摘要

主题建模分析文档，学习单词的有意义模式。然而，现有的主题模型在使用大词汇表和重尾词汇表时无法学习可解释的主题。为此，我们开发了嵌入式主题模型(ETM)，这是一种文档的生成模型，它将传统主题模型与单词嵌入结合在一起。特别是，它用分类分布对每个词进行建模，分类分布的自然参数是词嵌入与其指定主题嵌入之间的内积。为了适应ETM，我们开发了一种有效的平摊变分推理算法。ETM发现可解释的主题，即使是包含生僻词和停止词的大词汇表。它在主题质量和预测性能方面都优于现有的文档模型，如潜在Dirichlet分配

介绍

主题模型是发现文档集合中隐藏语义结构的统计工具(Blei et al.， 2003;布莱,2012)。主题模型及其扩展已经应用于许多领域，如市场营销、社会学、政治学和数字人文学科。BoydGraber等人(2017)进行了综述。

大多数主题模型建立在潜在Dirichlet分配(LDA)上(Blei et al.， 2003)。LDA是一个分层概率模型，它将每个主题表示为术语分布，并将每个文档表示为主题的混合。当适合文档集合时，主题将总结它们的内容，主题比例将提供每个内容的低维表示。利用变分推理和随机优化方法，LDA能够适应大型文本数据集。

LDA是一种功能强大的模型，被广泛应用。然而，它面临着一个普遍的技术问题——它在面对大型词汇表时失败了。从业者必须严格地精简他们的词汇表，以适应好的主题模型，也就是说，那些既可预测又可解释的模型。这通常是通过删除最常用和最不常用的单词来实现的。在大型集合上，这种修剪可能会删除重要的术语并限制模型的范围。大词汇量的主题建模问题在研究文献中尚未得到解决。

与主题建模并行的是词嵌入的想法。单词嵌入的研究始于Bengio等人(2003)的神经语言模型，该模型与Blei等人(2003)发表在同一年的期刊上。单词嵌入避免了单词的“one-hot”表示方式——单词长度为0的向量和一个单词长度为1的向量来学习分布式表示方式，即在低维向量空间中具有相似含义的单词接近(Rumelhart和Abrahamson, 1973;本吉奥等人，2006年)。对于主题模型，研究人员将嵌入方法扩展到大型数据集(Mikolov等人，2013a,b;Pennington等人，2014;Levy和Goldberg, 2014年;Mnih和Kavukcuoglu, 2013)。字嵌入(word embedding)已经在许多方面得到了扩展和发展。它们在自然语言处理的许多应用中变得至关重要(Li和Tao, 2018)，它们也被扩展到文本之外的数据集(Rudolph等人，2016)。

在本文中，我们开发了嵌入式主题模型ETM (embedded topic model)，这是一种用于单词嵌入的主题模型。该模型既具有良好的主题模型特性，又具有良好的词嵌入特性。作为主题模型，它发现了文本的可解释的潜在语义结构;作为一种词语嵌入，它提供了词语意义的低维表示。它能够稳健地容纳大型词汇表和语言数据的长尾。

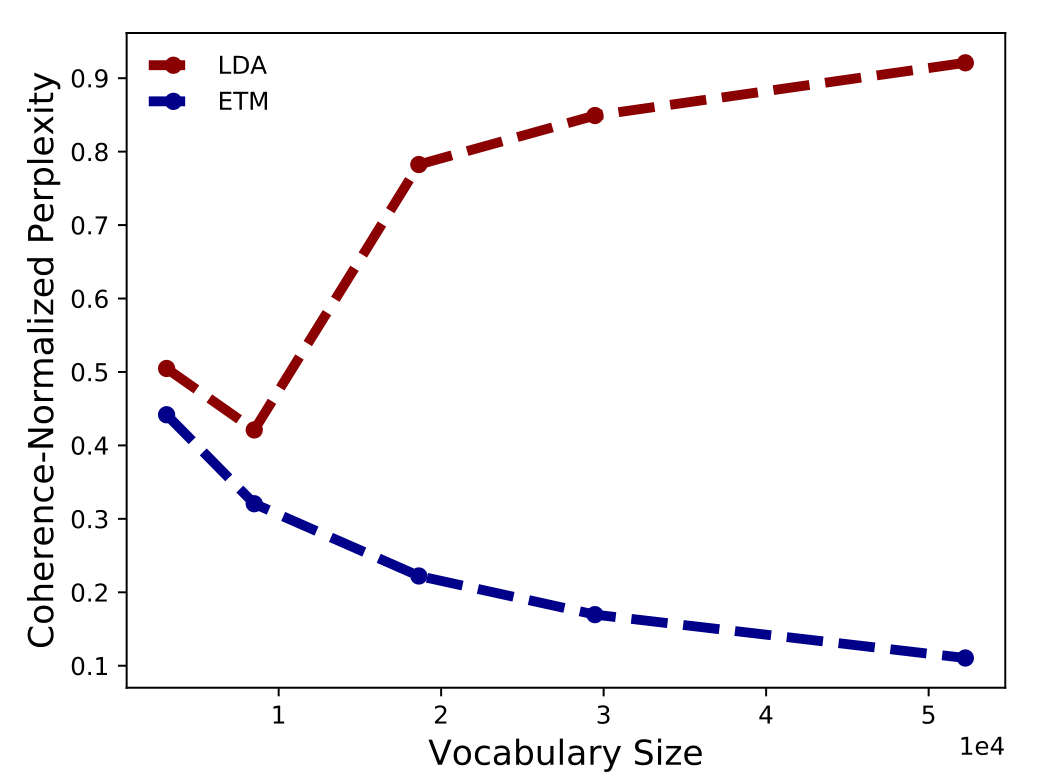


图1所示。在ETM和LDA中，文档完成的标准化保留困惑和主题连贯的比率作为词汇量的函数。当词汇量较大时，LDA的性能会下降，而ETM则保持了良好的性能。

图1说明了这些优点。这张图描绘了被搁置的文档的预测困惑和主题连贯之间的比率，作为词汇量大小的函数。(这个困惑已经被词汇量正常化了。)这是来自20新闻组的12000篇文章和100个主题的语料库。红线是LDA;它的性能会随着词汇量的增加而下降——预测性能和主题的质量会变差。蓝线是ETM;即使词汇量越来越大，它也能保持良好的性能。

像LDA一样，ETM是一个生成概率模型:每个文档都是主题的混合体，每个观察到的单词都被分配到一个特定的主题。与LDA相反，术语的每个主题条件概率具有对数线性形式，涉及词汇表的低维表示。每一项用一个内嵌表示;每个主题都是嵌入空间中的一个点;题目在项上的分布正比于题目的内嵌和每个项的内嵌的指数内积。图2和图3显示了来自《纽约时报》一个有300个主题的ETM的主题。图表显示了每个主题的嵌入和最接近的单词;这些话题是关于基督教和体育的

由于主题是按照嵌入空间中的一个点表示的，ETM对于停止词的存在也很健壮，这与大多数常见的主题模型不同。当停止词包含在词汇表中时，ETM将主题分配到嵌入空间的相应区域(我们在第6节中演示了这一点)。

对于大多数主题模型，主题比例的后验是难以计算的。我们推导了一种有效的算法，用变分推理逼近后验(Jordan等人，1999;Hoffman等人，2013年;Blei et al.， 2017)和另外使用平摊推理有效地逼近主题比例(Kingma和Welling, 2014;Rezende等，2014)。得到的算法使ETM适用于词汇量大的大型语料库。ETM算法既可以使用之前拟合的词嵌入，也可以将它们与其他参数联合拟合。(特别地，图1至图3是通过预先拟合的跳过图词嵌入ETM版本得到的。)

我们比较了ETM、LDA和神经变分文档模型(NVDM)(一种多项式矩阵分解形式)的性能。ETM提供了良好的预测性能，这是通过文档完成任务的延迟日志可能性来衡量的(Wallach等人，2009)。它还提供了有意义的主题，这是通过主题一致性(Mimno等人，2011)和主题多样性来衡量的，这也是一个指标，也表明了主题的质量。ETM对于大型词汇表尤其健壮。

相关工作

本工作开发了一个扩展LDA的新主题模型。LDA已经以许多方式进行了扩展，主题建模已经成为它自己的一个子领域。关于综述，见Blei(2012)和Boyd-Graber等人(2017)。

开发ETM的目标之一是将单词相似度纳入主题模型，以前也有类似的研究。这些方法要么修改主题先验，要么修改主题分配先验(Xie et al.， 2015)。例如，彼得森等人(2010)使用单词相似度图(由词典给出)使LDA倾向于将相似的单词分配到相似的主题。另一个例子是，Xie等人(2015)利用马尔可夫随机场对LDA的每个单词的主题分配进行建模，以考虑相似单词的主题比例和主题分配。这些方法使用单词相似性作为一种关于语言的“侧信息”;相比之下，ETM直接在词的生成过程中(通过嵌入)模拟相似性。

其他工作已经将LDA扩展到直接涉及单词嵌入。一种常见的策略是将离散的文本转换为连续的嵌入观察，然后调整LDA生成实值数据(Das et al.， 2015;Xun et al.， 2016;Batmanghelich et al.， 2016;熏等，2017)。使用这种策略，主题是具有潜在均值和协方差的高斯分布，嵌入的可能性采用高斯(Das et al.， 2015)或V - mises Fisher分布(Batmanghelich et al.， 2016)建模。ETM与这些方法的不同之处在于，它是一个分类数据的模型，一个经过嵌入矩阵的模型。因此，它不需要预先嵌入，实际上，它可以学习嵌入作为推理过程的一部分。

还有其他一些结合LDA和嵌入的方法。Nguyen等人(2015)将LDA定义的可能性与使用预拟合词嵌入的对数线性模型混合;Bunk和Krestel(2018)用高斯分布的嵌入词随机替换从一个主题中抽取的单词;和Xu等(2018)采用几何视角，联合使用Wasserstein距离学习主题和单词嵌入。

最近的另一项研究通过深度神经网络改进了主题建模推理(Srivastava和Sutton, 2017;卡德等人，2017;Cong等人，2017;Zhang et al.， 2018)。具体来说，这些方法通过平摊推理和变分自动编码器来降低文本数据的维数(Kingma和Welling, 2014;Rezende等，2014)。为了在ETM中进行推理，我们还利用了平摊推理方法(Gershman和Goodman, 2014)。

最后，作为一个文档模型，ETM还涉及到作为嵌入模型的一部分学习每个文档表示的作品(Le和Mikolov, 2014;喜怒无常,2016;苗等，2016)。与这些工作相反，ETM中的文档变量是一个更大的概率主题模型的一部分。

背景

ETM建立在两个主要思想之上，LDA和单词嵌入。考虑一个文档的语料库，其中词汇表包含个不同的术语。设表示第个文档中的第个单词。

**潜在狄利克雷分配** LDA是文档的概率生成模型(Blei et al.， 2003)。它假设个主题，每个主题都是词汇表上的一个分布。LDA假设每个文档都来自混合主题，其中主题在整个语料库中共享，混合比例对于每个文档都是唯一的。每个文档的生成过程如下:

1. 绘制主题比例。

2. 对于文档中的每个单词n:

(a)绘制主题分配。

(b)画单词。

其中为类别分布。LDA将Dirichlet置于主题之上， Dirichlet分布的浓度参数和是固定模型超参数。

**Word Embedding** 单词嵌入提供了使用单词向量表示的语言模型(Rumelhart和Abrahamson, 1973;本吉奥等人，2003年)。词语表征与意义相关联，意思相似的词语会有相近的表征。(在嵌入式应用中，一个词的“意义”来自它所使用的上下文。)

我们关注单词嵌入的连续词袋(CBOW)变体(Mikolov等人，2013b)。在CBOW中，每个词的似然为

嵌入矩阵是一个矩阵，它的列包含词汇表的嵌入表示，。向量是上下文嵌入。上下文嵌入是周围的词的上下文嵌入向量(每个词对应)的和。

4嵌入式主题模型

ETM是一个主题模型，它使用单词和主题的嵌入表示。它包含潜在维度的两个概念。首先，它将词汇表嵌入到l维空间中。这些嵌入在精神上类似于经典的单词嵌入。其次，它用K个潜在主题表示每个文档。

在传统的主题建模中，每个主题都是词汇表上的一个完整分布。而在ETM中，第主题是向量在嵌入空间中。我们称为主题嵌入，它是单词语义空间中第主题的分布式表示。

在生成过程中，ETM使用主题嵌入在词汇表上形成每个主题的分布。具体来说，ETM使用loglinear模型，取词嵌入矩阵和主题嵌入的内积。使用这种形式，ETM通过测量单词的嵌入和主题的嵌入之间的一致性，将高概率赋给主题中的单词。

用表示字嵌入矩阵;列是的嵌入。在ETM下，第个文档的生成过程如下:

1. 绘制主题比例.
2. 对于文档中的每个单词,
   1. 绘制主题分配。
   2. 画出单词。

第1步，表示logistic正态分布(Aitchison and Shen, 1980;布莱和拉费蒂，2007);它将一个标准的高斯随机变量转换为单纯形。从这个分布得到一个:

(我们用逻辑正态线代替了狄利克雷，从而更容易在推理算法中使用重参数化;见第五节。)

步骤1和2a是主题建模的标准步骤:它们将文档表示为主题的分布，并为每个观察到的单词绘制主题分配。步骤2b是不同的;它利用词汇表的嵌入和指定的主题嵌入从指定的主题中提取观察到的单词，如所给出的。

Step 2b中的话题分布反映了Eq. 1中的CBOW似然。回想一下，CBOW使用周围的单词构成上下文向量}。相反，ETM使用主题嵌入作为上下文向量，其中指定的主题是从每个文档变量中提取的。ETM从文档上下文中提取单词，而不是从周围单词的窗口提取单词。

ETM似然使用单词嵌入矩阵，表示词汇在低维空间中的表示。在实践中，它可以依赖于先前安装的嵌入或学习它们作为其整体安装程序的一部分。当ETM学习嵌入作为拟合过程的一部分，它同时找到主题和嵌入空间。

当ETM使用之前的拟合嵌入时，它在特定的嵌入空间中学习语料库的主题。当嵌入词中有语料库中未使用的词时，这种策略特别有用。ETM可以假设这些单词是如何适应主题的，因为它可以计算，即使单词没有出现在语料库中。

5 推理与估计

给我们一个文档的语料库，其中是单词的集合。我们如何适应ETM?

**边际似然** ETM的参数为嵌入矩阵和主题嵌入矩阵;每个是嵌入空间中的一个点。我们最大化文件的边际可能性，

问题是每个文档的边际似然是难以计算的。它涉及到对主题比例的一个困难的积分，我们把它写成公式2中未变换的比例的形式，

每个单词的条件分布使主题分配边缘化，

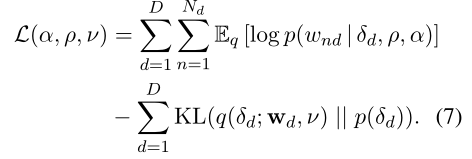
其中，表示(变换后的)话题比例(Eq. 2)， 表示传统的“话题”，即由单词嵌入和话题嵌入引起的单词分布。

公式4至6充实了公式3中的可能性。

**变分推理** 我们用变分推理回避了棘手的积分(Jordan等人，1999;Blei等人，2017)。V变分推理在公式4的边际似然对数上优化每个文档边界的和。有两组参数需要优化:模型参数，如上所述，和变分参数，它收紧边际似然的边界。

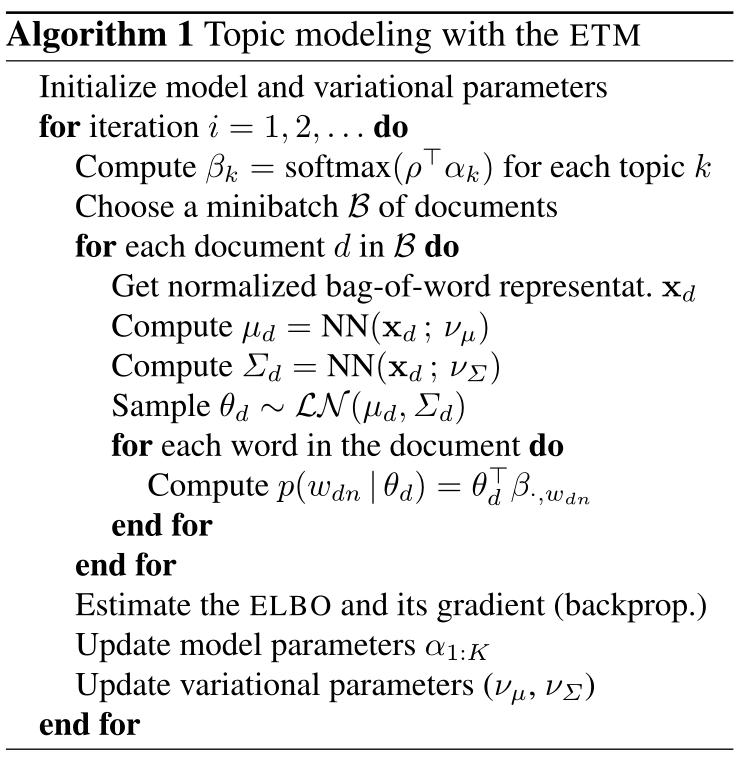
首先，假定一个未变换主题比例。我们使用平摊推理，其中的变分分布既依赖于文档，又依赖于共享的变分参数。特别是是一个高斯分布，其均值和方差来自一个“推理网络”，一个由参数化的神经网络(Kingma和Welling, 2014)。推理网络摄取文档并输出的平均值和方差。(为了适应不同长度的文档，我们通过单词的数量对文档的词袋表示进行规范化，从而形成推理网络的输入。)

我们用这一组变分分布来限制对数边际似然。证据下限(ELBO)是模型参数和变分参数的函数，



作为变分参数的函数，第一项鼓励他们将质量放在解释所观察词语的主题比例δ\_d上;第二项鼓励它们接近前一个。作为模型参数的函数，该目标最大化期望的完全对数似然，

我们根据模型参数和变分参数对ELBO进行了优化。我们使用随机优化，通过重参数化技巧对全梯度进行蒙特卡罗近似，形成噪声梯度(Kingma和Welling, 2014;Titsias和Lázaro-Gredilla, 2014;Rezende等，2014)。我们还使用数据子采样来处理大量的文档集合(Hoffman等人，2013年)。我们和Adam一起设定了学习速率(Kingma and Ba, 2015)。其过程如算法1所示，其中表示输入和参数的神经网络。



6 实证研究

我们研究了ETM的性能，并将其与其他无监督文档模型进行了比较。一个好的文档模型应该提供连贯的语言模式和准确的单词分布，因此我们在预测准确性和主题可解释性方面衡量性能。我们用对数似然来衡量文档完成任务的准确性(RosenZvi等人，2004;Wallach et al.， 2009);我们将主题的可解释性作为主题一致性和多样性的混合来衡量。我们发现，在可解释的模型中，ETM提供了更好的预测和主题。

在单独的分析中(第6.1节)，我们研究了每种方法在存在停止词时的鲁棒性。标准的主题模型在这种机制中失败了——因为停止词出现在许多文档中，每个学习过的主题都包含一些停止词，导致主题的可解释性很差。相比之下，ETM能够利用单词嵌入的信息提供可解释的主题

**语料库** 我们研究了20新闻组语料库和纽约时报语料库。

20Newsgroup语料库是新闻组帖子的集合。对语料库进行预处理，过滤停止词、文档频率在70%以上的词以及分词。为了形成词汇表，我们保留出现在一定数量以上文档中的所有单词，并且我们将阈值从100(较小的词汇表，其中V = 3102)更改为2(较大的词汇表，其中V = 52,258)。在预处理之后，我们进一步从验证和测试集中删除一个单词的文档。我们将语料库分为11,260个文档的训练集、7,532个文档的测试集和100个文档的验证集。

《纽约时报》语料库是一个更大的新闻文章集合。它包含超过180万篇文章，跨越1987年至2007年。我们遵循与20Newsgroups相同的预处理步骤。我们用V = 5921到V = 212,237的词汇构成了这个语料库的版本。在预处理之后，我们使用85%的文档进行培训，10%用于测试，5%用于验证。

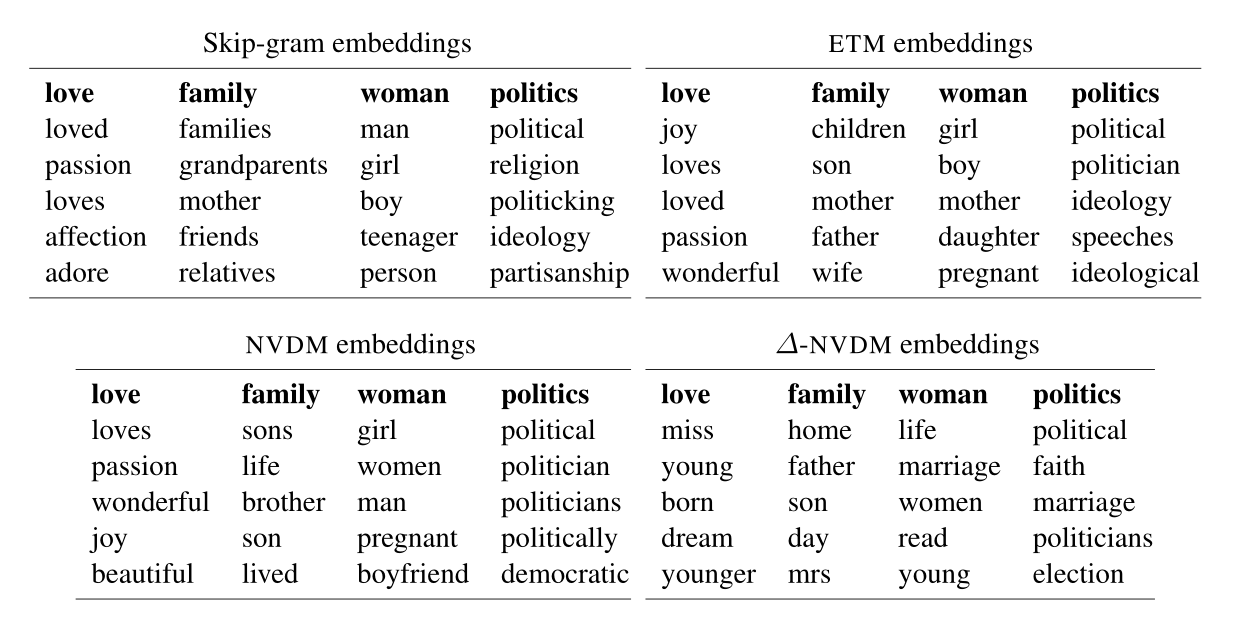
**模型** 我们比较了ETM与其他两种文档模型的性能:潜在Dirichlet分配(LDA)和神经变分文档模型(NVDM)。

LDA (Blei et al.， 2003)是一个标准的话题模型，它为话题和话题比例假定了Dirichlet先验。(我们将前面的超参数设置为1。)它是一个条件共轭模型，具有坐标上升的变分推理。我们之所以考虑LDA，是因为它是最常用的主题模型，并且与ETM有着相似的生成过程.

NVDM (Miao et al.， 2016)是文档的多项因子模型;它假设似然，其中维向量为每个文档变量，为大小为的实值矩阵。NVDM使用每个文档的实值潜向量对logit空间中的嵌入矩阵进行平均。与ETM类似，NVDM使用平摊变分推理联合学习文档表示和模型参数的近似后验。

NVDM不能作为主题模型来解释;它的潜变量是不受约束的。我们研究了NVDM的一个更可解释的变体，它约束处于单纯形中，用logistic正态代替其高斯先验(Aitchison和Shen, 1980)。(这可以被认为是半非负矩阵分解。)我们称这个文档模型为。

表1《纽约时报》上所有文档模型(和跳过图)都学会了单词嵌入，词汇量为118,363。



我们研究了ETM的两种变体，一种是预先拟合单词嵌入，另一种是它们与其他参数一起学习。带有预装嵌入物的变体称为“标记ETM”。我们使用跳跃图嵌入(Mikolov等人，2013b)。

**算法设置** 给定一个语料库，每个模型都有一个近似的后验推理问题。我们对所有模型使用变分推理，并使用随机变分推理(SVI) (Hoffman等人，2013)来加速优化。小批处理大小为1000个文档。对于LDA，我们设置了Hoffman et al.(2013)提出的学习速率:延迟为10，遗忘因子为0.85。

在SVI中，LDA具有协调的上升变分更新，有5个内部步骤来优化局部变量。对于其他模型，我们使用局部变量的平摊推理。我们使用三层推理网络，并设置局部学习率为0.002。我们对变分参数(权值衰减参数为)使用正则化。

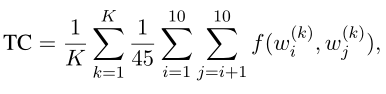
**定性的结果** 我们首先检查嵌入物。ETM、NVDM和都涉及到单词嵌入。我们通过固定一组术语并计算它们周围出现的单词来说明它们。为了进行比较，我们还说明了通过跳跃图模型学习的单词嵌入。

表1说明了不同模型的嵌入情况。所有的方法都提供了可解释的嵌入，即语义相近的词。ETM和NVDM学习的嵌入与跳过图中类似。的埋深不同;局部变量的单纯形约束改变了嵌入的性质。

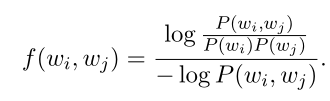
我们接下来看看所学的主题。表2显示了所有方法中最常用的7个主题，由主题比例的平均值给出。LDA和ETM都提供了可解释的主题。NVDM和都没有提供可解释的主题;他们的模型参数不能被解释为在混合形成文档的词汇表上的分布。

**定量的结果** 我们接下来定量地研究这些模型。我们测量主题的质量和模型的预测性能。我们发现，在具有可解释主题的模型中，ETM提供了最好的预测。

我们通过混合两个指标:话题一致性和话题多样性来衡量话题质量。话题连贯是对话题可解释性的定量测量(Mimno et al.， 2011)。它是从同一文档中随机抽取的两个词的点态互信息平均值(Lau et al.， 2014)，



其中表示主题中最可能出现的前10个单词。其中，为标准化点态互信息，

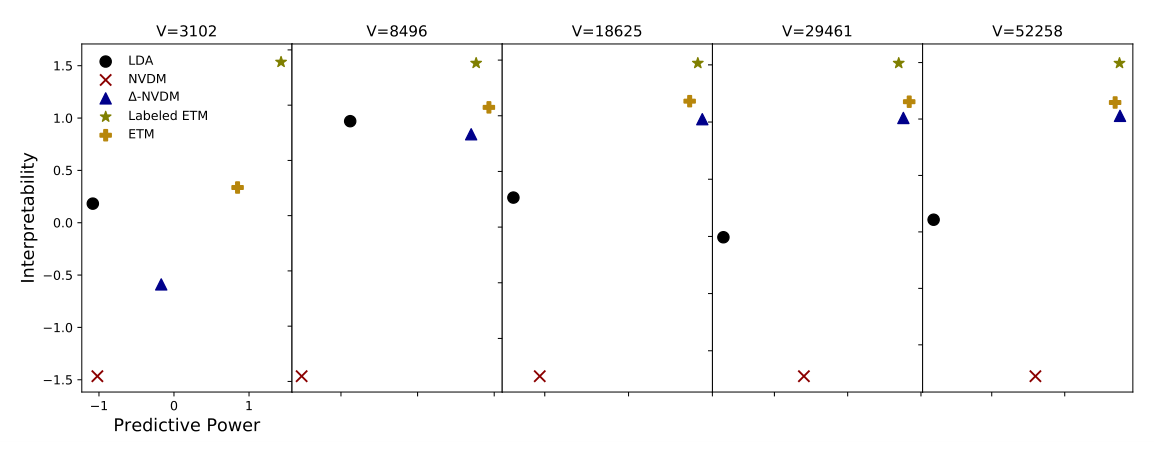


数量是单词和在文档中同时出现的概率，是单词的边缘概率。我们用经验计算来近似这些概率。

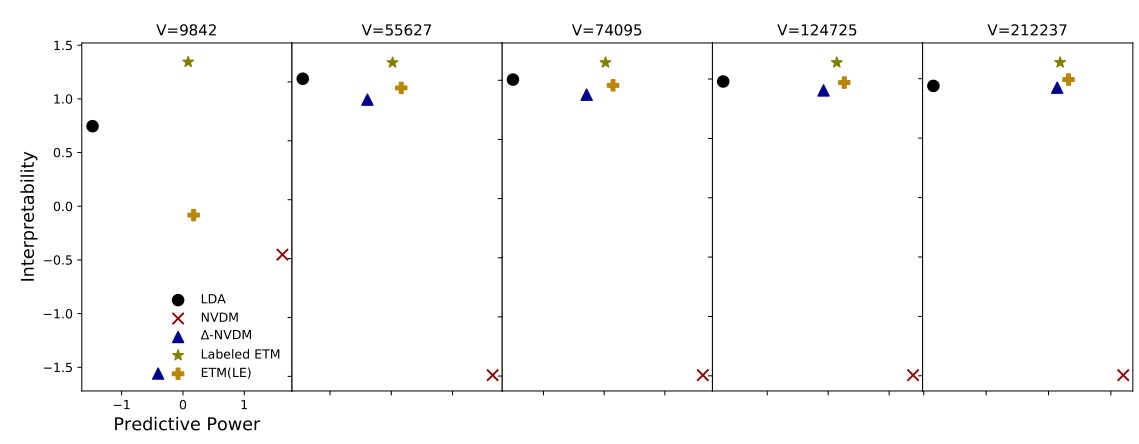
主题连贯背后的思想是，一个连贯的主题将显示往往出现在同一文档中的单词。换句话说，在一个连贯的话题中，最有可能的词应该具有高度的互信息。具有较高主题一致性的文档模型是更易解释的主题模型。

我们将连贯性与第二个指标——主题多样性结合起来。我们将话题多样性定义为所有话题前25个单词中唯一单词的百分比。多样性接近0表示主题冗余;接近1的多样性表示主题更加多样化。我们将模型主题质量的总体度量定义为主题多样性和主题一致性的产物。

好的主题模型还提供了良好的语言分布。为了度量预测质量，我们计算了文档完成任务的日志可能性(rosenzvi等人，2004;Wallach et al.， 2009)。我们将每个测试文档分成两组单词。观察前半部分:它在主题上形成分布，而主题又在文档中的下一个单词上形成分布。然后我们在这个分布下计算后半部分。一个好的文档模型应该在下半年提供更高的日志可能性。(对于所有方法，我们通过将设置为变分平均值来逼近可能性。)



(a) 在 20NewsGroup 数据集上，通过主题连贯性和主题多样性的归一化乘积（越高越好）与通过文档完成的归一化对数似然（越高越好）来衡量的预测性能相比，主题质量。



(b) 通过主题连贯性和主题多样性的归一化乘积衡量的主题质量（越高越好）与在纽约时报数据集上通过文档完成的归一化对数似然衡量的预测性能（越高越好）

图 4. 20NewsGroups 和纽约时报数据集在不同词汇量下的表现。在这两个图上，更好的模型位于右上角。总的来说，ETM 是一个更好的主题模型。

我们计算文档完成任务的对数似然度（Rosen-Zvi 等人，2004；Wallach 等人，2009）。我们将每个测试文档分成两组单词。观察到前半部分：它引起了主题的分布，而主题又引起了文档中下一个单词的分布。然后，我们在此分布下评估下半年。一个好的文档模型应该在下半年提供更高的对数似然。 （对于所有方法，我们通过将 设置为变分平均值来近似似然。）

我们研究语料库和不同的词汇。图 4 显示了主题质量作为预测能力的函数。 （为了便于可视化，我们通过减去均值并除以标准差来标准化这两个指标。）最好的模型在右上角。

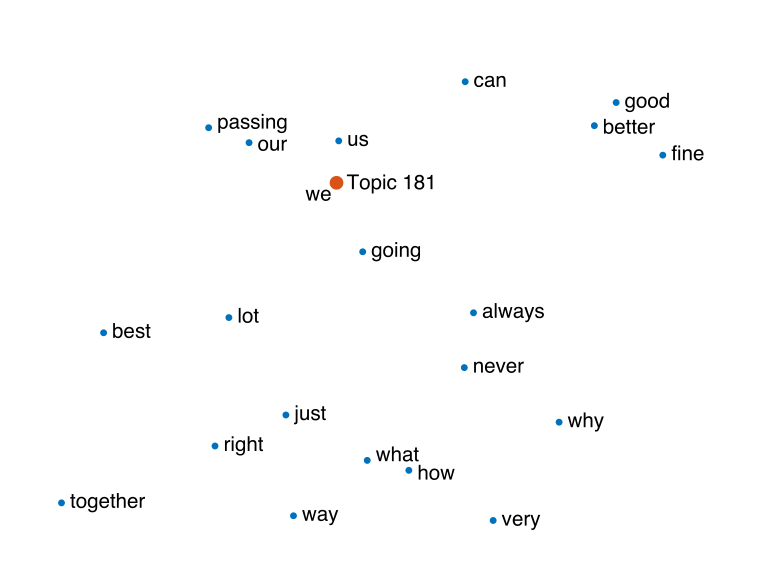
LDA 在几乎所有情况下都预测最差。在 20NewsGroups 上，NVDM 的预测总体上优于 LDA，但比其他方法差；在《纽约时报》上，NVDM 给出了最好的预测。然而，NVDM 的主题质量远低于其他方法。 （它不提供“主题”，因此我们评估其 矩阵的可解释性。）在预测中，两个版本的 ETM 至少与单纯形约束的 ∆-NVDM 一样好。

这些数字表明，在可解释模型中，ETM 在保持可解释主题的同时提供了最佳预测性能。它对大型词汇表是健壮的。

6.1 停用词

我们现在研究一个包含所有停用词的纽约时报语料库版本。我们删除不常用的单词以形成大小为 10,283 的词汇表。我们的目标是表明即使在存在停用词的情况下，标记的 ETM 也能提供可解释的主题，这是主题模型通常失败的另一种机制。特别是，鉴于停用词出现在许多文档中，传统的主题模型会学习包含停用词的主题，而不管主题的实际语义如何。这导致较差的主题可解释性。

图 5. ETM 在《纽约时报》上找到的包含停用词的主题。即使存在停用词，ETM 也是稳健的。



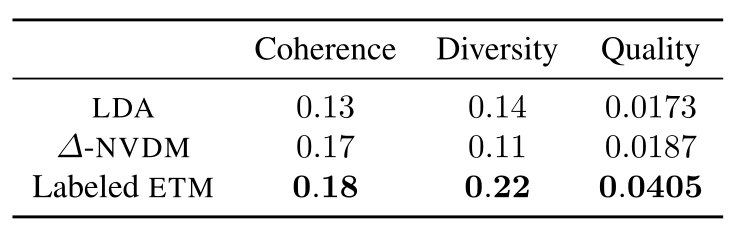
我们用 K = 300 个主题拟合 LDA、Δ-NVDM 和标记的 ETM。 （我们不报告 NVDM，因为它不提供可解释的主题。）表 3 显示了主题质量（主题连贯性和主题多样性的产物）。总体而言，带标签的 ETM 在主题质量方面表现最好。

虽然 ETM 有一些特定于停用词的“停用词”（参见例如图 5），但 ∆NVDM 和 LDA 几乎在每个主题中都有停用词。 （由于空间限制，此处未显示主题。）原因是停用词与其他单词在相同的文档中同时出现。因此，传统的主题模型难以区分实词和停用词。带标签的 ETM 识别嵌入空间中停用词的位置；它让他们开始讨论自己的话题。

7 结论

我们开发了 ETM，这是一种将 LDA 与词嵌入结合的文档生成模型。 ETM 假设主题和词存在于相同的嵌入空间中，并且词是从分类分布生成的，其自然参数是词嵌入和指定主题的嵌入的内积。

表 3. 存在停用词时《纽约时报》数据的主题质量。主题质量是主题连贯性和主题多样性的产物（越高越好）。标记的 ETM 对停用词具有鲁棒性；与没有停用词时相比，它实现了类似的主题连贯性。



ETM 学习可解释的词嵌入和主题，即使在具有大量词汇的语料库中也是如此。我们针对几个文档模型研究了 ETM 的性能。 ETM 学习连贯的语言模式和准确的单词分布。