# **Report of Deep Learning for Natural Language Processing**

曹思远

[3203998114@qq.com](mailto:3203998114@qq.com)

## **Abstract**

本次作业是基于中文语料库，利用LDA主题模型和自选分类器来实现小说语段主题分类。选择使用SVM分类模型。首先将提取的数据段落分为训练集合测试集，其次利用LDA和SVM针对训练集得到训练模型，最后通过测试集进行预测分析试验。

# **Introduction**

具体试验要求如下：

从下面链接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），

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？

# **Methodology**

## **M1:LDA模型介绍**

LDA（Latent Dirichlet Allocation）主题模型是一种无监督的机器学习技术，它可以从大量文档中发现隐藏的主题。LDA是一个三层贝叶斯概率模型，由词、主题和文档三层结构组成。在给定文档集合的情况下，LDA可以推断出每篇文档的主题分布，以及每个主题下的词汇分布。

一般的生成过程如下：

* 从整个文档集合中的主题分布（Dirichlet分布）中采样一个主题分布θ。
* 对于文档中的每个词汇，从主题分布θ中采样一个主题z。
* 根据主题z从词汇-主题分布β中采样一个词汇w。

数学表达如下：

𝑝(𝜃∣𝛼)=𝐷𝑖𝑟(𝜃∣𝛼)p(θ∣α)=Dir(θ∣α)

𝑝(𝑧∣𝜃)=𝑀𝑢𝑙𝑡𝑖𝑛𝑜𝑚𝑖𝑎𝑙(𝑧∣𝜃)p(z∣θ)=Multinomial(z∣θ)

𝑝(𝑤∣𝑧,𝛽)=𝑀𝑢𝑙𝑡𝑖𝑛𝑜𝑚𝑖𝑎𝑙(𝑤∣𝑧,𝛽)p(w∣z,β)=Multinomial(w∣z,β)

其中Dirichlet分布Dir(θ|α)表示主题分布θ的先验分布，α是超参数；Multinomial分布表示在给定主题分布θ和词汇-主题分布β的条件下，词汇w的生成分布。

## **M2:SVM模型介绍**

支持向量机（support vector machines, SVM）是一种二分类模型，它的基本模型是定义在特征空间上的间隔最大的线性分类器，间隔最大使它有别于感知机；SVM还包括核技巧，这使它成为实质上的非线性分类器。SVM的的学习策略就是间隔最大化，可形式化为一个求解凸二次规划的问题，也等价于正则化的合页损失函数的最小化问题。SVM的的学习算法就是求解凸二次规划的最优化算法。SVM在许多领域都有广泛的应用，包括文本分类、图像识别、生物信息学、金融预测等。它具有良好的泛化能力和鲁棒性，在处理高维数据和非线性数据时表现出色。

## **M3:实施过程**

（1）数据初始化，包括文件路径设定，段落1000，模式两种‘字’，‘词’，话题数量{2，5，10，20，50，100}，token数量{20，100，500，1000}，交叉验证次数10

（2）将实验数据分为10组，其中9组共900段作为训练集，剩余1组共100段作为测试集，相应小说名字作为标签。后将训练集经过LDA主题模型（主要是训练集形成词典和词袋作为输入）和SVM分类器（将LDA的主题分布，标签作为输入），得到训练模型。

（3）再利用上述的训练模型对测试集进行分类准确度测试，交叉验证10次后，取平均值，观察试验结果。

# **Experimental Studies**

将实验数据整理成表格如下：

分为字和词两组，其中L：表示tokens数量；T：表示LDA中的主题数量。

字测试集的准确度

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| L/T | 2 | 5 | 10 | 20 | 50 | 100 |
| 20 | 0.1550 | 0.1500 | 0.1510 | 0.1600 | 0.1780 | 0.1970 |
| 100 | 0.1720 | 0.1870 | 0.2180 | 0.2260 | 0.2450 | 0.2730 |
| 500 | 0.1910 | 0.3520 | 0.3880 | 0.4630 | 0.4830 | 0.4720 |
| 1000 | 0.2230 | 0.3860 | 0.4940 | 0.6000 | 0.6510 | 0.6310 |

词测试集的准确度

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| L/T | 2 | 5 | 10 | 20 | 50 | 100 |
| 20 | 0.1280 | 0.1390 | 0.1450 | 0.1440 | 0.1650 | 0.2060 |
| 100 | 0.1590 | 0.1840 | 0.1850 | 0.2150 | 0.2310 | 0.2920 |
| 500 | 0.2330 | 0.2430 | 0.2780 | 0.3540 | 0.3390 | 0.3990 |
| 1000 | 0.2170 | 0.3440 | 0.3150 | 0.4130 | 0.3090 | 0.4350 |

观察试验数据得到如下结论：

（1）在设定不同的主题个数T的情况下，‘字’和‘词’的分类性能均随着T的变大而变好。

（2）针对不同的取值的K的短文本和长文本，主题模型性能上是存在差异的，文本越长，‘字’和‘词’的分类性能也在变好。

（3）从整体来看，以‘字’为基本单元下的分类结果要比以‘词’为基本单元的分类结果要好，前者在L=1000下准确度要达到0.6以上，后者只有0.4。

# **Conclusion**

本次作业是基于中文语料库，利用LDA主题模型和SVM分类器来实现小说语段主题分类。首先将提取的数据段落分为训练集合测试集，其次利用LDA和SVM针对训练集得到训练模型，最后通过测试集进行分类预测试验。试验结果 ‘字’和‘词’的分类性能均随着主题数量的变大而变好；以‘字’为基本单元下的分类结果要比以‘词’为基本单元的分类结果要好；随着文本长度增加，‘字’和‘词’的分类性能也在变好。

# **Referances**

[1] https://blog.csdn.net/m0\_47220500/article/details/105765841

[2] https://www.cnblogs.com/liuxiaochong/p/14269313.html