深度学习与自然语言处理第二次作业

ZY2314222 魏智兴

Abstract:

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），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）。实现如下的方面的研究讨论：（1）在设定不同的主题个数T的情况下，分类性能的变化；（2）以"词"和以"字"为基本单元下分类结果差异；（3）不同的取值的K的短文本和长文本，主题模型性能上的差异。本次研究采用LDA主题模型和SVM分类器实现小说语段分类任务。首先，通过LDA主题模型训练文本，获得各部小说的主题分布。接着，利用训练集经过LDA模型提取的主题概率分布作为特征向量，训练SVM分类器以将概率分布向量分类为不同小说。最后，使用训练好的SVM分类器，通过LDA模型提取测试集的主题概率分布，并推断测试样本的小说类别。实验考察了不同主题数量T、基本单元选定（字或词）以及不同取值长度K对分类性能的影响。结果显示，一般情况下，增大T和K可提升分类性能，但T值存在边际效应。此外，以“字”为基本单元的分类性能普遍高于以“词”为基本单元的情况。

Introduction

一、LDA模型

LDA由Blei, David M.、Ng, Andrew Y.、Jordan于2003年提出，是一种主题模型，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，从而通过分析一些文档抽取出它们的主题（分布）出来后，便可以根据主题（分布）进行主题聚类或文本分类。同时，它是一种典型的词袋模型，即一篇文档是由一组词构成，词与词之间没有先后顺序的关系。

在LDA模型中，一篇文档生成的方式如下：

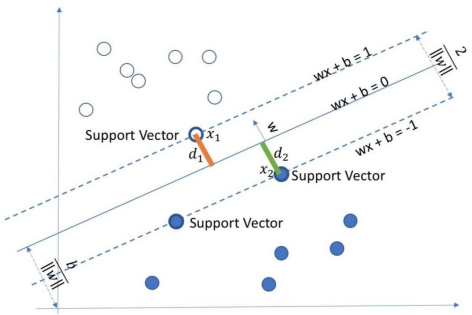
- 从狄利克雷分布α中取样生成文档 i 的主题分布θ_i
- 从主题的多项式分布θ_i中取样生成文档i第 j 个词的主题$z_{i,j}$
- 从狄利克雷分布β中取样生成主题$z_{i,j}$对应的词语分布$\phi_{z_{i,j}}$
- 从词语的多项式分布$\phi_{z_{i,j}}$中采样最终生成词语$w_{i,j}$

二、SVM 算法

支持向量机（support vector machine，SVM）是有监督学习中最有影响力的机器学习算法之一，该算法的诞生可追溯至上世纪 60 年代， 前苏联学者 Vapnik 在解决模式识别问题时提出这种算法模型，此后经过几十年的发展直至 1995 年， SVM 算法才真正的完善起来，其典型应用是解决手写字符识别问题。

SVM 是一种非常优雅的算法，有着非常完善的数学理论基础，其预测效果，在众多机器学习模型中“出类拔萃”。在深度学习没有普及之前，“支持向量机”可以称的上是传统机器学习中的“霸主”。

支持向量机是一种二分类模型，其基本模型定义为特征空间上的间隔最大的线性分类器，其学习策略便是间隔最大化，最终可转化为一个凸二次规划问题的求解。支持向量机的学习算法是求解凸二次规划的最优化算法。



以一个二维平面为例，判定边界是一个超平面（在本图中其实是一条线，但是可以将它想象为一个平面乃至更高维形式在二维平面的映射），它是由支持向量所确定的（支持向量是离判定边界最近的样本点，它们决定了判定边界的位置）。间隔的正中就是判定边界，间隔距离体现了两类数据的差异大小

基础的SVM算法是一个二分类算法，至于多分类任务，可以通过多次使用SVM进行解决。

Methodology

总体思路为：首先从给定的语料库进行文本预处理。删除所有的隐藏符号、非中文字符和标点符号中，接着均匀抽取1000个段落作为数据集（每个段落可以有 K 个 token, K 取20，100，500，1000，然后每个段落进行打标签，每个段落的标签为对应段落所属的小说。然后利用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选用支持向量机SVM模型进行分类，最后将分类结果使用 10 次交叉验证，得到训练集和测试集的分类精度。具体来说，方法步骤如下：

- 文本预处理：先将文档内的所有小说进行合并得到完整文本，然后对语料库进行预处理。删除所有的隐藏符号、非中文字符和标点符号。使用结巴分词和字符级标记化对文本进行预处理。
- 构建数据集：从预处理后的文本中均匀抽取1000个段落作为数据集，然后900 做训练，剩余100做测试的方式构建训练集和测试集，并将其转换为词袋表示。
- 调用LDA模型：使用构建好的数据集，放入LDA模型中得到每个段落的主题分布。
- SVM分类验证：使用支持向量机SVM模型对主题特征进行分类，并评估分类的精度。

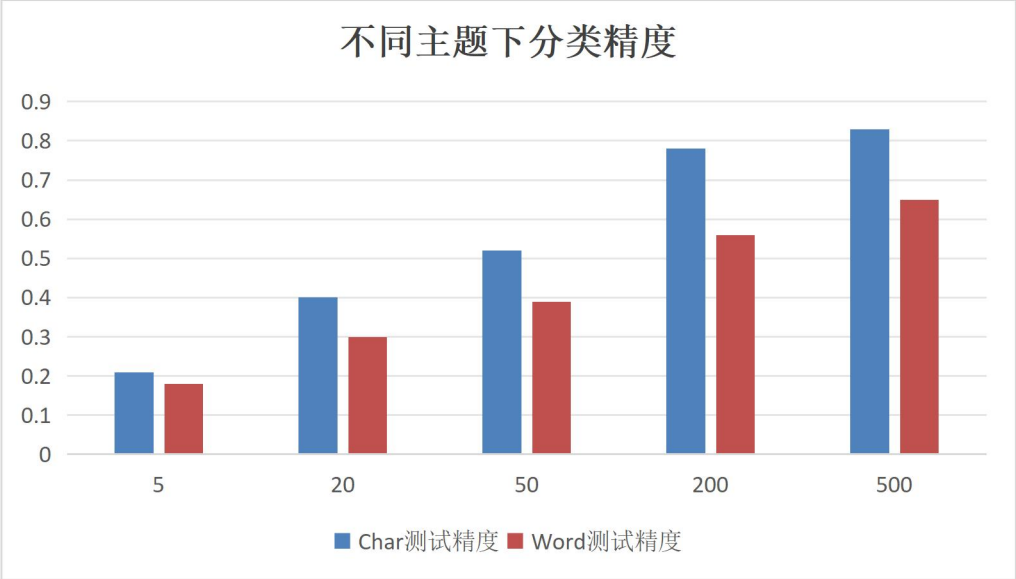
Experimental Studies

（1）问题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?

设置段落token为1000不变，通过分别设置实验主题数量为5，20，50，200，500，在以word和char两种模式下进行实验，实验结果如下：

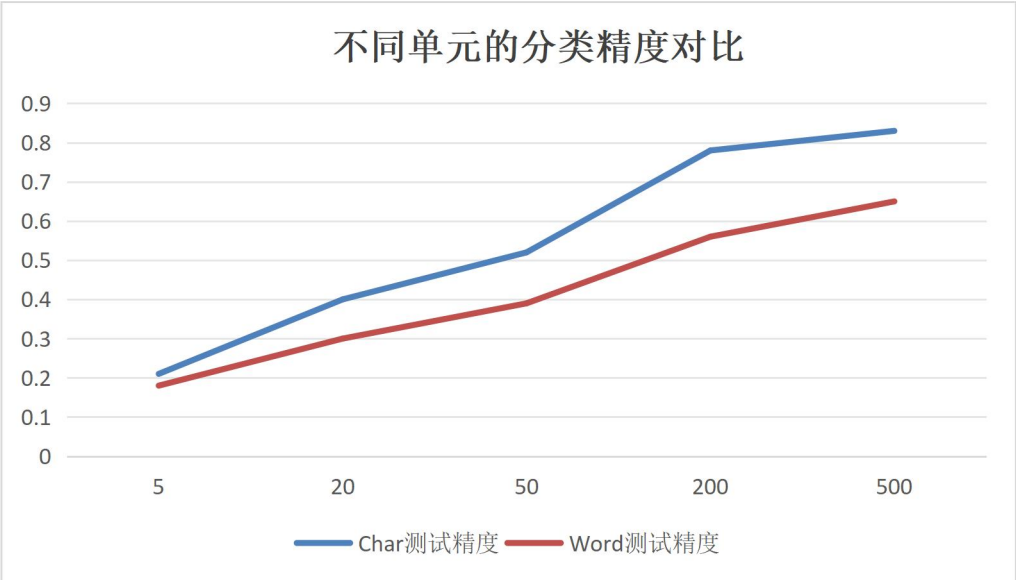
主题T的数量	5	20	50	200	500
Word训练精度	0.14	0.28	0.32	0.45	0.50
Word测试精度	0.18	0.30	0.39	0.58	0.65

Char训练精度	0.17	0.35	0.36	0.65	0.77
Char测试精度	0.21	0.40	0.52	0.78	0.83



可以看出，不论是以 word 或char为基本单元，分类准确率均随主题数量的增而增加。但是当主题个数提升到一定水平后，精度上升缓慢。增加特征数量通常会增加模型的复杂度，从而有助于提高模型的性能。但过大的主题数量也可能引入过拟合的风险，进而降低分类精度。

（2）问题2：以"词"和以"字"为基本单元下分类结果有什么差异?

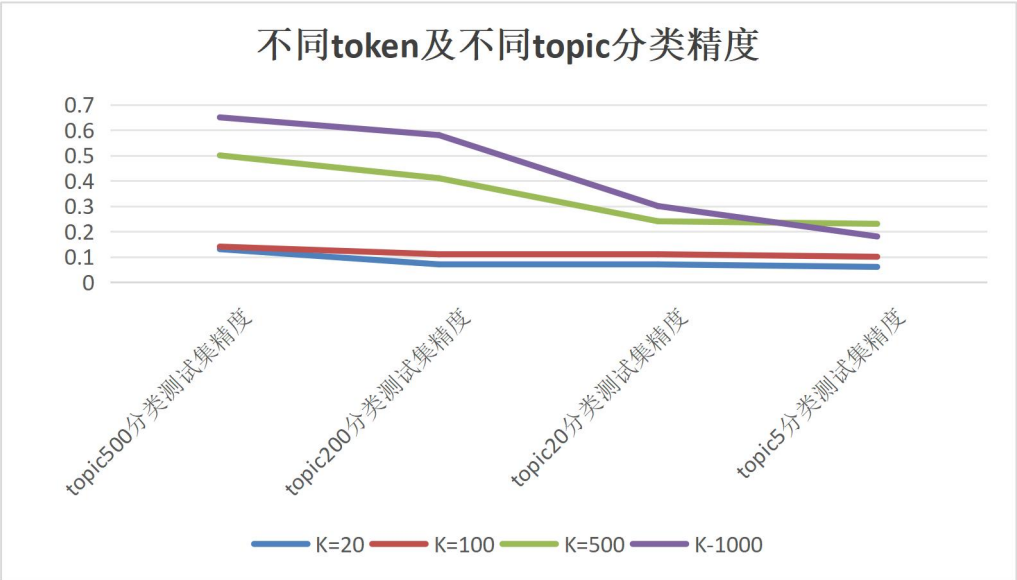


在问题1探究的实验中画图如上图所示，可以看出，通常情况下，以char字为基本单元的分类精度普遍高于以word词的分类精度。

（3）问题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？

以word词为基本单元进行实验，设置不同的token大小及不同的主题T个数，进行实验，得到的结果如下表：

Token:K / Topic:T	20	100	500	1000
5	0.06	0.10	0.23	0.18
20	0.07	0.11	0.24	0.30
200	0.07	0.11	0.41	0.58
500	0.13	0.14	0.50	0.65



可以看出：在保持主题不变的情况下，随着Token数目的增加，以字和词作为基本单元的分类准确率均呈上升趋势。这种现象的分析如下：随着Token数量的增加，提供给训练分类模型的信息也随之增加。以词为基本单元时，每

个词被视为一个独立特征。增加Token数量会增加特征空间的丰富度，更充分地描述文本的内容和特征。这同时包含更多语义和上下文信息，有助于提升模型对文本的理解和分类性能。

Conclusions

通过本次作业，使得我对在LDA模型确定主题分布后改变不同主题和不同基本单元下分类器模型的改变有了更深入的了解。