深度学习与自然语言处理第三次报告

郭子龙

1216199336@qq.com

摘要

本文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。并验证与分析分类结果，在不同数量的主题个数下分类性能的变化；研究在以"词"和以"字"为基本单元下分类结果的差异。

1. 介绍

LDA（Latent Dirichlet Allocation）模型是一种无监督学习算法，常用于文本分类和主题建模。它可以将文本数据集中的文档表示为一组主题的概率分布，而每个主题又由一组单词的概率分布表示。通过这种方式，LDA 模型可以帮助我们发现文本数据集中隐藏的主题结构，并根据文档中单词的出现情况进行分类。

LDA模型在文本分类中的应用是通过将文本看作由多个主题组成的混合物，并将每个主题看作一个类别或标签。然后，模型根据文本中每个单词所属的主题，计算每个文本属于每个类别的概率。最终，模型将文本分配给概率最高的类别或标签。LDA模型文本分类的可以处理大量的文本数据，并且可以自动识别和学习文本中的主题和类别。它可以用于许多应用程序，例如文本分类、情感分析、信息检索等。

2. 方法

M1: 数据读取与预处理模块

第一部分对文本进行读取，按照所给目录对小说进行读取和存储，之后对小说进行预处理，首先删去原文中的标点符号和广告词。随后依据停词语料库删去文本中相关停词。

M2: Gibbs采样

Gibbs采样是一种特殊的马尔可夫链算法，常被用于解决包括矩阵分解、张量分解等在内的一系列问题，也被称为交替条件采样(alternating conditional sampling)其中，“交替”一词是指Gibbs采样是一种迭代算法，并且相应的变量会在迭代的

过程中交替使用，除此之外，加上“条件”一词是因为Gibbs采样的核心是贝叶斯理论，围绕先验知识和观测数据，以观测值作为条件从而推断出后验分布。

LDA模型中，Gibbs算法流程如下

1).初始化参数

初始化主题数，主题权重等。

2).初始化主题分布和单词分布

3).遍历文档中的单词

对文档中每个单词进行遍历，依据条件概率分布，计算该单词在不同主题下的概率，然后根据概率进行采样。

对于每个单词 $w_{d,n}$，计算其属于每个主题k的概率 $p(z_{d,n}=k \mid w_{d,n},\phi,\theta)$，根据此结果更新单词的主题和分布。

4).重复步骤(3)直至满足迭代次数或者满足收敛要求。

5).输出结果

M3: K-Means 聚类算法

K-Means 是一种聚类算法，它将数据集划分为 k 个簇，其中每个簇包含具有相似特征的数据点。该算法的核心思想是将所有数据点分配给距离它们最近的簇中心，并将簇中心移动到簇中所有数据点的平均值的位置。重复这个过程，直到簇中心不再发生变化或者达到预设的迭代次数。

1).选择初始化的K个样本作为初始聚类中心 $a=a_1,a_2,...,a_k$

2).针对数据集中的每个样本 x_i 计算它到 k 个聚类中心的距离并将其分到距离最小的类中

3).针对每个类别 a_j，重新计算它的聚类中心 $a_j=\frac{1}{|c_i|}\sum_{x\in c_i}x$ (即属于该类的所有样本的质心)

4).重复2-3的操作，直至达成终止条件

3. 实验结果

3.1基于词分割的采样结果

当选择topic个数为5的时候，各主题最高词频的词语如下所示，不具备明显区分性：

```
topic: 1
[('道', 415), ('韦小宝', 186), ('说', 153), ('便', 125), ('听', 89)]
topic: 2
[('道', 305), ('说', 198), ('便', 136), ('中', 134), ('李文秀', 117)]
topic: 3
[('道', 314), ('一声', 129), ('剑', 126), ('便', 123), ('中', 119)]
topic: 4
[('道', 567), ('说', 221), ('便', 196), ('麽', 138), ('听', 136)]
topic: 5
[('道', 377), ('说', 130), ('中', 102), ('说道', 80), ('杨过', 77)]
```

当循环五百次后，部分段落隶属度如下所示：

```
topic1 [2] confidence [0.7755511] topic2 [1] confidence [0.14829659]
```

```
topic1 [0] confidence [0.63326653] topic2 [2 3] confidence [0.1743487 0.1743487]
```

```
topic1 [2] confidence [0.86372745] topic2 [1] confidence [0.04809619]
```

若设置主题数为15，结果如下，相较下有一定特征区别：

```
topic: 1
[('李文秀', 89), ('中', 87), ('教主', 45)]
topic: 2
[('道', 77), ('老人', 58), ('苏鲁克', 47)]
topic: 3
[('想', 58), ('当下', 40), ('道', 36)]
topic: 4
[('皇帝', 56), ('更', 51), ('道', 49)]
topic: 5
[('麽', 202), ('道', 159), ('听', 106)]
topic: 6
[('道', 72), ('没', 55), ('剑法', 52)]
topic: 7
[('中', 105), ('范蠡', 60), ('听', 55)]
topic: 8
[('道', 76), ('范蠡', 49), ('阿青', 41)]
topic: 9
[('一声', 148), ('石破天', 65), ('胡斐', 56)]
topic: 10
[('剑士', 114), ('之中', 58), ('道', 55)]
topic: 11
[('道', 210), ('说', 96), ('做', 69)]
topic: 12
[('道', 176), ('著', 113), ('说', 109)]
topic: 13
[('便', 44), ('吃', 41), ('见', 38)]
topic: 14
[('师父', 107), ('道', 90), ('事', 41)]
topic: 15
[('二人', 83), ('派', 73), ('道', 42)]
topic: 16
[('张无忌', 79), ('说道', 43), ('便是', 42)]
topic: 17
[('心想', 75), ('突然', 44), ('见', 36)]
topic: 18
[('一个', 83), ('说', 61), ('便', 56)]
topic: 19
[('韦小宝', 173), ('道', 84), ('说道', 71)]
topic: 20
[('少女', 79), ('道', 72), ('两位', 40)]
topic: 21
[('剑', 64), ('令狐冲', 63), ('便', 56)]
topic: 22
[('道', 38), ('两人', 36), ('起来', 29)]
topic: 23
[('道', 157), ('走', 64), ('说道', 59)]
topic: 24
[('只见', 61), ('见', 42), ('黄蓉', 38)]
topic: 25
[('知道', 75), ('手中', 55), ('郭靖', 53)]
topic: 26
[('道', 185), ('便', 68), ('听', 40)]
topic: 27
[('袁承志', 133), ('道', 68), ('说道', 67)]
topic: 28
[('说', 74), ('道', 71), ('笑', 46)]
topic: 29
[('见', 96), ('陈家洛', 68), ('说', 60)]
topic: 30
[('说', 81), ('爹爹', 75), ('右手', 65)]
```

3. 2基于字分割的采样结果

```
topic: 1
[('女', 162), ('头', 139), ('金', 124), ('花', 101), ('听', 96)]
topic: 2
[('年', 92), ('官', 91), ('天', 83), ('国', 78), ('成', 78)]
topic: 3
[('著', 175), ('死', 172), ('爹', 164), ('麽', 147), ('声', 124)]
topic: 4
[('道', 343), ('家', 165), ('李', 160), ('文', 137), ('会', 134)]
topic: 5
[('出', 263), ('头', 157), ('见', 147), ('力', 136), ('黄', 134)]
topic: 6
[('刀', 314), ('子', 243), ('娘', 124), ('袁', 121), ('相', 120)]
topic: 7
[('手', 461), ('身', 448), ('中', 313), ('时', 189), ('动', 181)]
topic: 8
[('道', 417), ('说', 399), ('胡', 237), ('师', 166), ('见', 155)]
topic: 9
[('中', 272), ('两', 156), ('镖', 101), ('山', 99), ('头', 92)]
topic: 10
[('马', 308), ('兵', 197), ('铁', 128), ('木', 92), ('奔', 82)]
```

```
topic: 11
[('士', 255), ('青', 228), ('名', 211), ('长', 169), ('中', 126)]
topic: 12
[('道', 509), ('老', 374), ('石', 211), ('掌', 191), ('便', 176)]
topic: 13
[('中', 236), ('事', 146), ('杨', 144), ('说', 121), ('龙', 120)]
topic: 14
[('子', 281), ('道', 188), ('然', 127), ('越', 119), ('行', 100)]
topic: 15
[('道', 261), ('宝', 237), ('主', 201), ('说', 179), ('太', 148)]
topic: 16
[('声', 199), ('令', 127), ('气', 122), ('子', 116), ('毒', 109)]
topic: 17
[('功', 282), ('武', 226), ('师', 144), ('位', 143), ('高', 140)]
topic: 18
[('回', 200), ('十', 189), ('四', 167), ('面', 150), ('众', 144)]
topic: 19
[('剑', 782), ('招', 202), ('法', 176), ('万', 170), ('手', 159)]
topic: 20
[('天', 220), ('道', 212), ('中', 164), ('雄', 144), ('说', 143)]
```

```
topic1 [7] confidence [0.17635271] topic2 [12] confidence [0.15631263]
```

```
topic1 [9] confidence [0.36873747] topic2 [17] confidence [0.08016032]
```

基于字分割的主题提取效果要弱于以词为分割。

部分聚类结果：

10 5 1 3 8 8 5 3 8 1 3 0 3

1 2 10 9 5 2 1 1 14 15 2 10 1

15 1 1 11 5 15 11 1 1 11 15 5 10

0 10 10 12 3 5 1 9 10 5 5 15 10

1 11 10 11 11 1 0 11 11 12 11 15 0

3.3基于词分割的Kmeans聚类算法结果

文本1	11 11 11 11 11 11 11 11 11 11 11 11 11
文本2	10 10 10 10 10 10 10 10 10 10 10 10 10
文本3	2 2 2 2 2 2 2 2 2 2 2 2 13
文本4	1 1 1 1 1 1 1 1 1 1 1 1 1

部分文本的K-MEANS聚类分析结果如上，在迭代次数足够多的条件下，能够对主题实现提取。

4.总结

LDA模型在迭代次数足够多的时候，能够较好的实现主题词的提取，当设置合适的主题词数量的时候，能够增强主题提取的效果，反之，效果会减少。以字为分割的主题提取效果要明显弱于以词为分割的效果，单词能表达的意思要多于字。这是由文字的熵所决定的。