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。

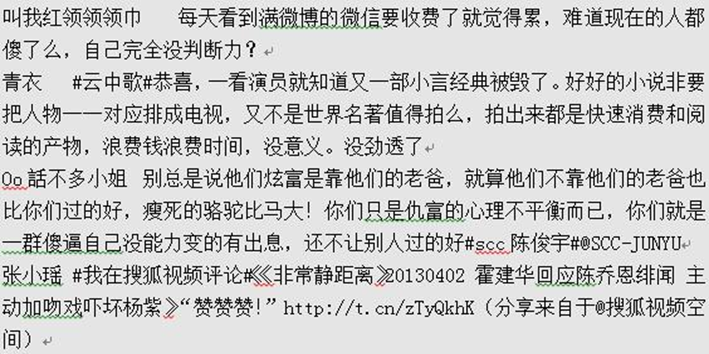
5、将分词后的文本放入到改进的LDA模型中进行微博主题提取。

6、用困惑度对该模型进行评价。

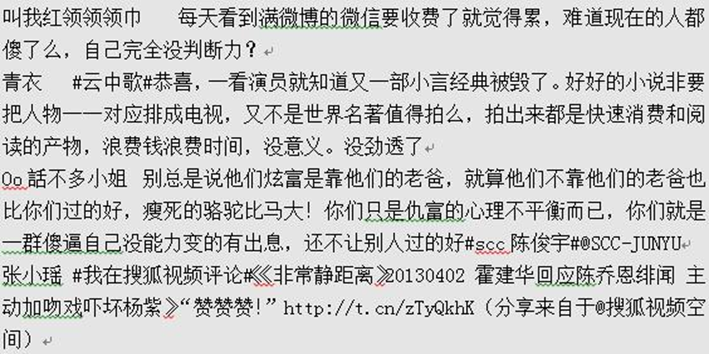
1. 技术方案

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

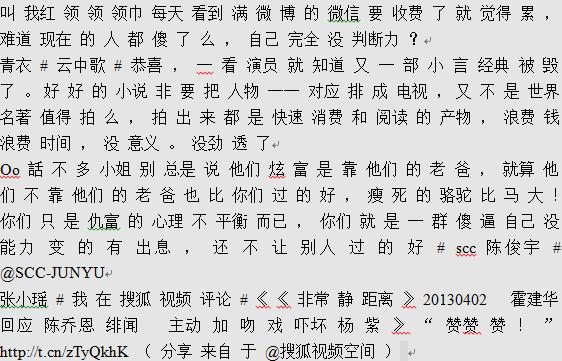
* 1. 数据爬虫结果：



* 1. 数据清洗结果：



* 1. 数据预处理结果



1. 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。

* 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：



* 1. 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：

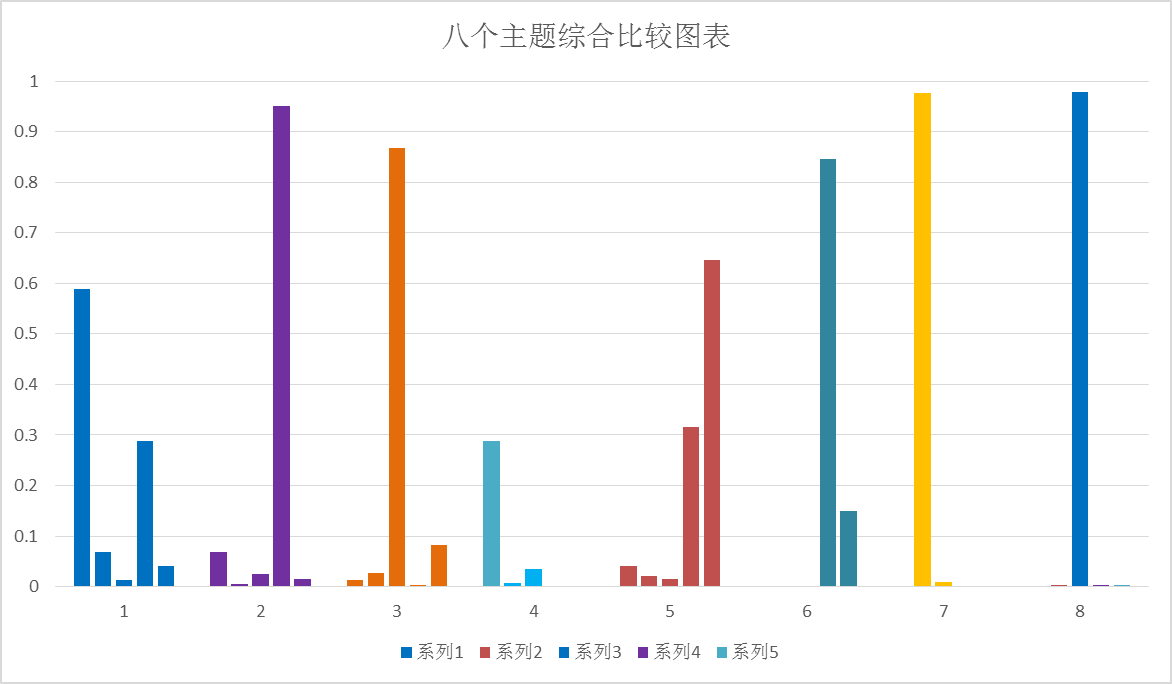


* 1. CU-LDA模型

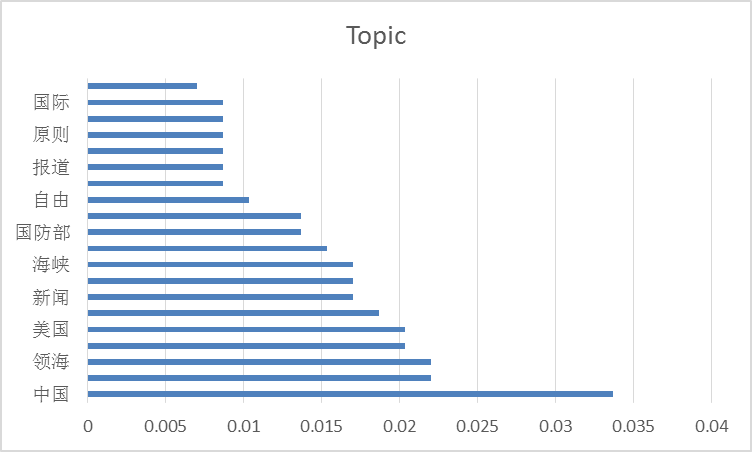
传统的LDA主题模型无法解决微博对于主题抽取的两点限制，解决噪音大的限制可以使用Content-LDA模型，解决微博文本短的限制可以使用User-LDA 模型。本文提出的新模型称之为CU-LDA模型，该模型结合了Content-LDA模型和User-LDA模型的优点，可以同时解决两个限制。



1. 实现
   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



* 1. 其中一个主题的展示



1. 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*﹛-﹜

其中Ｗ为测试集，为测试集中可观测到的单词,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。