的高频词为：
张无忌:0.001434 便:0.000865     少林:0.000538   说道:0.000463   上:0.000457     中:0.000408     恒山:0.000359   魔教:0.000311   掌门:0.000307   已:0.00
0305
--------------------------
主题2的高频词为：
去:0.001003      上:0.000699     只见:0.000650   见:0.000607     倒:0.000376     身上:0.000359   姑娘:0.000359   时:0.000355     已:0.000349     忽:0.00
0305
--------------------------
主题3的高频词为：
道:0.001214      教主:0.001048   都:0.000762     说道:0.000688   兄弟:0.000621   不知:0.000396   人:0.000396     江湖:0.000347   两位:0.000323   夫人:0.
000313
--------------------------
主题4的高频词为：
说:0.001555      便:0.000995     事:0.000500     去:0.000493     问:0.000378     却:0.000372     杀:0.000368     不知:0.000368   大:0.000364     只:0.00
0278
--------------------------
主题5的高频词为：
剑:0.000845      长剑:0.000770   便:0.000701     剑法:0.000565   使:0.000382     一剑:0.000357   师兄:0.000301   兵刃:0.000292   范蠡:0.000290   说道:0.
000270
--------------------------
主题6的高频词为：
道:0.002244      派:0.001610     说:0.000833     人:0.000686     都:0.000607     下:0.000579     做:0.000487     说道:0.000477   嵩山:0.000370   太:0.00
0366
--------------------------
主题7的高频词为：
韦小宝:0.003170 道:0.002997     . :0.000757     去:0.000611     不:0.000565     皇上:0.000528   做:0.000449     皇帝:0.000447   说:0.000447     说道:0.
000445
--------------------------
主题8的高频词为：
弟子:0.001259    师父:0.001235   武功:0.000879   便:0.000749     却:0.000623     功夫:0.000512   见:0.000414     李莫愁:0.000353 想:0.000315     一招:0.
000292
--------------------------
主题9的高频词为：
杨过:0.001336    麽:0.001216     甚:0.000814     不:0.000759     见:0.000625     小龙女:0.000593 却:0.000528     著:0.000506     二人:0.000457   一声:0.
000431
```

```
主题10的高频词为：
道:0.003007      不:0.001667     说:0.001545     好:0.000910     笑:0.000644     说道:0.000544   周伯通:0.000540 只:0.000532     想:0.000516     瞧:0.00
0514
--------------------------
主题11的高频词为：
道:0.001700      都:0.001103     一个:0.001086   说:0.000963     人:0.000847     去:0.000837     还:0.000648     不:0.000625     少女:0.000526   众人:0.
000491
--------------------------
主题12的高频词为：
都:0.000731      下:0.000607     上:0.000465     虚竹:0.000315   有人:0.000297   大:0.000258     敌人:0.000258   镖局:0.000250   众:0.000232     兵刃:0.
000225
--------------------------
主题13的高频词为：
人:0.000952      听:0.000429     请:0.000349     大声:0.000317   说道:0.000295   萧峰:0.000292   汉子:0.000288   女子:0.000254   便:0.000252     剑士:0.
000250
--------------------------
主题14的高频词为：
郭靖:0.001568    黄蓉:0.000977   □:0.000621      於:0.000508     甚:0.000461     後:0.000416     已:0.000408     麽:0.000351     罢:0.000347     丘处机:
0.000329
--------------------------
主题15的高频词为：
去:0.001363      道:0.000995     人:0.000891     一个:0.000861   中:0.000682     不:0.000662     上:0.000605     爹爹:0.000550   说:0.000473     没:0.00
0471
--------------------------
主题16的高频词为：
已:0.001113      上:0.000898     一声:0.000575   身子:0.000565   袁承志:0.000558 快:0.000502     都:0.000457     右手:0.000449   使:0.000443     手中:0.
000437
--------------------------
主题17的高频词为：
道:0.001034      却:0.000843     听:0.000841     去:0.000697     洪七公:0.000630 欧阳锋:0.000510 只:0.000445     心中:0.000445   说道:0.000351   段誉:0.
000345
--------------------------
主题18的高频词为：
见:0.000544      丐帮:0.000502   竟:0.000489     帮主:0.000463   武功:0.000406   蒙古:0.000392   已:0.000380     上:0.000364     再:0.000362     知:0.00
0353
--------------------------
主题19的高频词为：
道:0.003125      令狐冲:0.001781 便:0.001129     不:0.000654     大:0.000597     没:0.000435     师哥:0.000416   林平之:0.000412 师妹:0.000398   一声:0.
000390
```

以下是分类结果：

```
overall accuracy:0.077670
--------------------------
```

```
accuracy for each class: [0.04545455 0.12195122 0.         0.         0.         0.5
 0.         0.         0.         0.         0.         0.
 0.         0.         0.        ]
 ------------------------
 average accuracy:0.044494
```

可以看到模型在训练集和测试集的准确率上都有很多不足，由于时间关系会在后续改进，希望能得到更好的结果。

四．总结

通过本次作业，说明 LDA 模型能够较好地解决一词多义和多词一义的问题，验证了 LDA 的有效性，并通过 SVM 进行了验证，针对金庸的小说，应该能获得不错的效果，本程序会再次进行修改。