
Date：2018.4.26

Author：baby_qian

Function: 使用 LDA 主题模型的方法来对一定规模的微博评论进行情感词提取。

（1）装载所需的语料，将每条微博评论当成一个文档统并存储为矩阵的形式（每行代表一个文档，每列代表文档中的词）；

（2）文本预处理操作，主要包括分词、去停用词、去标点符号等；

（3）统计并得到文档的 TF-IDF 值矩阵；

（4）设置主题数目 K，用 TF-IDF 矩阵训练 LDA 模型至收敛，分别得到（文档--主题）以及（主题--词）的概率分布；

（5）统计得到文档矩阵中所有出现过的词集 C，载入知网的正负情感词典并与 C 取交集作为种子情感词集 P 和 N；

（6）分别计算非种子词与正负种子词集权重概率的绝对距离，得到两个距离列表 D1,D2；

（7）选取阈值 threshold，判断非种子词的情感极性，得到新的情感词，并将其保存为文件导出；

（8） 评估新情感词的质量（评估方法待确定）

一、方案思路

（1）LDA 的两个假设

文档是由多个主题以一定的概率分布构成的，主题是由语料库中所有的基本词汇以一定的概率分布构成的；

（2）将每条微博评论当成一个文档，得到一个文档矩阵，计算文档矩阵的 TF-IDF 矩阵值；

（3）人为设置需要的主题数量，并使用全量的语料进行 LDA 模型训练，得到构成每个文档的主题的概率分布，以及构成每个主题的词汇的概率分布；

（4）计算非种子词与正负种子词的相似度

两个出发点：

A. LDA 是从全局语料库中进行主题概括和词汇概率分布计算的，因此在考虑某个候选词的情感极性时，也应当放在全局的语料中进行考量；

B. 对于主题模型，从感性上来理解，对于一个文档，同一概率分布下权重越大的几个主题应当具有较强的文档概括能力，并且权重相当的几个主题的概括能力也相当（似）；对于一个主题，同一概率分布下权重越大的几个词汇应当具有较强的主题概括能力，并且权重相当的几个词汇的概括能力也相当（似）。

在基本的词汇中查找种子情感词（来自知网），分别计算非种子词与正负种子词集概率权重的绝对距离，选取阈值 threshold，判断非种子词的情感极性，得到新的情感词。

二、部分实验结果图

1. 统计主题—词汇分布图

查看构成某一主题的基本词汇的概率权重分布，横坐标代表按序号排列的基本词汇，纵坐标为其对应的概率值，概率值越大表示该词对于该主题的重要性（该主题出现时该词的出现次数越频繁）越大。

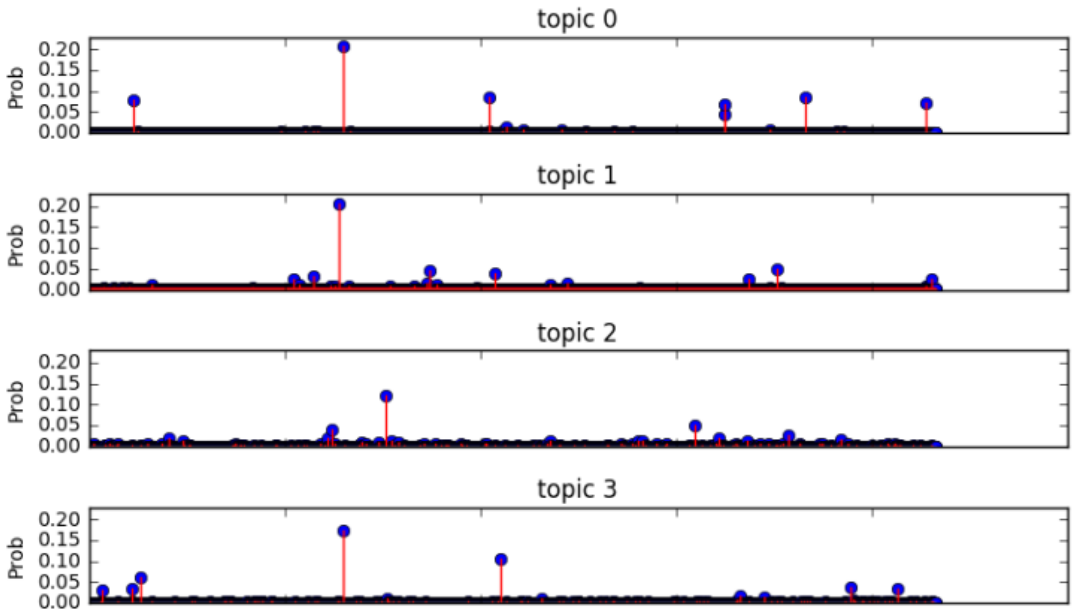

图 1(a)

```
# 获取主题--词汇概率分布

topic_word_distribute_proba_matrix = LDA_.topic_word_distribute(model=model, vectorizer=vectorizer)
```

```
type(topic_word): <class 'numpy.ndarray'>
shape: (100, 8648)
*Topic 0
- 回复 悠闲地 蘑菇 skm 魔王 破音 破破 我爱你 无明 老公
*Topic 1
- 喜欢 肯德基 小精灵 想要 可爱 黄金 精灵 单身汉 小朋友 明明
*Topic 2
- 女人 男人 哄哄 舒服 下次 呵护 眼睛 超好 浮云 精彩
*Topic 3
- 回复 憨憨 xkl 转圈 陌憨 s2017 ali 第一张 美好 摇尾巴
*Topic 4
- 杨洋 回复 围观 海洋 碧浪 偷乐 代言 微风 群体 带回家
*Topic 5
- 偷笑 艺人 地方 亚洲 佩服 身边 全能 尴尬 怒骂 多好
*Topic 6
- 陈伟霆 爱奇艺 恭喜 年度人物 明星 尖叫 vivo 演员 演技 偶像
*Topic 7
- 真的 东西 喜欢 理解 麻烦 前排 终于 萌萌 美的 微笑
*Topic 8
- 冬雨 兔子 表白 最美 影后 彩虹 冬叔 表演 裙子 情有独钟
*Topic 9
- 支持 不错 我会 网友 亲爱 一如既往 我家 钠盐 期待 爸爸
*Topic 10
- 宝宝 帮宝适 我家 拉拉 纸尿裤 放心 家里 分享 干爽 调皮
```

图 1(b)

图 1　主题 － 词汇分布统计图

2. 文档--主题统计分布图

查看构成某一文档的基本主题的概率权重分布，横坐标代表按序号排列的基本主题，纵坐标为其对应的概率值，概率值越大表示该主题对于该文档的重要性（该主题对该文档的概括能力越强）越大。

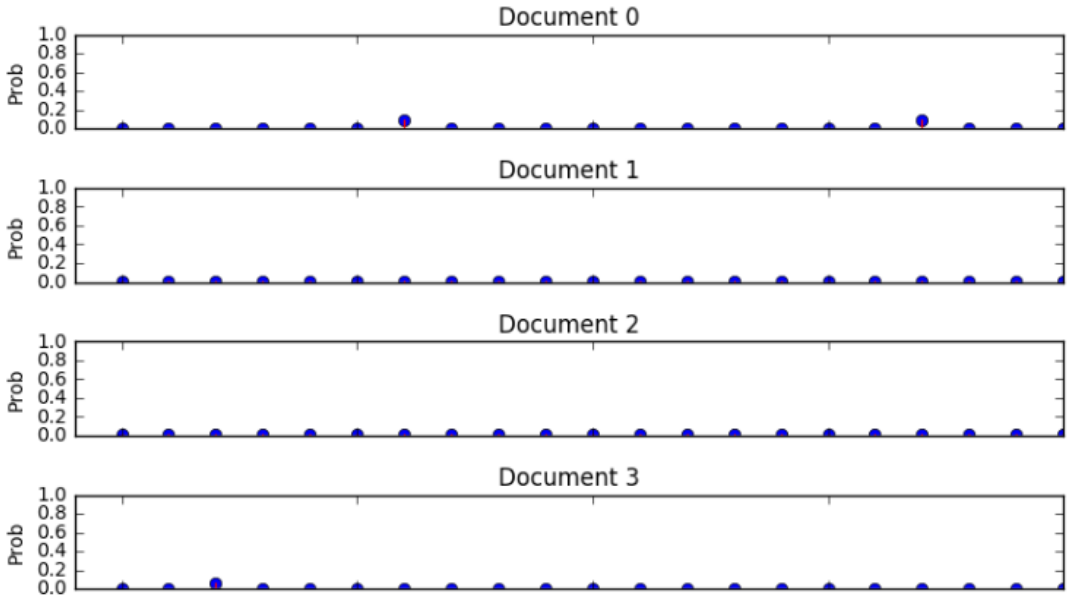

图 2(a)

```
# 获取文档--主题概率分布

doc_topic_distribute_proba_matrix = LDA_.text_topic_distribute(corpus=corpus, model=model)
```

```
type(doc_topic): <class 'numpy.ndarray'>
Shape: (9999, 100)


doc: 0 topic: [ 6 17  0 72 71]
doc: 1 topic: [22 89  0 71 70]
doc: 2 topic: [42 89  0 71 70]
doc: 3 topic: [89  2 42 58 83]
doc: 4 topic: [75 73 74 22  7]
doc: 5 topic: [65 12 72 71 70]
doc: 6 topic: [61 74 58 11 72]
doc: 7 topic: [20 67  0 72 71]
doc: 8 topic: [60 95 54 91  0]
doc: 9 topic: [60 39 20  0 72]
doc: 10 topic: [93  0 71 70 69]
```

图 2(b)

图 2　文档 － 主题分布统计图