深度学习与自然语言处理大作业

韩升华
Sy2303405
419130449@qq.com

摘要

LDA 模型，即潜在狄利克雷模型，是一种概率文本模型，用于分析文本集合的主题分布，LDA 通过狄利克雷分布来建模每个主题的词分布，从而实现文本集合中各个主体的推断。

LDA 的基本思想是通过迭代过程来更细主题概率和词在主题下的概率，从而得到一个分布。LDA 假定文本集合是由多个隐含的主题组成，每个文本是从某个给定的总主题中随机抽取。模型的目标是通过从潜在的主题分布中抽取文本集合来描述和解释这些文本集合中的内容。

实现 LDA 时，通过一种概率迭代的方法进行优化，通过逐步更新主题概率和词在主题下的概率来逐步逼近真实的文本分布集合。通过多次迭代，LDA 可以学习到文本集合中的主题分布，并生成新的主题。

引言

通过给定文本来训练 LDA 模型，使得 LDA 学习到文本的主题分布，进而通过 LDA 学到的主题分布，来实现基于 LDA 模型的文本主题分析。

由于中文存在字与词两个文本等级，因而基于字和词训练的 LDA 模型在文本主题分析的结果上存在一定的差异。

方法

首先对给定文本进行预处理，得到了测试数据集和模型训练数据集，随后，将模型训练数据集用于 LDA 模型训练，训练后，用测试数据集测试模型的功能。

实验过程

词模型训练

```
[[0.023      0.006      0.001      ... 0.005      0.002      0.012     ]
 [0.00601202 0.00783385 0.00564766 ... 0.00746948 0.00473675 0.00528329]
 [0.0106     0.005      0.0042     ... 0.0142     0.0056     0.0024    ]
 ...
 [0.00705513 0.00897134 0.00688093 ... 0.00583573 0.01071335 0.00722934]
 [0.00801603 0.00801603 0.00601202 ... 0.00801603 0.00200401 0.00200401]
 [0.016      0.004      0.012      ... 0.02       0.002      0.02      ]]
模型训练完毕!
```

词测试结果

```
number:193 label:倚天屠龙记                result:倚天屠龙记
number:194 label:倚天屠龙记                result:倚天屠龙记
number:195 label:倚天屠龙记                result:倚天屠龙记
number:196 label:倚天屠龙记                result:倚天屠龙记
number:197 label:倚天屠龙记                result:倚天屠龙记
number:198 label:倚天屠龙记                result:倚天屠龙记
number:199 label:鸳鸯刀                  result:鸳鸯刀
number:200 label:越女剑                  result:越女剑
分类正确数：198 分类错误数：2
```

字训练结果

```
[[0.05       0.008      0.011      ... 0.007      0.002      0.         ]
 [0.01886792 0.00784772 0.01235599 ... 0.01018534 0.00166973 0.00333946]
 [0.0212     0.0048     0.019      ... 0.0096     0.0058     0.0058    ]
 ...
 [0.0173913  0.00826087 0.01486957 ... 0.01017391 0.00843478 0.00556522]
 [0.026      0.018      0.016      ... 0.006      0.004      0.002     ]
 [0.028      0.         0.022      ... 0.         0.004      0.006     ]]
```

字测试结果

```
number:194 label:倚天屠龙记                result:鹿鼎记
number:195 label:倚天屠龙记                result:天龙八部
number:196 label:倚天屠龙记                result:倚天屠龙记
number:197 label:倚天屠龙记                result:倚天屠龙记
number:198 label:倚天屠龙记                result:倚天屠龙记
number:199 label:倚天屠龙记                result:倚天屠龙记
number:200 label:倚天屠龙记                result:倚天屠龙记
number:201 label:鸳鸯刀                  result:鸳鸯刀
number:202 label:越女剑                  result:越女剑
分类正确数：180 分类错误数：22
```

结论

通过字和词对 LDA 模型训练基本能够实现对文本主题的分析，基于词的 LDA 模型相较于基于字的 LDA 模型准确性更高、

References

[1] Zenchang Qin and Lao Wang (2023)，How to learn deep learning? Journal of Paper Writing, Vol. 3: 23: pp. 1-12.