**Report of Deep Learning for Natural Langauge Processing**

Jiabo Feng

775193638@qq.com

**Abstract**

对给定的语料库进行分析，具体使用LDA主题模型进行文本建模，分析主题个数、“字”“词”单位，文本长短对模型分类性能的影响。

所给定的语料库为金庸的16篇小说，从中均匀抽取1000个段落作为数据集（每个段落可以有 K 个 token, K 可以取20，100，500, 1000, 3000），每个段落的标签就是对应段落所属的小说。利用LDA模型在给定的语料库上进行文本建模，主题数量为 T，并把每个段落表示为主题分布后进行分类（分类器自由选择），分类结果使用 10 次交叉验证（i.e. 900 做训练，剩余100 做测试循环十次）。实现和讨论如下的方面：

1. 在设定不同的主题个数T的情况下，分类性能是否有变化？；
2. 以"词"和以"字"为基本单元下分类结果有什么差异？

（3）不同的取值的K的短文本和长文本，主题模型性能上是否有差异？

**Introduction**

LDA模型是在2003年所提出，主要用来推测文档的主题分布。它可以将文档集中每篇文档的主题以概率分布的形式给出。通过分析一些文档的主题分布后，便可根据主题分布进行主题聚类或文本分类。

LDA模型认为文章可以由主题分布来表示，主题可以由词汇分布来表示。它不关心文档中单词的顺序，使用词袋特征来代表文档。

**Methodology**

**M1: 文本处理**

对文档进行预处理，去除标点符号、停用词、 隐藏符号和非中文字符。读取小说的字数，按比例均匀抽取1000个段落作为数据集

**M2: LDA模型**

LDA模型的训练过程：

1. 对语料库中的每篇文档的每个词汇随机的赋予一个主题
2. 重新扫描语料库，对每个词，使用Gibbs Sampling公式对其采样，求出他的主题并在语料库中进行更新。
3. 重复步骤2，至Gibbs Sampling收敛
4. 统计语料库中的主题-词出现频率矩阵，该矩阵就是LDA模型。

对于一篇新的文档，可以认为Gibbs Sampling公式中的词语分布部分是稳定不变的，是由训练语料得到的模型提供的，所以采样过程中只要估计该文档的主题分布就好，具体过程：

1.对当前文档中的每个单词，随机初始化一个主题编号;

2. 使用Gibbs Sampling公式，对每个词, 重新采样其主题；

3. 重复以上过程，至Gibbs Sampling收敛；

4. 统计得该文档的主题分布。

**Experimental Studies**

设置LDA模型的主题个数T依次为10、20、50、100，段落长度K为20、100、500、1000、3000.以词为单位时得到分类平均准确率为

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| T K | 20 | 100 | 500 | 1000 | 3000 |
| 10 | 0.083 | 0.153 | 0.200 | 0.311 | 0.632 |
| 20 | 0.134 | 0.133 | 0.210 | 0.377 | 0.757 |
| 50 | 0.110 | 0.157 | 0.305 | 0.464 | 0.737 |
| 100 | 0.091 | 0.159 | 0.317 | 0.444 | 0.706 |

**Conclusions**

1. 在设定不同的主题个数T的情况下，分类性会产生变化，但主题数目没有固定的最优解，可根据训练出的结果，优化主题数目，优化结果；
2. 不同的取值的K的短文本和长文本，主题模型性能上存在差异，长文本中包含更多的语义信息，分类准确性更好。

**References**

[1] [一文详解LDA主题模型 - 知乎 (zhihu.com)](https://zhuanlan.zhihu.com/p/31470216)

[2] [LDA主题模型简介及Python实现-CSDN博客](https://blog.csdn.net/weixin_41168304/article/details/122389948)

[3][用python对单一微博文档进行分词——jieba分词（加保留词和停用词）\_python分词处理和停用词表设置-CSDN博客](https://blog.csdn.net/weixin_41168304/article/details/121707533?spm=1001.2014.3001.5501)