深度学习与自然语言处理第二次作业

汪婧昀
19231136@buaa.edu.cn

实验任务

从给定的语料库中均匀抽取 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）。

实验原理

1. LDA 模型

LDA 是一种基于贝叶斯概率的模型，在自然语言处理中常用于识别文档集中的潜在主题。LDA 的核心假设是一篇文档可以包含多个主题，而每个主题对应一组单词。这些主题以概率分布的形式表示在文档中，使得每个文档都可以通过其主题分布来进行主题聚类或文本分类。

LDA 是一种无监督机器学习技术，可以用来识别大规模文档集（document collection）或语料库（corpus）中潜藏的主题信息。它采用了词袋（bag of words）的方法，这种方法将每一篇文档视为一个词频向量，从而将文本信息转化为了易于建模的数字信息。但是词袋方法没有考虑词与词之间的顺序，这简化了问题的复杂性，同时也为模型的改进提供了契机。每一篇文档代表了一些主题所构成的一个概率分布，而每一个主题又代表了很多单词所构成的一个概率分布。LDA 的核心思想是寻找到最佳的投影方法，将高维的样本投影到特征空间(feature space)，使得不同类别间的数据“距离”最大，而同一类别内的数据“距离”最小。

2. LDA 模型生成

定义文章集合为文档集，文章主题集合为主题，每个文档可以看作一个单词序列$< w_1, w_2, \ldots, w_n >$，其中$w_i$表示第 i 个单词，文档集中所有不同单词构成一个词汇集。每个文档中对应到不同主题的概率为$\theta_d = < p_{t_1}, p_{t_2}, \ldots, p_{t_k} >$，其中$n_{t_i}$表示该文档中对应第 i 个主题的单词数，n 表示该文档中所有词的总数，则

$$p_{t_i} = \frac{n_{t_i}}{n}$$

对于每个主题，生成不同单词的概率为$\varphi_t = < p_{w_1}, p_{w_2}, \ldots, p_{w_k} >$，其中$N_{w_i}$表示对应到该主题的词汇集中的第 i 个单词数目，N 表示该主题的单词总数。

$$p_{w_i} = \frac{N_{w_i}}{N}$$

在 LDA 模型中，文档的生成过程是，首先从文档的主题分布中选取一个主题，然后从这个

主题的单词分布中选取单词，重复这个过程直到文档中的所有单词都被生成。核心公式如下：

$$P(\text{词}\mid\text{文档}) = P(\text{词}\mid\text{主题}) * P(\text{主题}\mid\text{文档})$$

公式以主题作为中间层，通过当前的θ_d和φ_t给出了文档 d 中出现单词 w 的概率，因此利用当前θ_d和φ_t，可以为一个文档中的单词计算它对应任意一个主题的 P(词｜文档)，然后根据这些结果更新这个词对应的主题。相应的，如果更新改变了这个单词对应的主题，反过来也会作用于θ_d和φ_t。

实验内容

1. 不同 topic 数（主题个数 T）对实验结果的影响

表 1 T 对实验结果的影响

K	T	单位	准确率
500	20	字	0.11
500	20	词	0.3
500	100	字	0.13
500	100	词	0.33
500	500	字	0.22
500	500	词	0.31
500	1000	字	0.08
500	1000	词	0.22

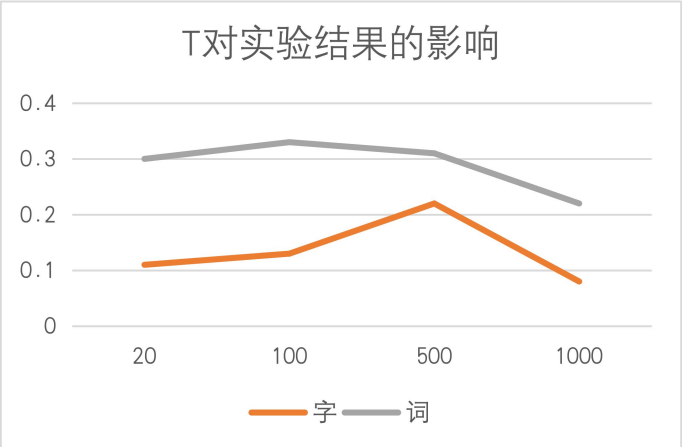

图 1 T 对实验结果的影响

2. 以词/字为单元对实验结果的影响
3. 不同 token 数（文本长度 K）对实验结果的影响

表 2 K 对实验结果的影响

K	T	单位	准确率
20	20	字	0.12
20	20	词	0.18
100	20	字	0.15
100	20	词	0.19
500	20	字	0.12
500	20	词	0.34
1000	20	字	0.11
1000	20	词	0.27
3000	20	字	0.05
3000	20	词	0.19

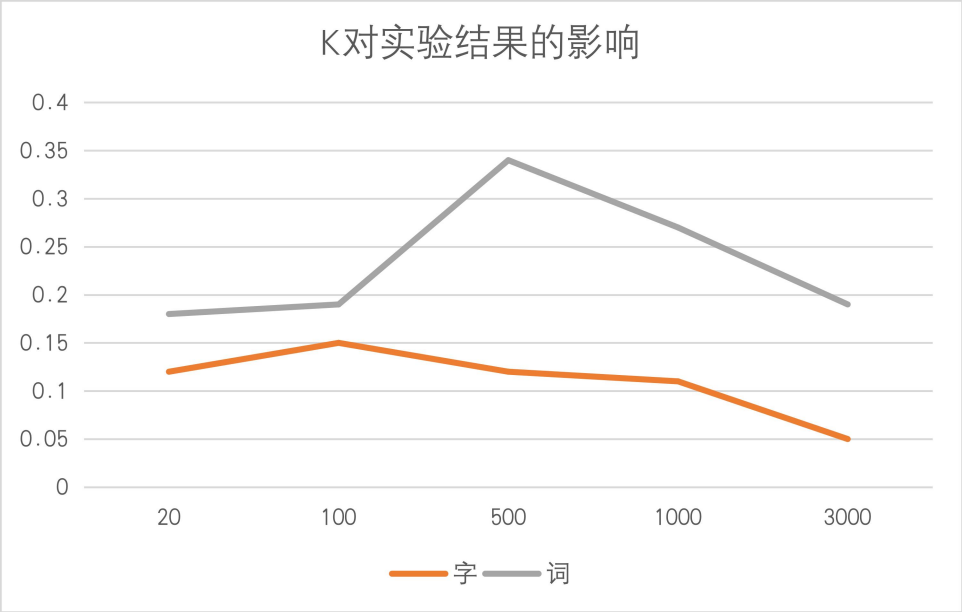

图 2 K 对实验结果的影响

实验结果

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？

通过表 1 观察发现，一定范围内，T 增大分类性能变好，但主题数不能无限增大。

2. 以"词"和以"字"为基本单元下分类结果有什么差异？

通过表 2 观察发现，以词为单位准确率比以字为单位准确率更高。

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？

通过表 2 观察发现，一定范围内，文本越长性能越好。