“智慧政务”中的文本挖掘应用

摘要

发展智慧政务已成为提升政府科学管理水平和社会治理能力的重要手段。随着技术进步和创新社会治理理念新要求的提出，发展智慧政务成为政府部门提升政府科学管理和社会服务水平的新举措。2015 年 8 月，国务院印发了《促进大数据发展行动纲要》，提出运用大数据完善社会治理、提升政府服务和监管能力为大势所趋，大力推动政务大数据的发展和应用，成为提升政府治理能力的新途径。

在本次赛题中，针对群众留言分类问题，团队采用基于自主力机制的 XLnet 模型进行文本分类，并将数据集按照 6:2:2 的比例划分训练集、检验集、测试集，最终模型在检验集和测试集上的准确率达到 89%的准确率，性能显著。

针对热点问题挖掘问题，团队采用 KMeans，Single-Pass,LDA 与 Single-Pass 结合，Single-Pass 与 Xlnet 结合等方法进行热度问题聚类。根据群众反映问题量，反映问题时间的集中度，问题的点赞数与反对数，同时结合了近年来的两会关注度，定义了并定义了合理的热度评价指标，并汇总出热点问题表和对应的热点问题明细表。

针对答复意见的评价问题，为了得到更合理的答复意见评价方案，团队将文本进行分词、词性标注等预处理工作，团队最终从完整性、相关性、可读性、实时性、内聚性五个角度制定了评价方法。

对于赛题的复杂要求，应用环境的不确定性，我们采用数据挖掘知识结合机器学习的理念为基础，基于自然语言处理知识构建了文本分类模型；结合多角度因素考量，运用 single-pass 算法结合多种文本处理工作, 进行热点问题评选; 结合《政府网站发展指引》以及内聚性等角度, 构建了 CRRTC 模型对政府部门回复意见予以评价。而为了评估相关问题的准确性，本论文采取借鉴了国家《“互联网+政务服务”技术体系建设指南》等标准。

关键词：LDA，single-pass，XLnet，文本分类，热度指标

一、简介

1.1 数据挖掘意义

时至今日，大数据、云计算、人工智能技术飞速发展，互联网+思维渗透到各个行业，互联网+政务惠民服务开展的如火如荼，各个政务服务领域正在经历着信息化带来的变革，百姓也享受到了前所未有的巨大便利。作为直接与群众沟通交流的途径，为民服务的窗口，面对新的百姓期盼，新的技术发展，如何利用这些信息化成果，代替低效的人工劳动，更加快速响应市民服务热线反应的社会问题，有效分类进行处置，提高服务效能，更好的发挥市民服务热线桥梁纽带作用，是“智慧政务”作为构建社会现代治理体系和服务型政府的一项新课题。

自然语言处理（NLP）和文本挖掘是以半结构或非结构的自然语言文本为对象，其基本思想是从文本中提取适当的特征，将文本标识成计算机能够理解的形式，采用各种语义分析和文本挖掘的方法发现隐藏的知识模式，以用户可以理解和接收的形式输出，成为指导现实活动的有用的知识。我们也使用了机器学习方法，通过构建具有很多隐层的机器学习模型和海量的训练数据，来学习更有用的特征，从而最终提升分类或判断的准确性。利用预测模型发现、聚类分析、分类与回归、关联分析、异常和趋势发现等数据挖掘技术，提供超大数据量的数据分析和结果展示，实现灵活分类、评价、就具体问题给出准确完整的回复意见。

1.2 数据挖掘目标

我们希望构建一个能够从大规模文本数据集中发现隐藏的、潜在的、新颖的、重要的规律的模型，能够针对群众反映的具体问题，进行准确分类，能够在庞杂的反馈意见中找到与群众关系最为紧密的矛盾问题所在，针对问题给出准确合理完整的答复，让人民群众感受到政府是真切的希望为人民群众做好事做实事。

发展智慧政务已成为提升政府科学管理水平和社会治理能力的重要手段。随着技术进步和创新社会治理理念新要求的提出，发展智慧政务成为政府部门提升政府科学管理和社会服务水平的新举措。2015 年 8 月，国务院印发了《促进大数据发展行动纲要》，提出运用大数据完善社会治理、提升政府服务和监管能力为大势所趋，大力推动政务大数据的发展和应用，成为提升政府治理能力的新途径。

随着当前政务管理的透明度和开放性也大大增强，信息越来越公开，数据越来越开放，使得绝大多数企业和老百姓都能够第一时间获取到政务信息。政府作为政务信息的采集者、管理者和占有者，具有其他社会组织不可比拟的信息优势。政府部门在出台社会规范和政策时，采用大数据进行分析，可以避免个人意志带来的主观性、片面性和局限性，减少因缺少数据支撑而带来的偏差，降低决策风险。通过大数据挖掘和分析技术，可以有针对性地解决社会治理难题，针对不同社会细分人群，提供精细化的服务和管理。

“智慧来自大数据”，在电子政务的推进中必需掌握大量的知识，并具有在大量的政务数据和信息中挖掘知识的能力。数据挖掘就是从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中，抽取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识的过程。从更广义的角度来说，数据挖掘就是在一些事实或观察数据的集合中寻找模式的决策支持过程。因此，它除了处理传统数据库中的数值型的结构化数据外，还可以对文本、图形、图像、WWW 信息资源等半结构、非结构的数据进行挖掘。大数据处理可以进一步提升电子政务价值，使政府决策建立在大量数据分析的基础之上，使政策更加透明，并且可以创造出更大的价值。

1.3 挖掘流程

本文主要分为四大部分，预处理部分和文本聚类分类部分、热度评价部分和模型构建部分。其中预处理包括分词，去停用词，将词组向量化（word2vec）。

① 中文语言复杂多变，存在一词多义、词语歧义、隐含表达等问题[1]，需要

结合语境和上下文来推断所要表达的意思，给文本处理带来了很大的困难。
② 用户评论具有很大的随意性，尤其是结合了英语和网络语言，再加上用户不会注意结构语法的使用，无法正确判断词语的词性，更是加大了特征提取的难度。
③ 评论质量的高低因人而不同，结合用户的偏好、经验和领域知识的评论质量评估值得深入研究[2]。

二、数据预处理【还用到那些呢】
针对题目中所给的原始数据而言，原始数据中存在着大量的冗余信息、不一致信息、非结构化数据和有异常的数据，这将会严重的影响数据挖掘和建模的执行效率，甚至可能会导致数据挖掘的结果出现偏差，因此，必须先对给定的原始数据进行预处理操作，提高数据集的质量，并使得数据能够更好地适应我们的挖掘算法和模型。文本预处理技术主要包括分词、停用词处理、词性标注、实体抽取等操作。
在分析数据之后,我们可以看出数据有哪些特点,以及数据有哪些缺点。通过数据清洗可以将重复、多余的数据筛选清楚,将缺失的数据补充完整,将错误的数据纠正或者删除,最后整理成为我们可以进一步加工、使用的数据。

2.1 分词

众所周知，英文是以词为单位的，词和词之间是靠空格隔开，而中文是以字为单位，句子中所有的字连起来才能描述一个意思其特点是词与词之间没有明显的界限，从文本中提取词语时需要分词。本文采用两种分词方法，Python 开发的一个中文分词模块——jieba 分词，对问题和回答中的每一句话进行分词进行中文分词，以及中国科学院计算研究所开发的 NLPIR 汉语分词系统（http://ictclas.nlpir.org/）中的分词模块进行分词。

jieba 分词用到的算法为最短路径匹配算法该算法首先利用词典找到字符串中所有可能的词条,然后构造一个有向无环图。其中,每个词条对应图中的一条有向边,并可利用统计的方法赋予对应的边长一个权值，然后找到从起点到终点的最短路径，该路径上所包含的词条就是该句子的切分结果。

1、 jieba 分词
2、 NLPIR 汉语分词系统进行分词操作。
在本论文中对于上述两种方法视不同的分词需求进行使用。

2.2 去停用词

在文本处理中，停用词是指那些功能极其普遍，与其他词相比没有什么实际含义的词，它们通常是一些单字，单字母以及高频的单词，比如中文中的“我、的、了、地、吗”等，英文中的“the、this、an、a、of”等。对于停用词一般在预处理阶段就将其删除，避免对文本，特别是短文本，造成负面影响。

通过官琴等人实验对比分析，比较不同停用词表对于不同类型的文本数据的作用效果，对比了三大停用词表后，我们决定采用百度停用词表以及四川大学停用词表，它们分别针对于新闻报道类和邮件文献等类型的文本效果更好更适合，结合我们选题所给出的情况来看，既需结合新闻报道类的简明易懂，也需要对于问题意见给予反馈这就需要邮件文献这一类，贴近生活口语形式的表述。

本文所用的停用词，取自哈工大停用词表、百度停用词表、四川大学机器智能实验室停用词库、中文停用词表，进行汇总、去重之后，停用词表共有 2317 个停用词。

2.3 词性标注

词性标注（POS Tagging）是根据上下文关系，结合语义语境，按照一定的规则，确定词语的性

质并标注的过程。在中文中，一些词语有多种词性，不能单独的指定这个词语是哪个词性，需要将其放在上下文中，根据语境来标注，此时该词语的词性是唯一的。如：“他正在向上级报告今天的行程。”和“这周我需要向老师提交一个报告。”在前一个句子中“报告”是动词，后一个句子中“报告”是名词。常用的词性标注方法有：基于规则的标注算法、随机标注算法和前两者混合型的标注算法。

本论文中使用的方法为借助使用 NLPIR 汉语分词系统（http://ictclas.nlpir.org/），利用其提供的 API 实现进行词性标注。

2.4 实体抽取

实体抽取，又称命名实体识别（Named Entities Recognition，NER），主要任务是识别命名实体的文本范围，并将其分类为预定义的类别，学术上所涉及一般包含三大类，实体类、时间类、数字类和 7 个小类，比如人、地名、时间、组织、日期、货币、百分比，是问答系统、翻译系统、知识图谱的基础，早期的 NER 的方法主要由语言学家手工构造规则模板，选用特定特征，包括统计信息、标点符号、指示词、方向词、中心词等，以模式与字符串相匹配为主要手段，但是此方法需要大量人力构建语言模型、系统周期较长、知识更新较慢、移植性较差。随着机器学习应用，提出了基于统计学的方法，主要包括隐马尔科夫模型(HMM)、最大熵马尔科夫模型(MEMM)、支持向量机(SVM)、条件随机场(CRF)，基于统计方法的对特征选择要求较高，对语料库的依赖较大。由于统计学方法对于语料库依赖较强，本论文中选取的方法为借助中国科学院计算研究所开发的 NLPIR 汉语分词系统（http://ictclas.nlpir.org/），利用其提供的 API 实现实体抽取，便于后续计算。

2.5 文本长度处理

首先注意到的一点是， 随着文本长度的增加，所需显存容量也会随之呈现线性增加， 运行时间也接近线性，因此，我们往往要做一个权衡，对于不同任务而言， 文本长度所带来的影响力并不相同。就分类问题而言， 到一定的文本长度后，模型表现就几乎没有变化了，这个时候再去提升文本长度意义就不大了。因此我们需要对于文本进行截断操作，以保证对于模型性能提升的效率。

2.6 缺失值处理

常用的方法有以下几种:

1.若缺失值并没有很多，可以考虑删除它们，因为删除后对整体数据影响不大.

2.使用一个全局常量填充 ——譬如将缺失值用”Unknown”等填充,但是效果不一定好，因为,因为算法可能会把它识别为一个新的类别,一般很少用

3.使用均值或中位数代替——优点:不会减少样本信息,处理简单。缺点:当缺失数据不是随机数据时会产生偏差。对于正常分布的数据可以使用均值代替，如果数据是倾斜的，使用中位数可能更好。

因为我们的数据集中需要用到时间等信息，对于日期的预处理共有两步，

1、有的日期没有时分秒，没有时分秒的话，默认 00:00:00。

2、日期格式混乱，有的日期是 2020/5/7 格式，有的是 2020-5-7

2.7 离群值处理

在统计学中，离群值是指与其他观测值有显著差异的数据点。在大多数较大的数据采样中，某些数据点与采样均值的距离会超出合理范围。这可能是由于偶然的系统错误或产生假设的概率分布族的理论中的缺陷所致，也可能是由于某些观测值离数据中心很远。因此，异常点可能指示错误的数据，错误的程序或某些理论可能无效的区域。但是，在大样本中，预期会有少量异常值（并非由于任何异常情况）。在一个较大的数据集里有极大值，是很正常的事情，就像姚明的身高和正常人的相比就是极大值，但是不能认为是异常值，这只是极大值。

在我们的数据集中经过反复验证检查发现，的确存在离群值，为了不影响到模型性能，根据其分布的特性采用数据分段处理的办法。

2.8 去重处理

数据集中属性值相同的记录被认为是重复记录，通过判断记录记录间的属性值是否相等来检测记录是否相等，相等的记录合并为一条记录(即合并/清除)，合并/清除是消重的基本方法。在第二题数据集附件 3 中，存在有些用户对于同一个问题重复反映为避免对于这类无效数据对于热度指数的影响，采取以下去重处理:

1.将文本中的标点符号去除，只留下纯文本进行对比，在聚类后的每一类中，分别对每一条留言中的每一个词进行对比，对于两个或多个文本中的相似度高达 98%以上的留言只留下其中一条，以减少重复留言对于热度评价的影响。

2.9 补齐数据

补全是更加常用的缺失值处理方式，通过一定的方法将缺失的数据补上，从而形成完整的数据记录对于后续的数据处理、分析和建模十分重要。在第三题数据集附件 4 中存在日期数据的缺失问题，为了避免对模型构建造成影响以及不影响结果的准确率，关于日期的预处理补全为，有的日期没有时分秒，没有时分秒的话，默认 00:00:00。通过利用时间戳的方式进行计算对比。

2、日期格式混乱，有的日期是 2020/5/7 格式，有的是 2020-5-7,

2.10 格式规范

数据格式不同会导致对数据无法进行同样的操作处理，引起较大误差，对于第二题数据集附件 3 中存在格式不统一的情况如下图 X 所示日期格式混乱, 有的日期是 2020/5/7 格式, 有的是 2020-5-7。在 excel 中使用公式将格式全部统一为 YYYY—MM—DD HH：MM：SS，这可以解决很多数据合并时可能遇到的问题。

2019/5/22 12:24:16
2019-08-01 00:00:00

关于日期的预处理共有两步，

1、有的日期没有时分秒，没有时分秒的话，默认 00:00:00。

2、日期格式混乱，有的日期是 2020/5/7 格式，有的是 2020-5-7,

2.11 数据规约

数据规约就是缩小数据挖掘所需的数据集规模，具体方式有维度规约与数量规约。在数据集成与清洗后，我们能够得到整合了多数据源同时数据质量完好的数据集。但是，集成与清洗无法改变数据集的规模。我们依然需通过技术手段降低数据规模，这就是数据规约（Data Reduction）。数据规约采用编码方案，能够通过小波变换或主成分分析有效的压缩原始数据，或者通过特征提取技术进行属性子集的选择或重造。降低无效、错误数据对建模的影响，提高建模的准确性；少量且具有代表性的数据将大幅缩减数据挖掘所需时间；降低存储数据的成本。

在大数据集上进行复杂的数据分析和挖掘将需要很长的时间，数据规约可以产生更小的但保持原数据完整性的新数据集。在规约后的数据集上进行分析和挖掘将更有效率。本论文中通过对第三题中涉及到的多个属性值进行数值规约已求达到更佳的模型性能，避免了一些计算数据对于结果的影响。

三、相关工作

3.1 任务重定义

官方赛题的要求是：聚焦于智慧政务”中的文本挖掘应用，对于群众所反映的意见问题，进行分类，选择甄别重点热点问题，在庞杂的众多问题中关注群众最关心的问题，以及针对所反映的问题给出的答复做一个合理完整的评价。

对于复杂多样的留言意见，我们需要找出是属于哪类问题，通过文本内容能够分析辨别出是属于哪一类民生问题。并且能够针对于所反映的问题给出一个热度评价指数，这需要结合多方面因素考量对于反映次数较高、支持关注人数多、群众关心度高的问题，应当给出一个合理的排名，让与群众生活密切相关的重点问题得到更为及时有效的解决。给出一套有理可循的评价指标能显著提升对于问题回复的质量，有助于规范相关部门回复时的专业性、严谨性，有利于“智慧政务”更长久的发展，让人民群众真切感受到便利便捷。

3.2 流程概览

本文选取来自互联网公开来源的群众问政留言记录，及相关部门对部分群众留言的答复意见的信息为数据集，对其中的群众文本评论的类别进行区分，构建一个智能合理的模型进行自动分类，不仅有助于提升效率，也有助于减轻工作人员负担，采用 F-Score 对分类结果进行评价。在繁杂的问题中找到热点难点重点问题及时响应是非常重要的，因此我们结合第二题的需求，通过时间、数量等多方面因素考量对于反馈的意见进行热度排名，让真正和群众利益密切相关的问题得以切实解决。无规矩不成方圆，合理的政策更需要完善的评级体系予以管理，我们建立了一套完整而详尽的评价方案用以对答复意见进行考量，提出了一种电子政务领域中文本评论数据的质量评估方法。

对于第一题中我们按照赛题给出的评判标准对于留言内容进行分类，为了尽可能通过学习完整留言内容进行分类，我们对比了近年来较为先进的预训练模型，为了获得尽可能多的上下文记忆信息，经过对 BERT/XLnet 模型尝试后，我们最终选择了 XLnet 这一模型。接下来我们将在 3.3.1 章节详细介绍 XLnet 在本题中的应用。

在第二题，热度评价指标建立需要对于意见反映数量，关注度，时间范围等角度考虑问题，首先需要对各留言详情进行聚类分析我们分别采用三种办法，从机器学习的角度到数据挖掘的方法分别进行了评估和建立。在下文 3.3.2 中将简要介绍所尝试的三种方法。

第三题主要是从五个方向考虑第三题的文本:全面性、相关性、可读性、实时性、内聚性。具体操作为先将所有文本打上标签，不同类型的标签分别处理，然后计算每条文本的 5 个方向的分数，并按照不同的标签分别进行数据规约操作，规约至[0,10]分，然后五个属性之间进行加权平均，作为本留言的成绩。

本文的主要工作如下：

1.政务特征词的提取

本文对评论质量的分析方法是基于政务特征词进行的，因此需要对问题反馈和答复意见的特征词进行提取。本文首先对选取的数据集进行分词、词性标注和停用词处理等预处理操作，分析用户评论的语法结构，提出一定的句法规则，并根据这些句法规则提取答复意见中的名词或名词性短语作为政务意见的特征词。

2.根据留言内容进行分类

根据群众留言的内容进行一级分类标签的标注，有利于减轻基层工作人员的负担，提升准确率，也方便于后续题目的解决。例如，在第二题中可以先进行分类，针对于每一个二级分类下的留言问题再进行聚类操作，可以有效减少工作量利于提升工作效率与准确性，让重点问题更为突出。

3.通过数据分析得出热度评价指标

通过对于数据集的多角度分析，借鉴新闻热度评价、微博热度评价以及电子商务信息分析的方法，针对于某一时段内群众集中反映的某一问题可称为热点问题。从多种因素角度考虑，以时间数量频率群众关心度等多个方面考虑，构建一个完善的热度评价指标，有助于重点问题，热点问题，难点问题得以快速解决，真正让智慧政务帮助到每一个需要帮助的群众。

4.答复意见评论质量分析模型的构建

分析了答复意见文本评论的特点，根据工作人员给出的的答复意见质量的 XX 隐含函数，提出了影响评论质量的五个因素：完整性(Completeness)、相关性（Relativity）、可读性（Readability）、实时性（Real-time）、内聚性（Cohesion）；相关性（Relativity）、完整性(Integrity)和可解释性（Interpretability），由此提出了一个对于的答复意见评价的质量分析模型（Comment Quality Model，CQM）来计算每条评论的质量分数，该模型的计算是在【特征主题层次格】的基础上进行的。为了验证本文提出的答复意见质量分析模型 CQM 的正确性，

3.2.3 第三题

3.3 模型介绍【比如所用到的模型是什么样的 每一部分是什么、作用应用】

3.3.1 【第一题构建模型】

文本分类（Text Categorization）是指计算机将一个文本划分为预先给定的某一个类别或某几个类别的过程。常用的文本分类方法有以下几种，我们对于以下模型经过多方面尝试，选取表现最好的结果作为最终模型的构建。

5.bert——nlp 领域的里程碑式模型

预训练模型的出现将 NLP 带入了一个全新时代，而 BERT 将为 NLP 带来里程碑式的改变。BERT 模型，在机器阅读理解顶级水平测试 SQuAD1.1 中表现出惊人的成绩：全部两个衡量指标上全面超越人类，并且还在 11 种不同 NLP 测试中创出最佳成绩，包括将 GLUE 基准推至 80.4%（绝对改进 7.6%），MultiNLI 准确度达到 86.7% 。从诞生之初就展现出惊人的性能，并且预训练的 BERT 表示可以通过一个额外的输出层进行微调，适用于广泛任务的最先进模型的构建，比如问答任务和语言推理，无需针对具体任务做大幅架构修改。

BERT 模型是以 Transformer 编码器来表示，所以我们将在下文介绍 Transformer 编码器。使用该模型在神经机器翻译及其他语言理解任务上的表现远远超越了现有算法。

在 Transformer 之前，多数基于神经网络的机器翻译方法依赖于循环神经网络（RNN），后者利用循环（即每一步的输出馈入下一步）进行顺序操作（例如，逐词地翻译句子）。尽管 RNN 在建模序列方面非常强大，但其序列性意味着该网络在训练时非常缓慢，因为长句需要的训练步骤更多，其循环结构也加大了训练难度。与基于 RNN 的方法相比，Transformer 不需要循环，而是并行处理序列中的所有单词或符号，同时利用自注意力机制将上下文与较远的单词结合起来。通过并行处理所有单词，并让每个单词在多个处理步骤中注意到句子中的其他单词，Transformer 的训练速度比 RNN 快很多，而且其翻译结果也比 RNN 好得多。

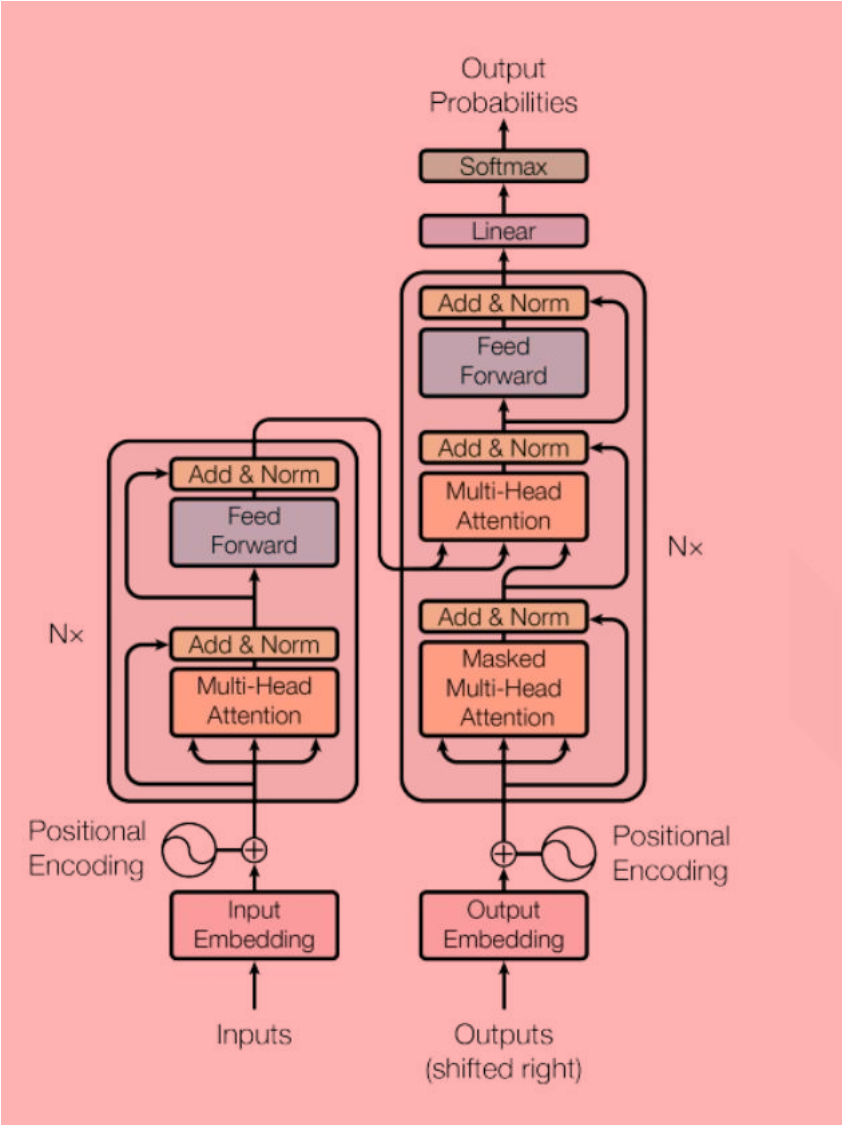


图 1. Transformer 结构图

其中 Encoder 层，由 N=6 个相同的 layers 组成，每一层包含两个 sub-layers. 第一个 sub-layer 就是多头注意力层（multi-head attention layer）然后是一个简单的全连接层。 其中每个 sub-layer 都加了 residual connection（残差连接）和 normalisation（归一化）。

Decoder 层，则是由 N=6 个相同的 Layer 组成，但这里的 layer 和 encoder 不一样， 这里的 layer 包含了三个 sub-layers，其中有一个 self-attention layer, encoder-decoder attention layer 最后是一个全连接层。前两个 sub-layer 都是基于 multi-head attention layer。这里有个特别点就是 masking, masking 的作用就是防止在训练的时候 使用未来的输出的单词。比如训练时，第一个单词是不能参考第二个单词的生成结果的。Masking 就会把这个信息变成 0，用来保证预测位置 i 的信息只能基于比 i 小的输出。

有两种常用的注意力函数，一种是加法注意力(additive attention)，另外一种是点乘注意力(dot-productattention)，论文所采用的就是点乘注意力，这种注意力机制对于加法注意力而言，更快，同时更节省空间。

“Scaled dot-product attention”具体的操作有三个步骤：

1.每个 query-key 会做出一个点乘的运算过程，同时为了防止值过大除以维度的常数

2.最后会使用 softmax 把他们归一化

3.再到最后会乘以 V (values) 用来当做 attention vector

数学公式表示如下：

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$$

网络的表达能力还有一些简单所以提出了多头注意力机制（multi-head attention）。multi-head attention 则是通过 h 个不同的线性变换对 Q，K，V 进行投影，最后将不同的 attention 结果拼接起来，self-attention 则是取 Q，K，V 相同。

$$\begin{aligned}\text{MultiHead}(Q, K, V) &= \text{Concat}(\text{head}_1, ..., \text{head}_\text{h}) \\ \text{where head}_\text{i} &= \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)\end{aligned}$$

Transformer 架构中除了主要的 Encoder 和 Decoder，还有数据预处理的部分。Transformer 抛弃了 RNN，而 RNN 最大的优点就是在时间序列上对数据的抽象，所以文章中作者提出两种 Positional Encoding 的方法，将 encoding 后的数据与 embedding 数据求和，加入了相对位置信息。

$$\begin{aligned}PE_{(pos,2i)} &= sin(pos/10000^{2i/d_{\text{model}}}) \\ PE_{(pos,2i+1)} &= cos(pos/10000^{2i/d_{\text{model}}})\end{aligned}$$

这里使用了两个构造函数 sin、cos。pos 用来表示单词的位置信息，比如第一个单词啦，第二个单词什么的。而 i 用来表达 dimension 现在的例子里，dmodel 是 512， 那 i 应该是 0 到 255. 这里为了好说明，如果 2i= dmodel, PE 的函数就是 sin(pos/10000)，那它的波长就是 10000*2pi， 如果 i=0，那么他的波长就是 2pi. 这样的 sin, cos 的函数是可以通过线性关系互相表达的。

Transformer 是第一个用纯 attention 搭建的模型，不仅计算速度更快，在翻译任务上也获得了更好的结果。该模型彻底抛弃了传统的神经网络单元，为我们今后的工作提供了全新的思路。

与 Peters et al. (2018) 和 Radford et al. (2018)不同，论文不使用传统的从左到右或从右到左的语言模型来预训练 BERT。相反，使用两个新的无监督预测任务对 BERT 进行预训练。

任务 #1：Masked LM

从直觉上看，研究团队有理由相信，深度双向模型比 left-to-right 模型或 left-to-right and right-to-left 模型的浅层连接更强大。遗憾的是，标准条件语言模型只能从左到右或从右到左进行训练，因为双向条件作用将允许每个单词在多层上下文中间接地“see itself”。 为了训练深度双向表征，我们采取了一个直接的方法，随机遮蔽输入 token 的某些部分，然后预测被遮住的 token。我们将这一步骤称为「masked LM」(MLM)，不过它在文献中通常被称为 Cloze 任务 (Taylor, 1953)。在这种情况下，对应遮蔽 token 的最终隐藏向量会输入到 softmax 函数中，并如标准 LM 中那样预测所有词汇的概率。在所做的所有实验中，我们随机遮住了每个序列中 15% 的 WordPiece token。与去噪自编码器 (Vincent et al., 2008) 相反，我们仅预测遮蔽单词而非重建整个输入。

虽然这确实能让研究团队获得双向预训练模型，但这种方法有两个缺点。第一个是，如果常常把一些词 mask 起来，未来的 fine tuning 过程中模型有可能没见过这些词，这个量积累下来还是很大的。因为作者在他的实现中随机选择了句子中 15%的 WordPiece tokens 作为要 mask 的词。为了解决这个问题，作者在设计 mask 的时候，使用如下的方法。

1.80%的概率真的用[MASK]取代被选中的词。比如 my dog is hairy -> my dog is [MASK]

2.10%的概率用一个随机词取代它：my dog is hairy -> my dog is apple

3.10%的概率保持不变: my dog is hairy -> my dog is hairy

为什么要以一定的概率保持不变呢？ 如果 100%的概率都用[MASK]来取代被选中的词, 那么在 fine tuning 的时候模型可能会有一些没见过的词。那么为什么要以一定的概率使用随机词呢？这是因为 Transformer 要保持对每个输入 token 分布式的表征，否则 Transformer 很可能会记住这个[MASK]就是"hairy"。至于使用随机词带来的负面影响，文章中说了,所有其他的 token(即非"hairy"的 token)共享 15%*10% = 1.5%的概率，其影响是可以忽略不计的。

使用 MLM 的第二个缺点是每个 batch 只预测了 15%的 token，这表明模型可能需要更多的预训练步骤才能收敛。团队证明 MLM 的收敛速度略慢于 left-to-right 的模型（预测每个 token），但 MLM 模型在实验上获得的提升远远超过增加的训练成本。

任务 #2：下一句预测

很多重要的下游任务（如问答（QA）和自然语言推断（NLI））基于对两个文本句子之间关系的理解，这种关系并非通过语言建模直接获得。为了训练一个理解句子关系的模型，我们预训练了一个二值化下一句预测任务，该任务可以从任意单语语料库中轻松生成。具体来说，选择句子 A 和 B 作为预训练样本：B 有 50% 的可能是 A 的下一句，也有 50% 的可能是来自语料库的随机句子。

可以看到在许多基准测试中都有 BERT 的身影，如此强大的模型，在本题中，我们 XX。

6.XLnet——通用自回归预训练方法

Xlnet，则是 bert 的改进版，在这之前也有很多公司对 bert 进行了优化，包括百度、清华的知识图谱融合，微软在预训练阶段的多任务学习等等，但是这些优化并没有把 bert 致命缺点进行改进。xlnet 作为 bert 的升级模型，主要在以下三个方面进行了优化：

1.采用 AR 模型替代 AE 模型，解决 mask 带来的负面影响

2.双流注意力机制

3.引入 transformer-xl

首先介绍两种语言模型：

1.自回归语言模型（Autoregressive LM，AR）

AR 的思路很简单，对于一个文本序列，AR 单方向地逐个进行预测，也就是最大化概率

$$p(x) = \prod_{i=1}^{n} p(x_i | x_{<i})$$

可以看出，这就是传统的单向语言模型（Language Model, LM）。这样的模型缺点很明显，就是只能看到单方向的信息（上文或下文），不过这种单方向的模式更加适用于一些生成类的 nlp 任务，因为这些任务在生成内容的时候就是从左到右的，这和 AR 的模式天然匹配。

2.自编码语言模型（Autoencoder LM，AE）

BERT 采用了不同于 AR 的模式——随机 mask 掉一些单词，训练过程就是根据上下文对这些单词做预测最大化，而这个其实就是（Denoising Autoencoder ,DAE）的典型思路。那些被 Mask 掉的单词就是加入的噪音, BERT 就可以认为是对其“去噪”。这样的做的好处就是能够结合上下文的信息，但是由于在预训练过程中额外地加入了[Mask]标记，这就导致了预训练和后面的微调阶段的数据的不一致。

相比较于 AR，采用了 AE 模型的 BERT 具有利用到双向上下文信息的优点，但是这样也导致了

一些问题的出现，一个是预训练和微调阶段的不一致；另一个是 BERT 模型有一个假设并不符合实际情况，它认为所有的[Mask]之间是相互独立的，举一个例子，对于“自然语言处理”这句话，假设 BERT 的输入是“自然语言[Mask] [Mask]”，那么 BERT 要优化的概率就是 P(处|自然语言)×P(理|自然语言)，而对于传统 LM 来说，则应是 P(处|自然语言)×P(理|自然语言处)，我们可以看出 BERT 忽略了[Mask]之间的相关性。

既然两种模型都有各有所长，那么能不能用一个模型将两者的优点结合起来呢？而这就是 XLNet 的出发点，它在实现双向 Transformer 编码的同时，还能够避免 BERT 所产生的那些问题。它基于 AR 模型，采用了一种全新的方法叫做 Permutation Language Modeling，先简单说一下流程，因为采用的是 AR 的模式，因此它是从左向右的输入，也就是说只能看到预测单词的上文，而我们希望在看到的上文中能够出现下文单词，这样就能在只考虑上文的情况下实现双向编码，为了到达这样的效果，XLNet 将句子中的单词随机打乱顺序，这样的话对于单词 xi，它原先的上下文单词就都有可能出现在当前的上文中了。

在改变了顺序以后，模型又产生了新的问题，那就是不知道要预测句子中的哪一个单词，这种问题在传统的 AR 模型中是不存在的，因为 AR 预测的永远是序列的下一个位置的单词，但是对于单词顺序被打乱的 XLNet，我们无法根据上文知道要预测的是哪一个位置的单词。具体来说，对于句子[x1,x2,x3,x4]，打乱顺序后对第三个位置进行预测，得到的上文信息是 x2,x4，可是我们却并不知道要预测的是哪一个单词，若打乱后顺序为 2413，那就是 x1，若是 2431，就变为 x3，所以只根据上文 x2,x4 无法准确预测出某一个单词，为了解决这个问题，XLNet 引入了 Two-Stream Self-Attention 机制。另外，其实对于排列在比较靠前位置的单词，它的上下文单词数量较少，预测难度较大， 因此 XLNet 只做部分预测，比如说只预测后 1/K 个单词，这样可以加快模型收敛速度。

XLNet 利用了 Transformer-XL，以此来获得更长距离的单词依赖关系，Transformer-XL 的详细内容可以参见之前的总结，这里说一下它在 XLNet 的用法，将一个长序列分成两个片段，

$\tilde{x} = s_{1:T}, x = s_{T+1:2T}$，将它们分别被重新排列，首先对于片段 $\tilde{x} = s_{1:T}$，利用 Content Stream 可以得到 $\tilde{h}(m)$，那么对于片段 $x = s_{T+1:2T}$，$h_{z_t}^{(m)}$ 可计算为：

$$h_{z_t}^{(m)} \leftarrow \text{Attention}\left(\mathbf{Q} = h_{z_t}^{(m-1)}, \mathbf{KV} = \left[\tilde{\mathbf{h}}^{(m-1)}, \mathbf{h}_{\mathbf{z} \le t}^{(m-1)}\right]; \theta\right)$$

序列之间建模，上面的内容都是关于序列中的单词之间的关系，而 nlp 中还有一些下游任务是在句子层级上的，在 BERT 中是通过 Next Sentence Prediction 的方法来进行适应的，而 XLNet 论文经过研究后认为 Next Sentence Prediction 对其没有什么帮助，因此 XLNet 提出了 Relative Segment Encoding，这个可以说借鉴了 Transformer-XL 中相对位置的思想，只判断两个单词是否在同一个 segment 中，而不是判断它们各自属于哪个 segment。

XLNet 模型可以说是现在预训练模型中的一个集成之作，它通过 Permutation Language Modeling 巧妙地将 LM 与 BERT 模型中各自的优点结合了起来，再加上引入的 Transformer-XL 和 Relative segment encoding 等技术，使得 XLNet 在模型表现上更上一层楼。在本题中，我们 XX。

文本聚类

聚类是把相似的对象通过静态分类的方法分成不同的组别或者更多的子集（subset），这样让在同一个子集中的成员对象都有相似的一些属性，常见的包括在坐标系中更加短的空间距离等。

数据聚类算法可以分为结构性或者分散性。结构性算法利用以前成功使用过的聚类器进行分类，而分散型算法则是一次确定所有分类。结构性算法可以从上至下或者从下至上双向进行计算。从下至

上算法从每个对象作为单独分类开始，不断融合其中相近的对象。而从上至下算法则是把所有对象作为一个整体分类，然后逐渐分小。
分布式聚类算法，是一次性确定要产生的类别，这种算法也已应用于从下至上聚类算法。
基于密度的聚类算法，是为了挖掘有任意形状特性的类别而发明的。此算法把一个类别视为数据集中大于某阈值的一个区域。DBSCAN 和 OPTICS 是两个典型的算法。
许多聚类算法在执行之前，需要指定从输入数据集中产生的分类个数。除非事先准备好一个合适的值，否则必须决定一个大概值，关于这个问题已经有一些现成的技术。

聚类技术将数据元组(即记录，数据表中的一行)视为对象，它将对象划分为簇，使一 个簇中的对象相互“相似”，而与其他簇中的对象“相异”。在数据规约中，用数据的簇替换实际数据。该技术的有效性依赖于簇的定义是否符合数据的分布性质 R 中常用的聚类函数有 hdust()、kmeans(),前者在使用系统聚类法时使用，后者为快速聚类的函数。

统计分类
统计分类是机器学习非常重要的一个组成部分，它的目标是根据已知样本的某些特征，判断一个新的样本属于哪种已知的样本类[1]。分类是监督学习的一个实例，根据已知训练集提供的样本，通过计算选择特征参数，创建判别函数以对样本进行的分类。与之相对的是无监督学习，例如聚类分析。

3.3.2【热度挖掘评价】

热点话题是一段时间内，围绕某一事件的相关新闻报道、微博信息被大量用户讨论和分享，造成该事件被广泛关注，最终形成全网范围内的话题焦点。热点话题检测是舆情监控及引导工作中的重要任务之一，它通过对海量的实时数据进行及时有效的处理，挖掘文本数据中的话题结构，展示当前互联网中用户关注的话题焦点及其相关内容，为舆情监控者及普通用户掌握当前的热点话题发展趋势提供便捷准确的参考。

但由于网络中的消息冗余繁杂，仅仅依靠人工查找新闻话题难以应对网络中海量信息的处理并对其中的敏感主题及时做出反应。尤其对于决策者，要监控网络中所有相关的信息是不现实的，如果没有自动化的工具支持，很难及时的做出正确的决断，所以人们希望可以通过计算机来自动获取热门新闻话题，从而提高网络监管能力及处置网络舆情突发事件的能力。更为重要的是，在一些安全机构针对网络犯罪的检测和预防过程中，能快速准确地检测出相关话题并及时应对就显得尤为重要。

1.LDA

在机器学习领域，LDA 是两个常用模型的简称：Linear Discriminant Analysis 和 Latent Dirichlet Allocation。本文的 LDA 仅指代 Latent Dirichlet Allocation. LDA 在主题模型中占有非常重要的地位，常用来文本分类。

LDA 由 Blei, David M.、Ng, Andrew Y.、Jordan 于 2003 年提出，用来推测文档的主题分布。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后，便可以根据主题分布进行主题聚类或文本分类。

LDA 涉及到许多先验知识有：二项分布、Gamma 函数、Beta 分布、多项分布、Dirichlet 分布、马尔科夫链、MCMC、Gibs Sampling、EM 算法等。LDA 采用词袋模型。所谓词袋模型，是将一篇文档，我们仅考虑一个词汇是否出现，而不考虑其出现的顺序。

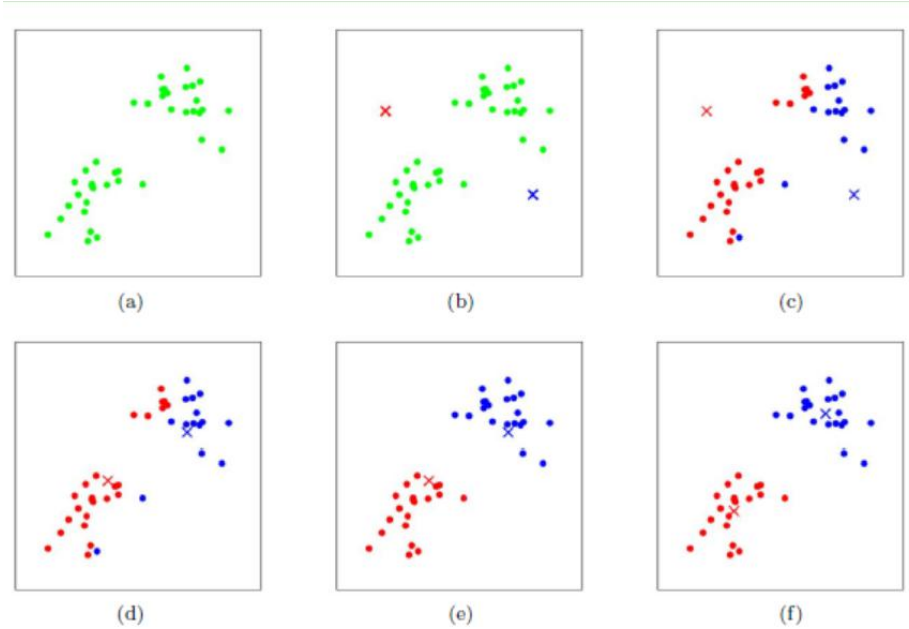
它可以将文档集 中每篇文档的主题以概率分布的形式给出，从而通过分析一些文档抽取出它们的主题（分布）出来后，便可以根据主题（分布）进行主题聚类或文本分类。同时，它是一种典型

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。此外，一篇文档可以包含多个主题，文档中每一个词都由其中的一个主题生成。

隐含狄利克雷分布主题模型（以下简称 LDA）是一种典型的词袋模型，即其先不考虑新闻文档的组成结果，而是先考虑如何通过词袋来生成一篇文章；首先它认为一篇文章是由一组词语构成的集合，词语与词语之间没有顺序及先后的关系，因此主题模型更加简单；一篇文章可以包含很多个主题，而文档中的每一个词都由其中的一个主题来生成。

3.K-means

聚类算法有很多种，K-Means 是聚类算法中的最常用的一种，算法最大的特点是简单，好理解，运算速度快，但是只能应用于连续型的数据，并且一定要在聚类前需要手工指定要分成几类。K-Means 算法的思想很简单，对于给定的样本集，按照样本之间的距离大小，将样本集划分为 K 个簇。让簇内的点尽量紧密的连在一起，而让簇间的距离尽量的大。K-Means 采用的启发式方式很简单，用下面一组图就可以形象的描述



上图 a 表达了初始的数据集，假设 k=2。在图 b 中，随机选择了两个 k 类所对应的类别质心，即图中的红色质心和蓝色质心，然后分别求样本中所有点到这两个质心的距离，并标记每个样本的类别为和该样本距离最小的质心的类别，如图 c 所示，经过计算样本和红色质心和蓝色质心的距离，得到了所有样本点的第一轮迭代后的类别。此时对当前标记为红色和蓝色的点分别求其新的质心，如图 4 所示，新的红色质心和蓝色质心的位置已经发生了变动。图 e 和图 f 重复了在图 c 和图 d 的过程，即将所有点的类别标记为距离最近的质心的类别并求新的质心。最终得到的两个类别如图 f。

当然在实际 K-Means 算法中，我们一般会多次运行图 c 和图 d，才能达到最终的比较优的类别。

4.single-pass

较常使用的是 K-means 聚类算法，需要事先设定聚类数，对于数据量较小的模型来看比较好选定设置簇数数据，但对于比赛中给定的全部数据集，其数据量较大难以确定簇数给定一个合理的阈值范围。针对这种情况不确定到底能聚出来多少主题，需要一种聚类算法不需要预设聚类数。

Single-pass clustering，中文名一般译作“单遍聚类”，由 R Baeza-Yates 等学者[3]于 1999 年提出，它是一种简洁且高效的文本聚类算法。在文本主题聚类中，Single-pass 聚类算法比 K-means 来的更

为有效。Single-pass 聚类算法不需要指定类目数量，可以通过设定相似度阈值来限定聚类数量。主要用于新主题发现、文章聚类等领域。其核心算法思想是流式数据聚类；给定按照时间次序到来的新闻数据流，该算法按顺序处理一次新输入的流式数据；根据当前数据集与已生成的聚类簇进行比较；如果按照一定规则与阈值寻找到了确定的近似簇，则将新数据归入此类簇中；如果没有合适的归入类簇，则将这个新数据作为一个新类；如此反复执行，知道完成所有的数据聚类任务。整个过程中，只对数据进行一次读取操作，所以称为单次（Single）。下图 X 给出了 Single-Pass 聚类算法对流数据处理的过程。

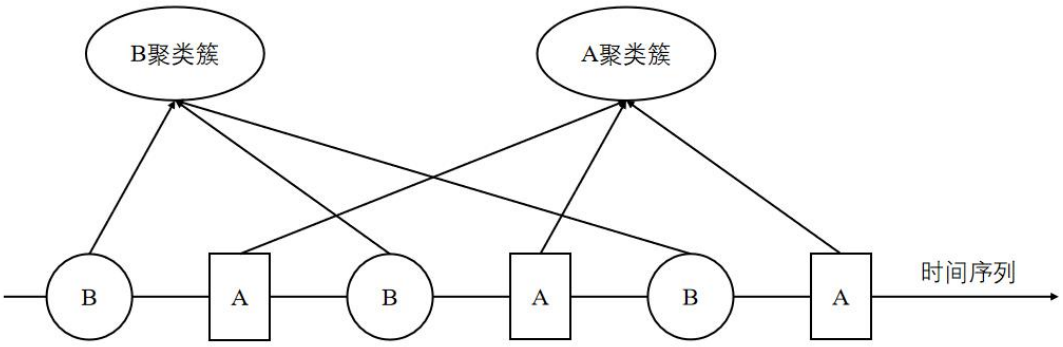


图 X.Single-pass 流程

Single-pass 聚类算法同时是一种增量聚类算法（Incremental Clustering Algorithm），每个文档只需要流过算法一次，所以被称为 single-pass，效率远高于 K-means 或 KNN 等算法。它可以很好的应用于话题监测与追踪、在线事件监测等社交媒体大数据领域，特别适合流式数据（Streaming Data），比如微博的帖子信息，因此适合对实时性要求较高的文本聚类场景。对于本题中，群众所反映的问题覆盖面广，涉及范围大的情况，使用该算法更有助于得到更好的结果。

Single-pass 算法顺序处理文本，以第一篇文档为种子，建立一个新主题。之后再进行新进入文档与已有主题的相似度，将该文档加入到与它相似度最大的且大于一定阈值的主题中。如果与所有已有话题相似度都小于阈值，则以该文档为聚类种子，建立新的主题类别。其算法流程如下：

（1）以第一篇文档为种子，建立一个主题；

（2）基于词袋模型将文档 X 向量化；

（3）将文档 X 与已有的所有话题均做相似度计算，可采用欧氏距离、余弦距离，或者上篇文章中提到的 kullback_leibler, jaccard, hellinger；

（4）找出与文档 X 具有最大相似度的已有主题；

（5）若相似度值大于阈值 theta，则把文档 X 加入到有最大相似度的主题中，跳转至 7；

（6）若相似度值小于阈值 theta， 则文档 X 不属于任一已有主题， 需创建新的主题类别，同时将当前文本归属到新创建的主题类别中；

（7）聚类结束。

点话题是一段时间内，围绕某一事件的关注人群较多、涉及领域与群众生活关系度是否密切、支持呼声较高、时间集中度高与事件急需解决处理息息相关。热点话题检测是舆情监控及引导工作中的重要任务之一，它通过对海量的实时数据进行及时有效的处理，挖掘文本数据中的话题结构，展示当前群众生活最为关心关注的话题焦点及其相关内容，为政府部门及相关决策者掌握当前的热点话题发展趋势提供便捷准确的参考，也让群众的生活愈为便利便捷。如何判断什么是热点问题需要从多个角度全方面考察衡量，本论文中从留言数量、时间集中度、支持率、结合两会群众关心热词反映出的群众关心度进行合理评判，构建一个合理的热度评价指标。

本论文中结合所给数据表内信息以及参考网络舆情评价体系[4]微博热度评价体系[5]，并调研选取了近十五年来两会热词给予权重分配，结合以上内容选取四个方面的指数用与构建热度评价指数：

1.留言数：留言数量，即为同一个问题反映次数，通过留言内容判断类别，对于归纳于相同类别进行统计，这里需运用到第一题的分类结果，在分类后的每一类留言问题中进行文本聚类操作，得到反映同一主题事件的留言数量。

2.时间集中度：对于相同时间段中反应次数，其数量越集中相应重要程度越高。针对于每一主题的留言问题都进行时间范围前后两个月时间的计算，再求取其和得到一个相对集中的事件反映情况，因此得出该事件的时间集中程度。

3.点赞反对数：对于每一主题的留言都进行聚类，对数据集进行去重处理，通过去除标点符号的方式，分别对每一条留言中的每一个词进行对比，对于两个或多个文本中的相似度高达 98%以上的留言只留下其中一条，以减少重复留言对于热度评价的影响，但是重复的留言信息也可能被其他群众用户看到并对其进行点赞或点反对意见，因此重复留言的点赞反对信息依旧需要保留并予以计算。

4.群众关心度： 本论文中通过对于留言内容运用第一题的算法进行分类得到该留言所属类别，通过调研 2006-2019 年两会群众关心热词统计得到，诸如房价、教育、养老就业等问题一直是人民群众最为关心的事情，结合近期 2020 年两会热议词对于上述得到的分类类别赋予不同的权值，以得到更为精准的群众关心度让与群众最关心的问题得到更快更好的解决。

通过下述方法我们可以，XXX

在本题中，通过尝试 LDA 算法、K-means 两种方法发现其效果在数据量较小的数据上给定簇数时表现较好，但对于数据量过大无法自主判断簇数阈值时其结果较差难以应用，其结果对比如下图 X 所示，因此换为不用确定簇数的 single-pass 方法并对于文本处理采取三种方法。分别对于三种文本处理方法结合 single-pass 算法进行介绍：

$$\begin{cases} \text{原文本} + \text{single pass} \\ LDA\text{主题词} + \text{single pass} \\ \text{带标签文本} + \text{single pass} \end{cases}$$

尝试的方法一为，使用原文本和改进 Single-Pass 聚类算法的热点话题评价方法。其具体步骤如下：

1.对于原文本进行去重操作。

2.因为涉及到时间数据的计算，因此需要统一数据格式，统一转化为国务院下发的《政府网站发展指引》中规定的 YYYY—MM—DD HH：MM：SS 格式以方便计算，都转化为时间戳的形式用于后续计算。

3.通过设计 python 程序，进行提取纯语料数据。

4.再利用 single-pass 方法进行文本聚类。

5.对于聚类结果选取前 20 类，使用上述四个角度的评价指标求出排名前 5 的热点问题。

其流程示意如下图 X 所示：

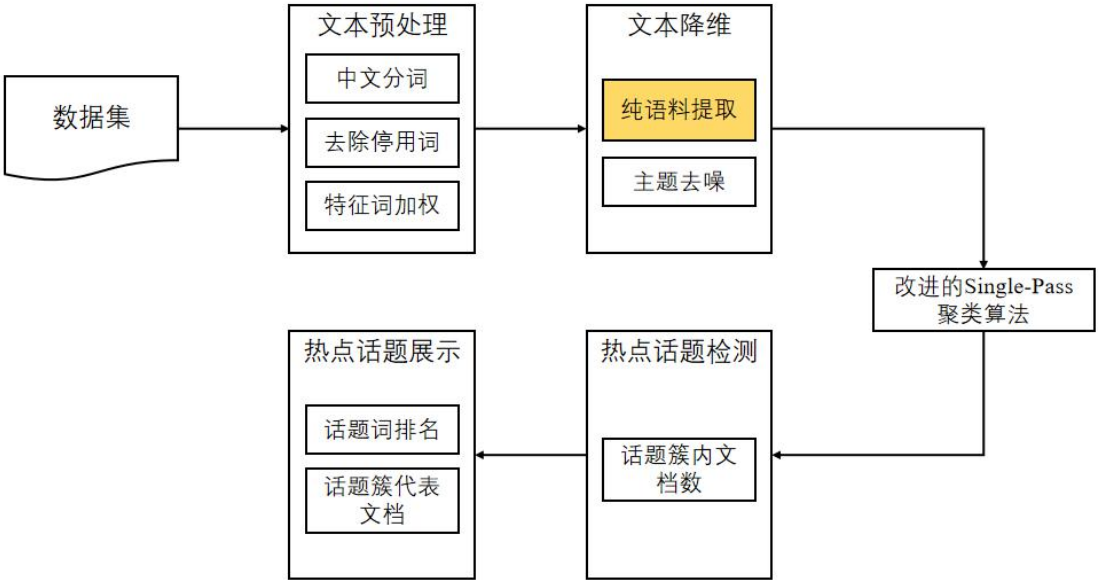


图 X.原文本结合 single-pass 方法求出热点问题

尝试的方法二为，基于特征词加权的隐含狄利克雷分布(Latent Dirichlet Allocation，LDA)主题模型和改进 Single-Pass 聚类算法的热点话题评价方法。

针对隐含狄利克雷分布主题模型聚类算法与 Single-Pass 聚类算法有着各自的优势与不足，在推广到实际的网络新闻系统中时，都不适合作为单独的数据挖掘工具。为了设计一种适合于网络新闻系统的新闻聚类算法，应当明确本题中对于不同留言内容涉及到的不同领域需要政府部门给予合理答复完善的大体需求。

1.对于原文本进行去重操作。

2.因为涉及到时间数据的计算，因此需要统一数据格式，统一转化为国务院下发的《政府网站发展指引》中规定的 YYYY—MM—DD HH：MM：SS 格式以方便计算，都转化为时间戳的形式用于后续计算。

3.通过 LDA 算法，进行提取主题词。

4.再利用 single-pass 方法进行文本聚类。

5.对于聚类结果选取前 20 类，使用上述四个角度的评价指标求出排名前 5 的热点问题。

其流程示意如下图 X 所示：

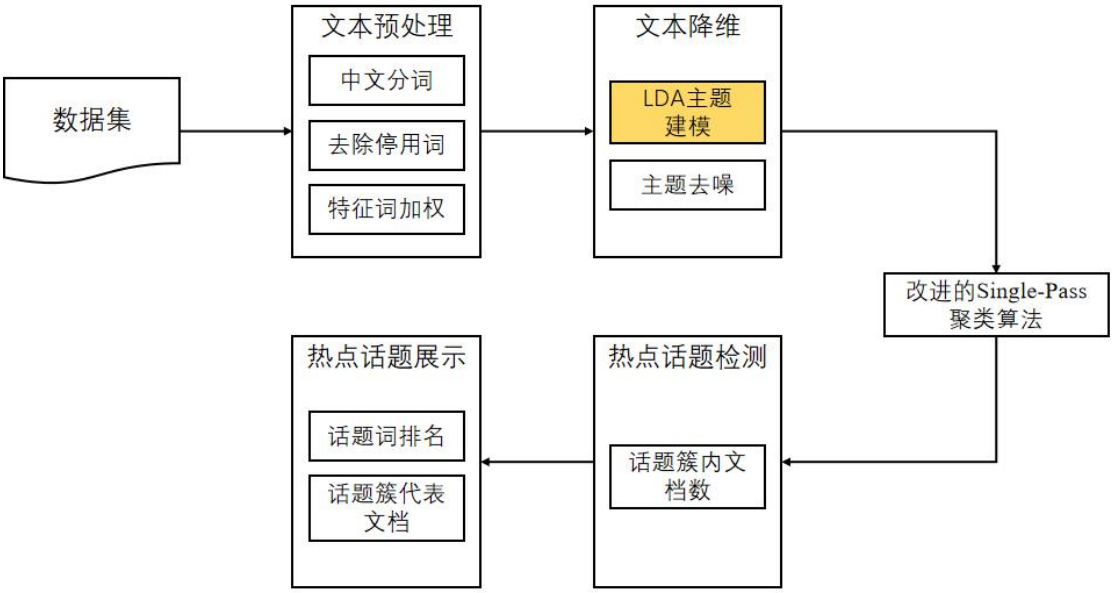


图 X.LDA 主题词结合 single-pass 方法求出热点问题

尝试的方法三为，运用第一题给留言内容打上标签，以处理后的文本结合改进 Single-Pass 聚类算法的热点话题评价方法。

1.对于原文本进行去重操作。

2.因为涉及到时间数据的计算，因此需要统一数据格式，统一转化为国务院下发的《政府网站发展指引》中规定的 YYYY—MM—DD HH：MM：SS 格式以方便计算，都转化为时间戳的形式用于后续计算。

3.通过第一题分类方法，对于留言内容进行打上所属类标签，提取纯语料数据。

4.再利用 single-pass 方法进行文本聚类。

5.对于聚类结果选取前 20 类，使用上述四个角度的评价指标求出排名前 5 的热点问题。

其流程示意如下图 X 所示：

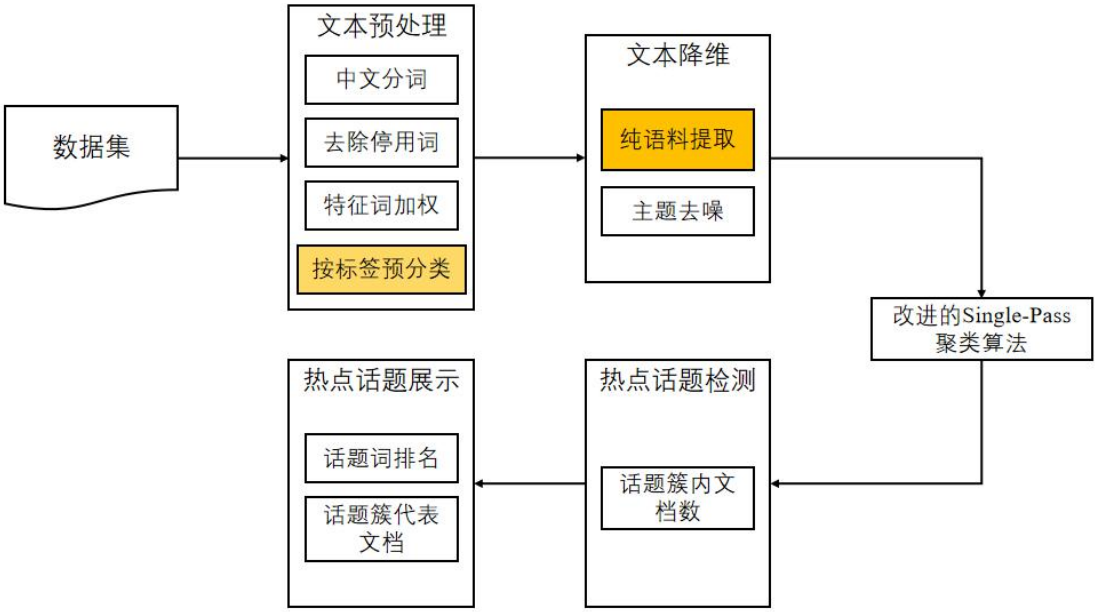


图 X.带标签文本结合 single-pass 方法求出热点问题

3.3.3 第三题——答复意见的评价

本文中评估评论的质量主要是量化评论对用户的价值程度。根据人工分析评论质量的隐含因素和以往的研究[6,7,8,9]，可以从以下五个方面来评估评论的质量。

1、完整性

完整性（Completeness）衡量的是评论覆盖回复留言内容各个方面的程度，评论评价答复的方面越多，该答复解答就越全面。

首先进行分词、标注词性标签操作，随后直接提取其中的/n 和/vn 两种词性的词作为特征词。然后按类型统计其中的高频特征词汇，例如 500 个、1000 个，统计每条留言当中的高频词汇数量作为相关性分数。

$$S\_comp(i)=\frac{TextCount(i)}{LabelCount(L_i)}$$

其中，TextCount(i)指的是第 i 条数据的文本的特征词数量，L_i 指的是第 i 条数据的文本分类类型，LabelCount(i)指的是第 i 条数据的特征词数量。

参考《福建省公安厅关于 2019 年度福建公安公众服务网与政务新媒体建设工作情况的函》其中给出了关于省级审批服务事项和便民服务事项按照要素。以及在当今依法治国的背景下，要求答复内容须包含有关法律、法规、政策、文件依据，因此无论是解答疑惑，针对留言内容进行答复都完整性代表留言回复必须包含指定的要素，根据上述信息得出完整性是一个非常重要的指标。我们具体的设计思想如下：

1.先对于答复意见详情采用第一题的思路进行分类，针对于每一个类别进行政务特征词提取，每一类别的政务特征词均为 500 个，能较好覆盖到该类问题中的方方面面。

2.针对于每一条留言进行特征词提取，再利用留言特征词与该领域类整体的政务特征词进行比对，当其包含的数量越多时，认为该留言的完整性越强。

2、相关性

相关性（Relativity）是指针对于留言答复的内容与所涉及的留言问题的相关程度。留言答复所用的特征词是用来评价答复质量的参考因素之一，这些特征词与答复有一定的关联，这个关联程度即为答复意见的相关性。相关性越好的评论，提供的有用信息越多，对留言用户以及关注此问题的用户越有用。

WMD 算法是 2015 年提出的，是基于 word2vec 基础上通过计算文本间词的距离来衡量文本相似度的算法，是基于 word embeddings 计算距离的算法。文本间的距离一般被用于信息检索，新闻分类等等。原有的计算方法中，文本的表示形式一般分为两种，一种是 BOW(bag of words)，另一种是 TF-IDF(term frequency-inverse document frequency)。这两种表示方法都存在缺点，影响最大的方面为，不能捕捉各个单词间的距离，WMD 算法可以较好的解决以上问题。结合本题的 WMD 公如下所示：

$$\mathrm{S\_rela(i)} = \mathrm{WMD(Text\_1(i), Text\_2(i))}$$

其中，$Text\_1(i)$ 指的是第 i 条数据的“留言详情”内容，$Text\_2(i)$ 指的是第 i 条数据的“答复意见”内容，$WMD(a,b)$ 指的是使用 WMD 算法计算文本 a，文本 b 的文本相似度。

评价答复意见质量的重要因素之一是相关性，只有对于针对问题回答问题，才能认为该答复意见对于群众反映的事件有切实帮助，当相关性越高时，可以认为与该留言密切愈为相关，针对反映的问题也更好的解决与帮助。

3、可读性

可读性（Readability）是指评论适合用户阅读的程度，是评论所具有的阅读价值。一条可读性比较好的评论可以直接简洁的答复内容，陈述观点，回复中的用词错误比较少，而且不会多次重复无意义用词。可通过评论的长度或平均长度、拼写错误的词语数目等来度量。也可以通过各种可读性指标来衡量，如 Gunning fog index[10]、SMOG（Simple Measure of Gobbledygook）[11]、Flesch – Kincaid readability tests [12]等。 本题设计的可读性公式如下所示：

$$\mathrm{S\_read(i)} = \mathrm{EffectiveLength(Text\_2(i))}$$

其中，$Text\_2(i)$ 指的是第 i 条数据的“答复意见”内容，$\mathrm{EffectiveLength(a)}$ 指的是文本 a 的有效长度。

本论文中的方法是将本文分词之后去除停用词并去重，统计文本有效长度作为可读性分数。

4、实时性

在新闻热度中、网络舆情热度的分析中，实时性都是一个非常重要的影响因素之一，王文妤等人对政府实时监控网络舆情热度的研究中指明，面对监控困境和监控的必要性，强调网民心态的主观性对舆情热度的主观性，及时性是一个重要的影响因素之一。及时的答复才对群众留言及反馈的意见有了更好的指导与帮助，对于政府工作效率的改进有了强有力的规范依据，也提升了群众对于政府部门的信赖。

$$\mathrm{S\_time(i)} = \frac{\mathrm{Time\_label(L_i)}}{\mathrm{Time\_2(i) - Time\_1(i)}}$$

其中，$Time\_1(i)$ 指的是第 i 条数据的“答复时间”转化成的时间戳数据，$Time\_2(i)$ 指的是第 i 条数据的“留言时间”转化成的时间戳数据。$Time\_label(c)$ 指的是 c 类数据当中时间的最高值。

在本题中，从数据集附件 4 中统计数据“留言时间”和“答复时间”之间的时间间隔，精确到秒，转换为时间戳的数据类型，计算两者差值以求得实时性的分数。当“答复时间”与“留言时间”的差值越小时，代表答复越及时，因此该条答复的实时性得分越高。

5、内聚性

内聚性（Cohesion）反映的是评论中出现的特征词之间的语义关系，评估一条评论的内聚度可以用组成评论的特征词之间的语义联系来衡量。若这些特征词联系越紧凑或越切合同一主题，评论的内聚性就越强。根据 Ransdell[13,14]的定义，一个文本片段的语义关系指的是它代表一些相互独立的概念的能力。

本文定义函数 XX 来计算答复意见的内聚性，内聚性只考虑匹配到政务特征主题的词语，如果答复意见中的词匹配不到该类留言的特征主题词，则该答复的内聚性值为 0。本文中内聚性的计算采用 S_cohe 函数[15]，Word2vec 函数被用于衡量语言学本体中两个词语之间的语义间距离，本论文中设计的基本思想是通过计算本体中两个词语之间的距离来度量答复内容之间的内聚性，两个词语之间的距离越小，它们的相似度就越大。如果两个词语的距离为本体的最大深度，那么它们的相似度为 0，如果两个词语直接相连或者两个词语相同时，对其进行数值规约，它们的相似度为 10。本文的内聚性函数如下：

$$S\_cohe(i)=\frac{\sum_{(w_j,w_k\in Words_i)w_j\neq w_k} Length(W_j,W_k)}{(CountWord(Words_i)-1)!}$$

其中，$Words_i$ 指的是第 i 条数据的特征词数组，W_j 指的是 $Words_i$ 中的第 j 个特征词，$Length(a,b)$ 表示单词 a 和单词 b 的欧式距离，$CountWord(c)$ 表示特征词数组的 c 的长度。

本论文中的设计思想为，计算特征词之间的距离，然后求平均值作为内聚性分数。当内聚性分数越高时，认为该条答复意见的内聚性越好其内容越紧凑主题相关度越好。

针对附件 4 相关部门对留言的答复意见，从答复的完整性、相关性、可读性、实时性、内聚性五个角度对答复意见的质量给出一套评价方案，通过对于上述五个角度计算得出的数据，在运用第一题的分类计算后，对同一类别的数值规约计算，减少极端数据对于评价方案造成的影响。

四、实验[实验评估]

在此部分，主要介绍群众留言分类、热点问题挖掘两个问题的实验内容以及实验结果。

4.1 超参数选取

4.1.1 群众留言分类

在本题中，系统采用 xlnet 模型作为分类模型，模型的相关超参数设置如下：

超参数	数值
Train epochs	8
Train batch size	8
Eval batch size	8

Learning rate	5e-6

4.1.2 热点问题挖掘

在本题中，不同的算法需要设置不同的超参数，相关超参数设置如下：

4.2 数据集划分

4.2.1 群众留言分类

在第一题当中，团队采用的分类模型为基于自注意力机制的 xlnet 模型。为了更好的评价模型的性能，团队将泰迪杯组委会提供的“附件 2”数据集按照不同的类别进行划分，并将不同类别的数据按照 6:2:2 的比例划分不同类别训练集、检验集、测试集，最终汇总不同类别的训练集作为最终模型的训练集，校验集相同。

4.2.2 热点问题挖掘

在该题中，团队手工标记了五个热点问题的类别，并以此作为第二问的测试集。

4.3 实验评价指标

团队采用准确率和 F-Score 对模型的分类结果进行评价。其中 F-Score 评价方法如下：

$$F_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{2P_iR_i}{P_i+R_i},$$

其中，Pi 为第 i 类的差准率，Ri 为第 i 类的查全率。

4.4 实验结果及分析

4.4.1 群众留言分类

团队使用 train 数据集训练 xlnet 模型，并使用测试集和校验集评估模型的性能。

Model	f_scores_eval	eval_accuracy	f_scores_test	test_accuracy
Bert	0.884	0.891	0.875	0.882
xlnet	0.886	0.892	0.888	0.893

4.4.2 热点问题挖掘

团队使用测试集对热点问题挖掘模型的性能进行评估，结果如下：

Model	accuracy
K-means	0.944
原文本+single-pass	0.924
LDA 主题词+single-pass	0.821