基于 LDA 模型的微博主题抽取

魏林静 2120151045　王丹 2120151036

李克南 2120151004　郭一迪 2120150986

简介： 在这个 project 中要实现微博主题的抽取，输入微博内容，然后输出是这些微博的主题。解决这个问题我们小组采用的是 LDA 主题模型，并对其进行改进，使得系统更加适合微博的特征，挖掘出更准确的主题。

1　问题陈述

本次任务数据集来源是爬虫获取，将获得的数据集挖掘潜在的主题，该系统会输出设置的主题个数的单词的概率，并且由高到低排列，最终的评价标准是困惑度。本次任务可以分成如下几个步骤：

1、用数据爬虫从新浪微博上面抓取微博文本内容。
2、数据预处理，将用户信息和微博内容分开。
3、使用 NLPIR 汉语分词系统对微博内容文本进行分词。
4、对分词后的微博内容文本进行去停用词处理。
5、将分词后的文本放入到改进的 LDA 模型中进行微博主题提取。
6、用困惑度对该模型进行评价。

adfa, p. 1, 2011.

2 技术方案

解决问题所采用的技术和方法包括：数据爬取、数据清洗、数据预处理、LDA、困惑度

2.1 数据爬虫结果：

叫我红领领领巾　每天看到满微博的微信要收费了就觉得累，难道现在的人都傻了么，自己完全没判断力？
青衣　#云中歌#恭喜，一看演员就知道又一部小言经典被毁了。好好的小说非要把人物一一对应排成电视，又不是世界名著值得拍么，拍出来都是快速消费和阅读的产物，浪费钱浪费时间，没意义。没劲透了
Oo話不多小姐　别总是说他们炫富是靠他们的老爸，就算他们不靠他们的老爸也比你们过的好，瘦死的骆驼比马大！你们只是仇富的心理不平衡而已，你们就是一群傻逼自己没能力变的有出息，还不让别人过的好#scc陈俊宇#@SCC-JUNYU
张小瑶 #我在搜狐视频评论#《非常静距离》20130402 霍建华回应陈乔恩绯闻 主动加吻戏吓坏杨紫》“赞赞赞！”http://t.cn/zTyQkhK（分享来自于@搜狐视频空间）

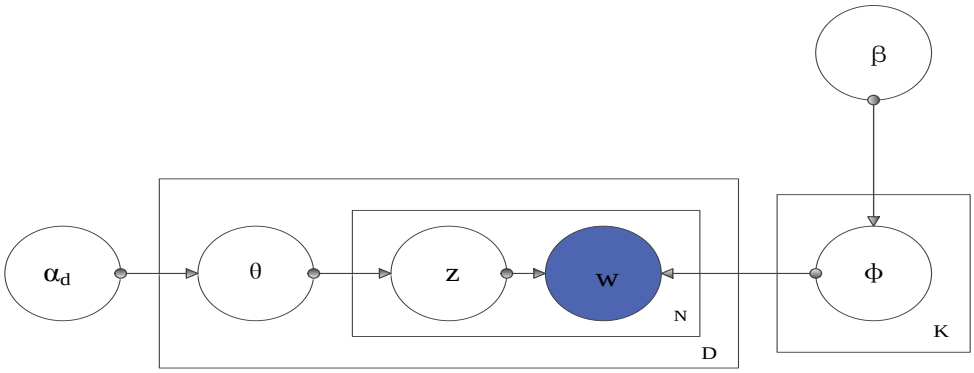
2.2 数据清洗结果：

叫我红领领领巾　每天看到满微博的微信要收费了就觉得累，难道现在的人都傻了么，自己完全没判断力？
青衣　#云中歌#恭喜，一看演员就知道又一部小言经典被毁了。好好的小说非要把人物一一对应排成电视，又不是世界名著值得拍么，拍出来都是快速消费和阅读的产物，浪费钱浪费时间，没意义。没劲透了
Oo話不多小姐　别总是说他们炫富是靠他们的老爸，就算他们不靠他们的老爸也比你们过的好，瘦死的骆驼比马大！你们只是仇富的心理不平衡而已，你们就是一群傻逼自己没能力变的有出息，还不让别人过的好#scc陈俊宇#@SCC-JUNYU
张小瑶 #我在搜狐视频评论#《非常静距离》20130402 霍建华回应陈乔恩绯闻 主动加吻戏吓坏杨紫》“赞赞赞！”http://t.cn/zTyQkhK（分享来自于@搜狐视频空间）

2.3 数据预处理结果

叫 我红 领 领 领巾 每天 看到 满 微 博 的 微信 要 收费 了 就 觉得 累 ，
难道 现在 的 人 都 傻 了 么 ， 自己 完全 没 判断力 ？
青衣 # 云中歌 # 恭喜 ， 一 看 演员 就 知道 又 一 部 小 言 经典 被 毁
了 。 好 好 的 小说 非 要 把 人物 —— 对应 排 成 电视 ， 又 不 是 世界
名著 值得 拍 么 ， 拍 出 来 都 是 快速 消费 和 阅读 的 产物 ， 浪费 钱
浪费 时间 ， 没 意义 。 没劲 透 了
Oo 話 不 多 小姐 别 总是 说 他们 炫 富 是 靠 他们 的 老 爸 ， 就算 他
们 不 靠 他们 的 老 爸 也 比 你们 过 的 好 ， 瘦 死 的 骆驼 比 马 大 ！
你们 只 是 仇富 的 心理 不 平衡 而已 ， 你们 就 是 一 群 傻 逼 自己 没
能力 变 的 有 出息 ， 还 不 让 别人 过 的 好 # scc 陈俊宇 #
@SCC-JUNYU
张小瑶 # 我 在 搜狐 视频 评论 # ≤ ≤ 非常 静 距离 ≥ 20130402 霍建华
回应 陈乔恩 绯闻 主动 加 吻 戏 吓坏 杨 紫 ≥ “ 赞赞 赞 ！ ”
http://t.cn/zTyQkhK （ 分享 来自 于 @搜狐视频空间 ）

3 LDA 方法简介：



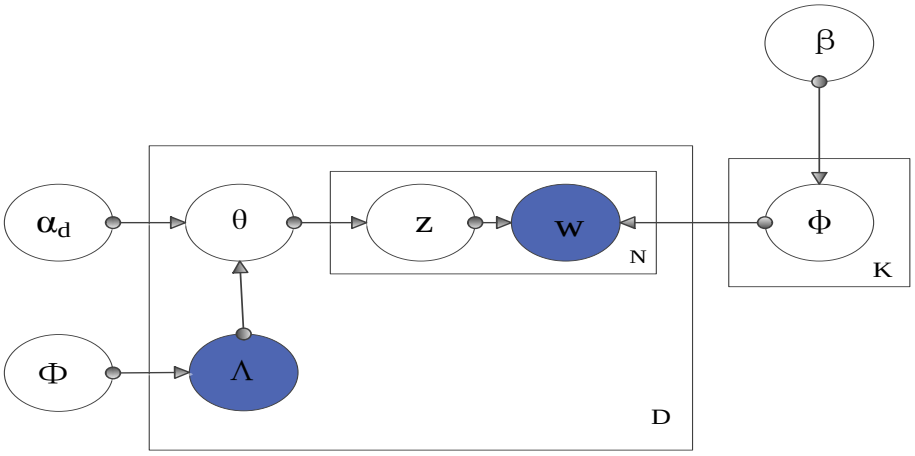
LDA 是一个多层的产生式概率模型，包含词、主题和文档三层结构。通过浅层的主题将词和文档关联起来[11]。文档可以由潜在主题的多项式分布来表示，主题可以由词语的集合的多项式分布来表示。文档中的每一个主题 Topic 的分布都是基于 Multinomial 分布，先验是基于 Dirichlet 分布（Multinomial 分布的共轭先验）；同样，每个主题 Topic 下单词都是基于 Multinomial 分布，先验是基于共轭先验的 Dirichlet 分布。

LDA 模型是发展至今最完备的主题模型，克服了 LSA 以及 PLSA 模型的缺陷，凭借着概率理论以及贝叶斯理论基础，在文本检索、

文本分类、图像识别、社交网络等领域得到了广泛的应用。但是利用 LDA 主题模型对微博进行实验分析，抽取出具有代表性的主题聚类的结果。从各个主题的关键词中可以看出主题聚类的效果不明显，没有办法从关键词中看出主题事件。而且，各个主题概率的值都较小，而且比较均衡，这些说明判别出该微博所属主题比较困难。并且在关键词中，发现有噪音词语，说明 LDA 主题模型不能够解决微博噪音大的问题。这个模型很难再一步进行主题聚类，也无法解决文本短的限制。目前对 LDA 模型的扩展主要基于三个方面：对参数的扩展；引入上下文信息；面向特定任务。

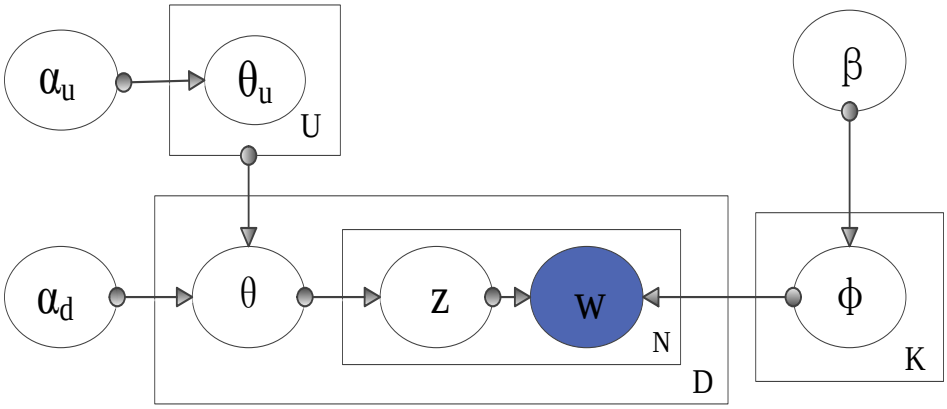
3.1 针对噪声大的 Content-LDA 模型

Content-LDA 模型通过建立两层主题结构，根据词语类别，在所有 K 个主题上，事先人为划分为 n（n<K）个主题类别，在此模型中称为 Label，视为主题层次的第一层，K 个主题视为主题层次的第二层,如图所示：



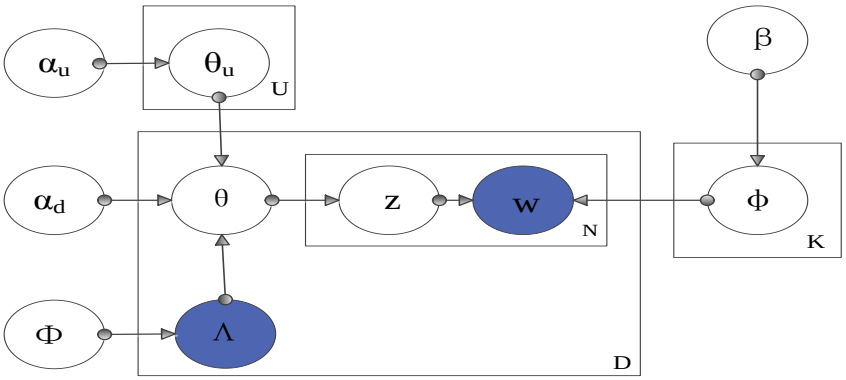
3.2 针对文本短的 User-LDA 模型

User-LDA 模型中，认为每个用户发布微博基于自己的兴趣，每条微博和用户的兴趣点有较大相关性。认为每个用户都有基于主题的概率分布，将用户发布的所有历史微博数据作为一个整体，利用此历史微博数据集对用户的主题概率分布进行分析，利用公式进行计算，得到用户基于主题的概率分布，如图 3.2 所示：



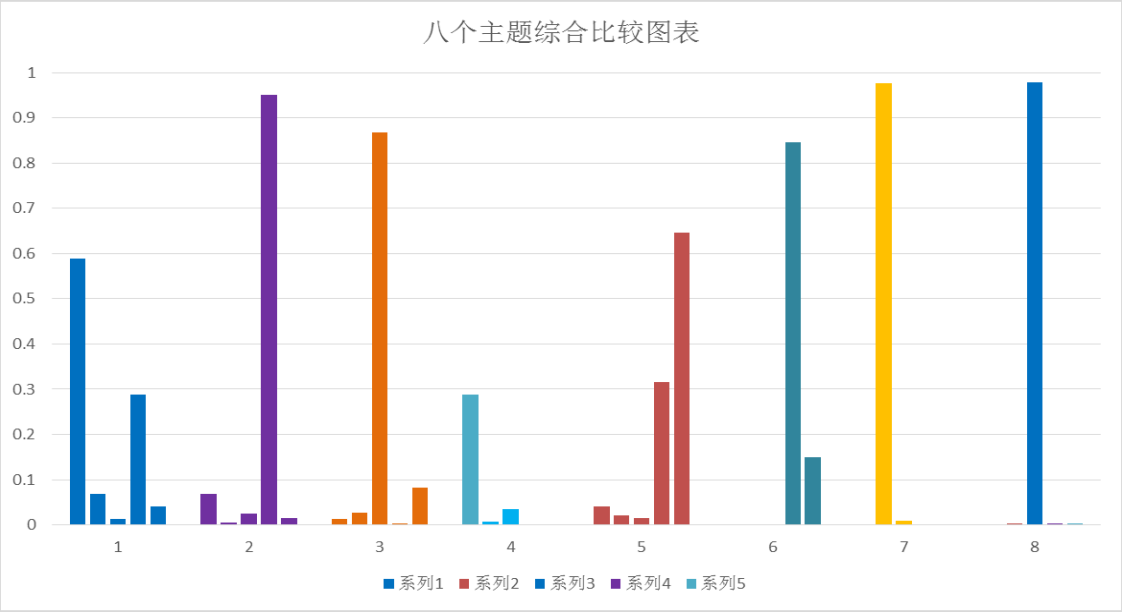
3.3 CU-LDA 模型

传统的 LDA 主题模型无法解决微博对于主题抽取的两点限制，解决噪音大的限制可以使用 Content-LDA 模型，解决微博文本短的限制可以使用 User-LDA 模型。本文提出的新模型称之为 CU-LDA 模型，该模型结合了 Content-LDA 模型和 User-LDA 模型的优点，可以同时解决两个限制。

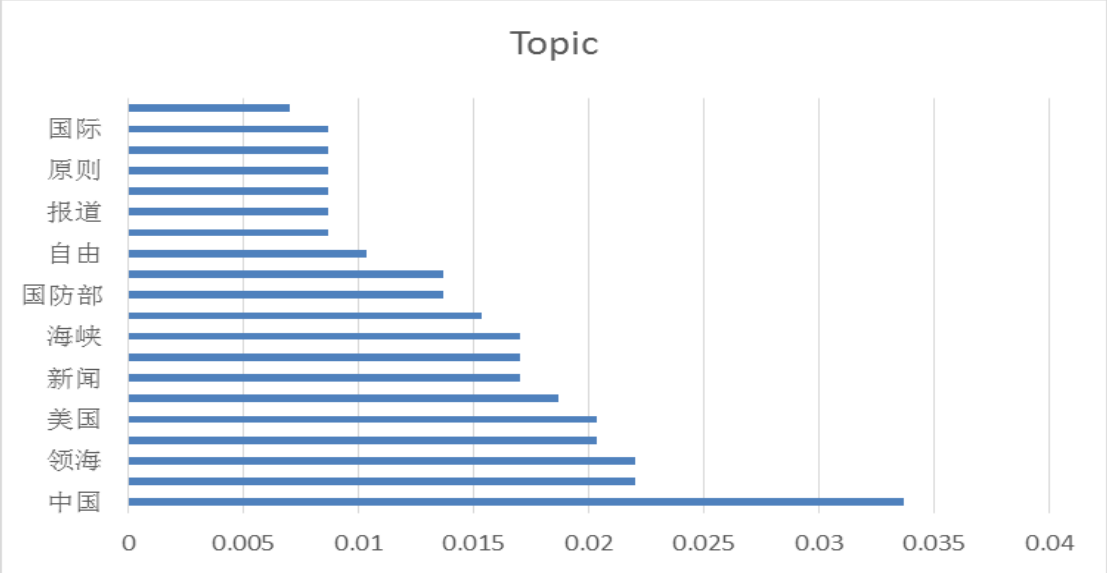


4 实现

4.1 八个主题结果展示，可以看出该系统主题的分布比较突出



4.2 其中一个主题的展示



5 评价结果：困惑度

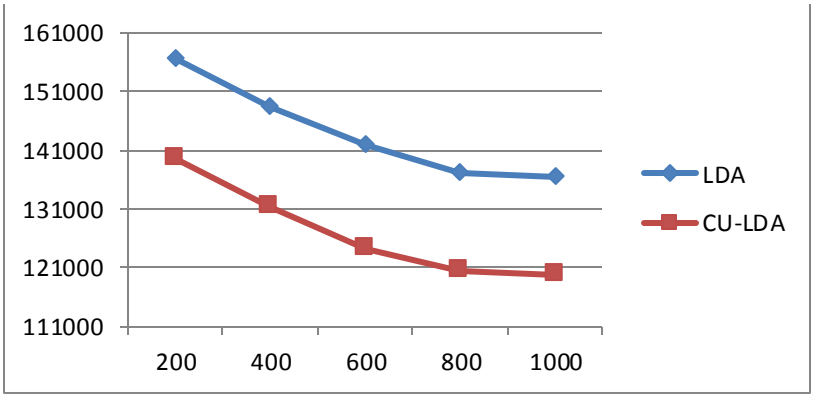
模型的困惑度（Perplexity）是验证模型泛化能力的一种有效方法。这种方法用以说明模型对于预测非测试数据是否有较好的能力。本实验采用 Perplexity 指标对实验结果进行度量．Perplexity 是度量概率图模型性能的常用指标，也是主题建模界常用的衡量方法，表示预测数据时的不确定度，取值越小表示性能越好，模型的推广性越高．Perplexity 定义如下：

$$Perplexity（W）=exp\{-\frac{\sum_m lnp(w_m)}{\sum_m N_m}\}$$

其中W为测试集，w_m为测试集中可观测到的单词, N_m为单词数。在相同的参数设置下，通过计算 Perplexity 来分析模型的推广能力，计算得出 LDA 与 CU-LDA 模型的 Perplexity

LDA 与 CU-LDA 模型的 Perplexity

迭代次数	LDA	CU-LDA
200	156420.4	139620.7
400	148289.1	131420.5
600	141860.9	124256.4
800	137170.7	120324.3
1000	136577.3	119830.6



模型的 Perplexity 对比图

通过与 LDA 模型的对比实验发现，在相同的参数条件下，随着迭代次数的增加，直到模型趋于收敛时，CU-LDA 模型的 Perplexity 均要小于 LDA，证明 CU-LDA 模型对微博进行分析，确实能够提高模型的性能和推广性。

综上所述，CU-LDA 的 Perplexity 指标优于传统的 LDA 模型，整体效果较好。