信息内容安全相关论文阅读笔记

一、基于LDA—加权Word2Vec组合的机器学习情感分类模型研究

主要内容

目的与目标

- 研究评论文本的目的——挖掘潜在市场、了解用户需求等
- 两种研究目标——情感分析（用户态度分析）&情感分类（积极/消极等）

Jieba

- 对数据集进行分词

LDA训练主题模型

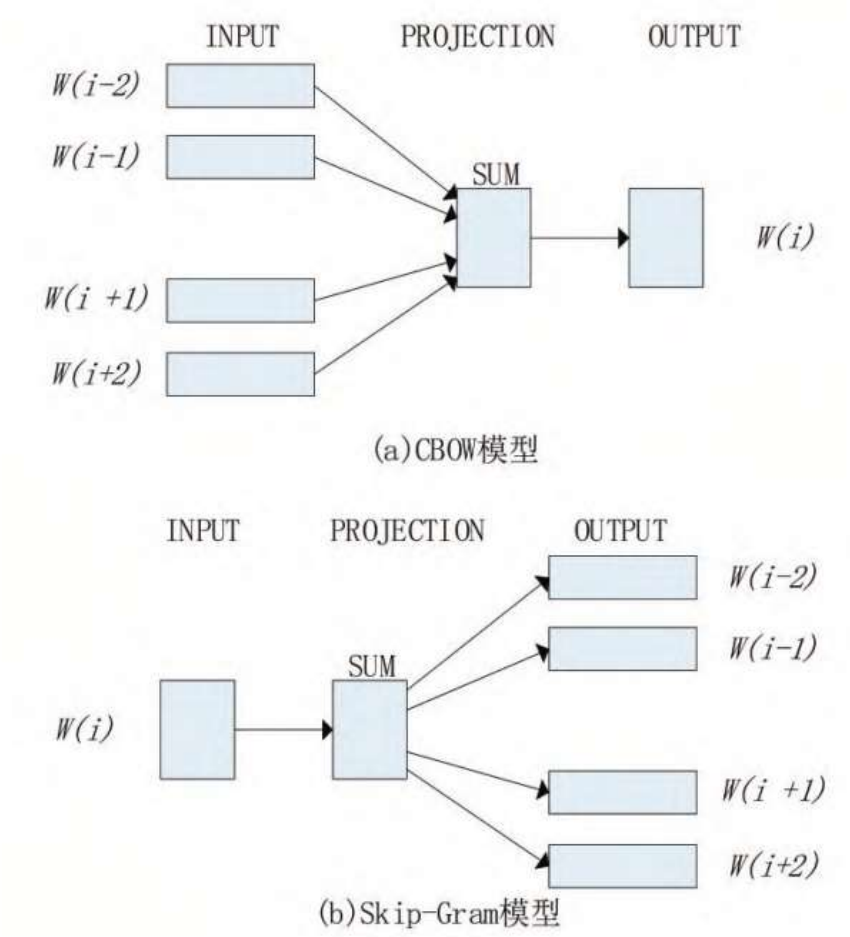
- 借助困惑度确定最佳主题数K
- 计算每个词与主题的相关性，取前300个作为特征词代表当前主题
- 对每一个文本，只提取出其中的特征词

	长评论文本
LDA特征表达前	宾馆在小街道上，不大好找，但还好北京热心同胞很多~宾馆设施跟介绍的差不多，房间很小，确实挺小，但加上低价位因素，还是物超所值的；环境不错，就在小胡同内，安静整洁，暖气好足-_-ll。。。呵还有一大优势就是从宾馆出发，步行不到十分钟就可以到梅兰芳故居等等，京味小胡同，北海距离好近呢。总之，不错。推荐给节约消费的自助游朋友~比较划算，附近特色小吃很多~
LDA特征表达后	“宾馆”“还好”“北京”“宾馆”“设施”“介绍”“房间”“很小”“环境”“不错”“胡同”“安静”“整洁”“暖气”“优势”“宾馆”“出发”“步行”“不到”“胡同”“距离”“不错”“推荐”“消费”“朋友”

Word2Vec

- 实现在大规模语料数据中通过CBOW 或 Skip-gram 训练生成每个词的向量特征。
- CBOW: 通过输入词的上下文向量来预知当前词出现的概率

- Skip-gram则是通过当前词预测出上下文出现的概率



二、Latent Dirichlet Allocation

TF-IDF文本模型

- TF：词频，文本中某个词在该文本中出现的频率
- IDF：逆文本，语料库中文本包含该词的对数频率
- TD-IDF：TF与IDF的乘积
- 假设一个语料库有N篇不同的文档，M个不同的词，通过TD-IDF的计算方式获得(M, N)的矩阵
- 构造完矩阵后，每一篇文章就由矩阵中对应的列向量来表示
- 缺陷：没有将原有的文本信息压缩了很多，而且单纯地统计词频也没有很好地挖掘词语间、文本间的信息

LSI文本模型

- 将TF-IDF中得到的矩阵进行分解
- (M, N)分解为（M, r）、（r, r）、（r, N）
- r表示主体的个数
- 第一个子矩阵：词和主题的关系
- 第二个子矩阵：主题之间的关系
- 第三个子矩阵：主题和文档的关系
- 缺陷：矩阵分解的不可解释性

Unigram文本模型

- 文档中的每个词都是从一个单独的多项分布中独立采样而得的

$$p(\vec{w}) = p(w_1, w_2, ..., w_N) = \prod_{n=1}^{N} p(w_n)$$

- 缺陷：显然没有考虑到上下文直接的关系，也没有考虑不同主题的文章的差别

mixture of unigrams文本模型

- 假设每篇文章只属于一个主题
- 文本的生成过程是先选择一个主题z，然后从条件多项分布 $p(w|z)$ 中独立地生成N个词，

$$p(\vec{w}) = p(w_1, w_2, ..., w_N) = \sum_{z} p(z) * \prod_{n=1}^{N} p(w_n|z)$$

- 缺陷：一篇文档往往属于多个主题

pLSA（pLSI）文本模型

- 引入文本变量来使得对于一个特定的目标文本，可以有多个主题以加权的形式结合在一起
- 作者认为：一篇文档由多个主题混合而成，每个主题都是词汇上的概率分布，文档中每个单词都是先确定一个主题后，然后在该主题下生成。
- 生成文章可以认为是如下过程
  - 在K个主题中随机确定一个主题
  - 在确定主题的条件下随机生成一个词
  - 重复以上操作
- 文章 d_m 中词 w_j 的概率

$$p(w_j|d_m) = \sum_{k=1}^{K} p(w_j|z_k) * p(z_k|d_m)$$

- 缺陷：泛化能力差，主题词只能来自于训练集中的文档，只能提取出在训练集文档中出现过的词语，对于一个里面大部分词都没在训练集文档中出现过的“未知”文档很无力。

LDA文章生成过程

- 文章之间生成是独立的
- 对于某个文章，它的生成可以认为是这样的：
  - 确定文章单词个数 N（参数为 ϵ 的泊松分布）
  - 确定文章的主题分布 θ（参数为 α 的狄利克雷分布）
  - 对于每一个单词，首先从主题分布中随机选择一个主题 z_n（参数为 θ 的多项分布）
  - 选好每个单词的主题之后，以 $p(w_n|z_n, \beta)$ 的概率生成这个单词

LDA补充说明

- 对于语料库，文章数为M，单词个数V，这两个值是确定的
- 对于单个文章，它的单词个数N是不确定的，但是符合参数为 ϵ 的泊松分布
- 每个词语在实际处理的时候都被表示为长度为V的向量
- 每篇文章都是由N（变量）个词向量组成
- K：语料库全部文档需要训练出的主题总数
- β：(K, V)矩阵，每一行代表该主题的单词分布（狄利克雷分布）

LDA概率公式

$$p(\theta, \vec{z}, \vec{w}|\alpha, \beta) = \underbrace{p(\theta|\alpha)}_{\mathbf{Dir(\alpha)}} * \prod_{n=1}^{N} \underbrace{p(z_n|\theta)}_{\mathbf{Mult(\theta)}} * \underbrace{p(w_n|z_n, \beta)}_{\mathbf{Mult(\beta_{z_n})}}$$