◂ BIT ▸

话题发现

成员：李艳涛 冯嘉伟 刘伟杰 严睿逸 刘溪伟 王铭远

时间：2021/11/15

北京理工大學
BEIJING INSTITUTE OF TECHNOLOGY

德以明理　学以精工

北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

目录

CONTENTS

1 概述

2 文本向量化表示

3 文本主题模型

4 基于短文本的话题发现

5 技术前沿

6 Demo展示

北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

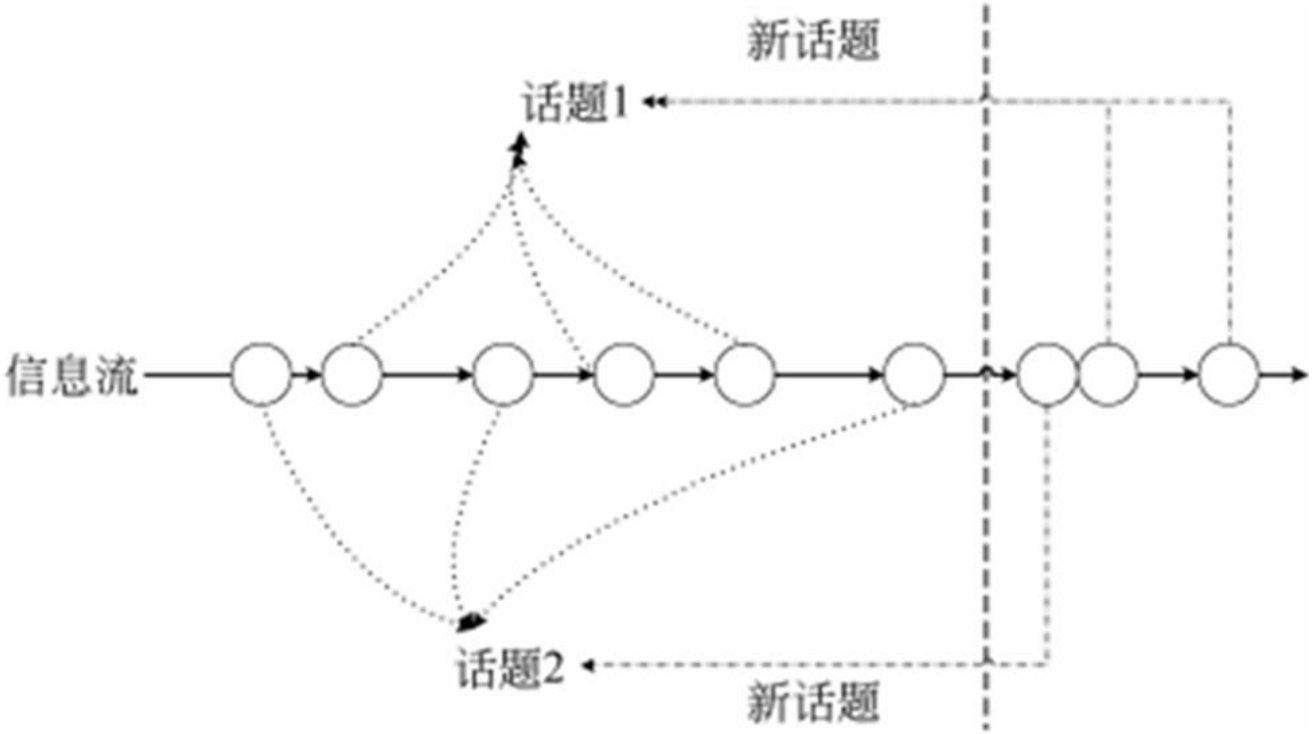
1

概述

讲解：李艳涛

话题发现

又称话题检测，是指将新闻专线和新闻报道等来源的数据流中的信息归入不同的话题，并在必要时建新话题的技术。



话题发现流程

【重庆万州大巴坠桥事件】#重庆大巴车坠江# 虽然现在事故原因还在等待最后的调查结果，但听说是因为私家车上女司机逆行+穿高跟鞋引起，魔王实在忍不住想吐槽几句，难道这位女司机的驾照是马路上捡的吗？！！一点交通常识没有！而且招招害人害己！！！当然，这也给我们司机朋友们又提了一个醒，遵守交规并不是儿戏，对待生命也不是儿戏，别拿生命开玩笑！

幸亏不是工作日，希望大巴车上人不多🙏🙏🙏，都相安无事，为他们祈祷❤️🙏🙏🙏

重庆万州公交车坠江事故原因查明

新华社 2018-11-03 新华社官方帐号 关注

新华社重庆11月2日电(记者陈国洲、韩振)11月2日上午,重庆万州公交车坠江事故调查处置部门发布消息,此次事故原因已经查明,系乘客与驾驶员发生争执互殴引发。

公安机关先后调取监控录像2300余小时、行车记录仪录像220余个片断,排查事发前后过往车辆160余车次,调查走访现场目击证人、现场周边车辆驾乘人员、涉事车辆先期下车乘客、公交公司相关人员及涉事人员关系人132人。10月31日零时50分,潜水人员将车载行车记录仪及SD卡打捞出水后,公安机关多次模拟试验,对SD卡数据成功恢复,提取到事发前车辆内部监控视频。

- “重庆万江公交车坠江事件”
- 发生之初，关于女司机驾驶红色轿车逆行导致事故发生的新闻铺天盖地
- 经过调查，发现事故起因是乘客与司机激烈争执互殴导致车辆失控，证明了女司机的清白
- 针对该事件的舆论给当事人带来了巨大的伤害

意义

- 由于网络传播的不可控性，部分新闻话题给民众传播了与事实不符的信息，进而产生不良的社会影响
- 话题发现能够有效地将数据流中的未知话题进行提取识别。
- 话题发现用于舆情监测，可以帮助政府有效地了解并引导社会舆论走向
- 还可以帮助公司企业掌握其在社会的最新动态，制定营销策略

文本表示：为文本数据建立计算机可处理的结构化表示模型

传统文本聚类：通过利用文档之间的一些相似度度量来集群文档。将文档内容描述相似、在语义上可能是对同一话题进行报道的文本分类到一个文本集合中。

主题模型：以非监督学习的方式对文本集的隐含语义结构进行聚类的统计模型 。

4 发展

- SinglePass
- 改进的SinglePass
- 层次聚类
- LDA
- 基于gibbs采样的LDA
- ATM
- DTM
- KeyGraph(Sayyadi&Raschid，2013)
- KeyGraph+
- Idea-Graph(Zhang，2016)
- LDA-IG：综合Idea-Graph和LDA
- ……

- James Allan等人于1998年提出一种基于单遍聚类算法（SinglePass）的话题检测框架，为话题发现研究奠定了关键基础。
- 后续研究都是通过改进文本表示方式、聚类算法、引入其它自然语言处理技术等各个方面来提升话题发现算法。

北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

2

文本向量化表示

讲解：冯嘉伟

向量表示分类

原始文本→ 分词→ 清洗→ 标准化→ 特征提取→ 建模

离散表示

向量的每个元素代表了一个词。

常见的离散表示模型有：

One-hot，词袋模型（BOW），TF-IDF、N-Gram。

分布式表示

用一个词周围词来表示该词。

常见的分布式表示模型有：

共现矩阵，Word2Vec，ELMO，GPT，BERT。

One-hot

向量化步骤：
•用构造文本分词后的字典
•对词语进行One-hot编码

减少维度

缺点：
维数过高：语料增加导致维度灾难
矩阵稀疏：词向量只有1维为1，其他全为0
不能保留语序和语义：“我帮你”和“你帮我”

BOW

向量化步骤：
•用构造文本分词后的字典
•对词语每个词语出现的频次向量化

缺点：
矩阵稀疏：语料过多，不频繁词过多
不能保留语序和语义：认为“我喜欢北京”和“我不喜欢北京”相似

TF-IDF

主要思想：
字词的重要性随着它在文件中出现的次数成正比增加，但同时会随着它在语料库中出现的频率成反比下降。

$$\text{TF-IDF}(t,d)=\text{TF}(t,d)\times\text{IDF}(t)$$

$$\text{IDF}(t)=\log\frac{\text{文章总数}}{\text{包含单词}t\text{的文章总数}+1}$$

优点：考虑了**文档与文档之间的联系**
缺点:不能保留语序和语义

TF(t,d) 表示 词语t 在 文档d 中出现的频率
IDF(t)是逆文本频率指数，它可以衡量单词t用于区分这篇文档和其他文档的重要性

N-Gram

主要思想：
为了保持词的顺序，做了一个滑窗的操作，例如2-gram模型，也就是把2个词当做一组来处理，然后向后移动一个词的长度。以此来生成字典。

当N=2时
你帮我->字典词语(你，你帮，帮，帮我，我)
我帮你->字典词语(我，我帮，帮，帮你，你)

{"你": 1, "你帮": 2, "帮": 3, "帮我": 4, "我": 5, "我帮": 6, "帮你": 7}
你帮我[1, 1, 1, 1, 1, 0, 0]
我帮你 [1, 0, 1, 0, 1, 1, 1]

优点：一定程度上**考虑了词的顺序**
缺点:随着N的增大，词表迅速膨胀，数据出现大量稀疏的问题

离散表示方法对比

向量表示方法	优点	缺点
One-hot	将每个词离散化	维数过高,矩阵稀疏,不能保留语序
BOW	减少特征维数	稀疏性，不能保留语序，不能保留语义
TF-IDF	考虑了文档与文档之间的联系	不能保留词语在句子中的位置关系
N-Gram	一定程度上考虑了词的顺序	随着N的增大，词表迅速膨胀，数据出现大量稀疏的问题

共现矩阵

主要思想：
设置滑动窗口大小，可以得到一个词典。考虑词组共同出现的次数。词文档的共现矩阵主要用于发现主题。

I like deep learning.
I like NLP.
I enjoy flying.

{"I like","like deep","deep learning","like NLP“,
"I enjoy","enjoy flying“}

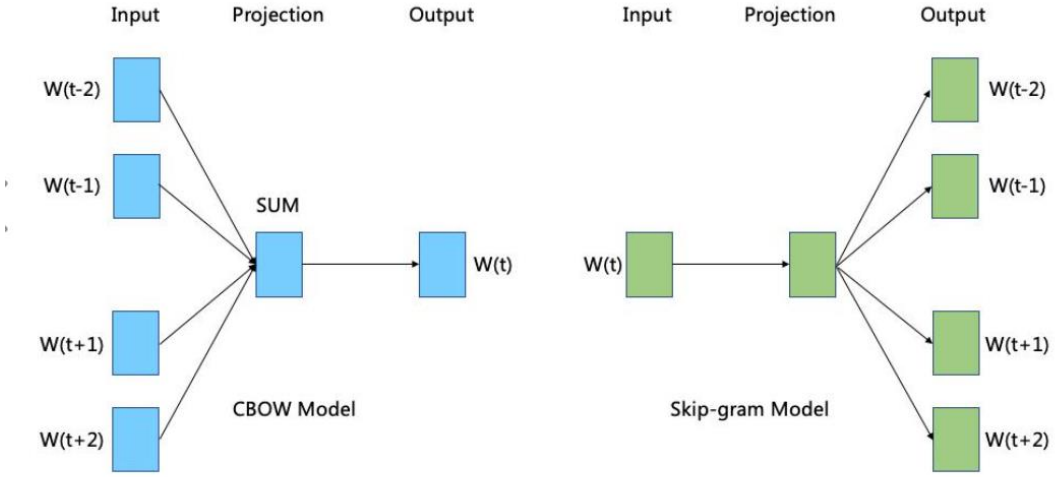
优点：
一定程度上**考虑了词的顺序**
缺点:
向量维数随着词典大小线性增长
存储整个词典的空间消耗非常大
一些模型如文本分类模型会面临稀疏性问题
模型会欠稳定，每新增一份语料进来，稳定性就会变化

北京理工大學 BEIJING INSTITUTE OF TECHNOLOGY

Word2Vec

主要思想：
谷歌2013年提出的词嵌入模型之一，是一种浅层的神经网络模型，它有两种网络结构，分别是CBOW和Skip-gram。

CBOW是用周围词预测中心词，从而利用中心词的预测结果情况，而Skip-gram是用中心词来预测周围的词。



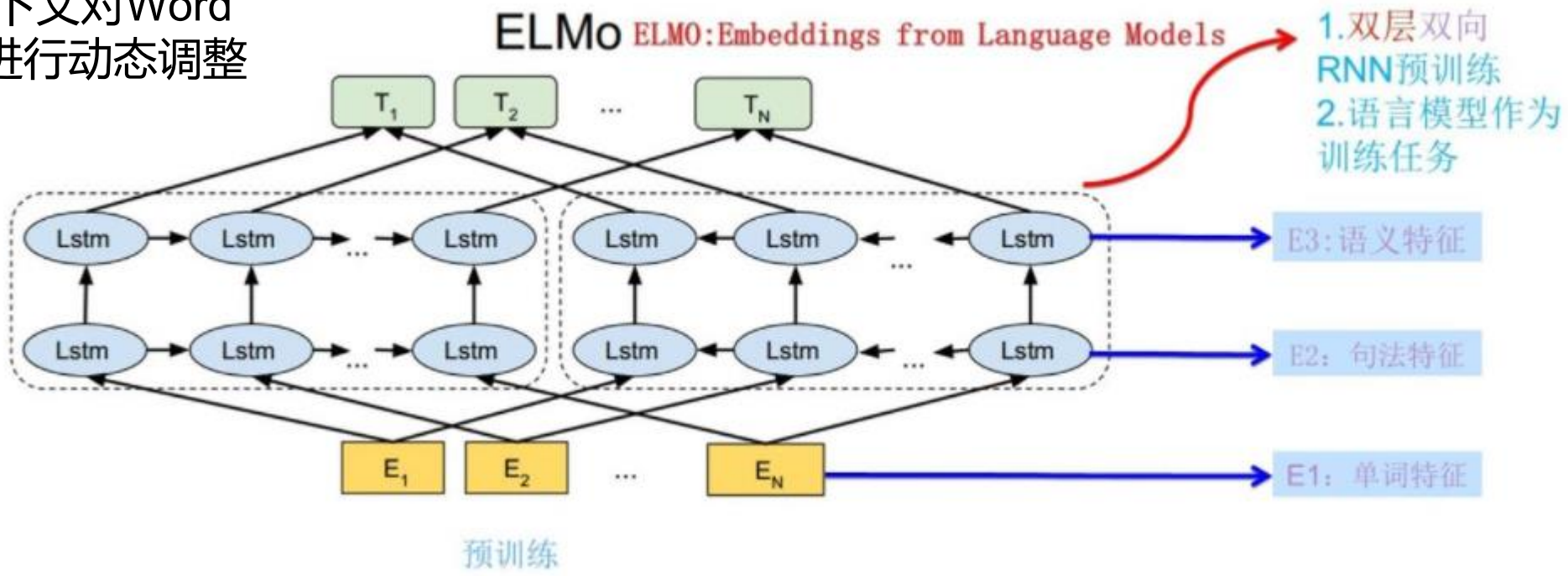
Skip-gram学习的词向量更细致，利于表征大量低频词
CBOW效率高，速度更快

优点：考虑上下文，效果好，维度更少，速度快，通用性很强
缺点：无法解决**一词多义**

ELMO

解决了一词多义问题

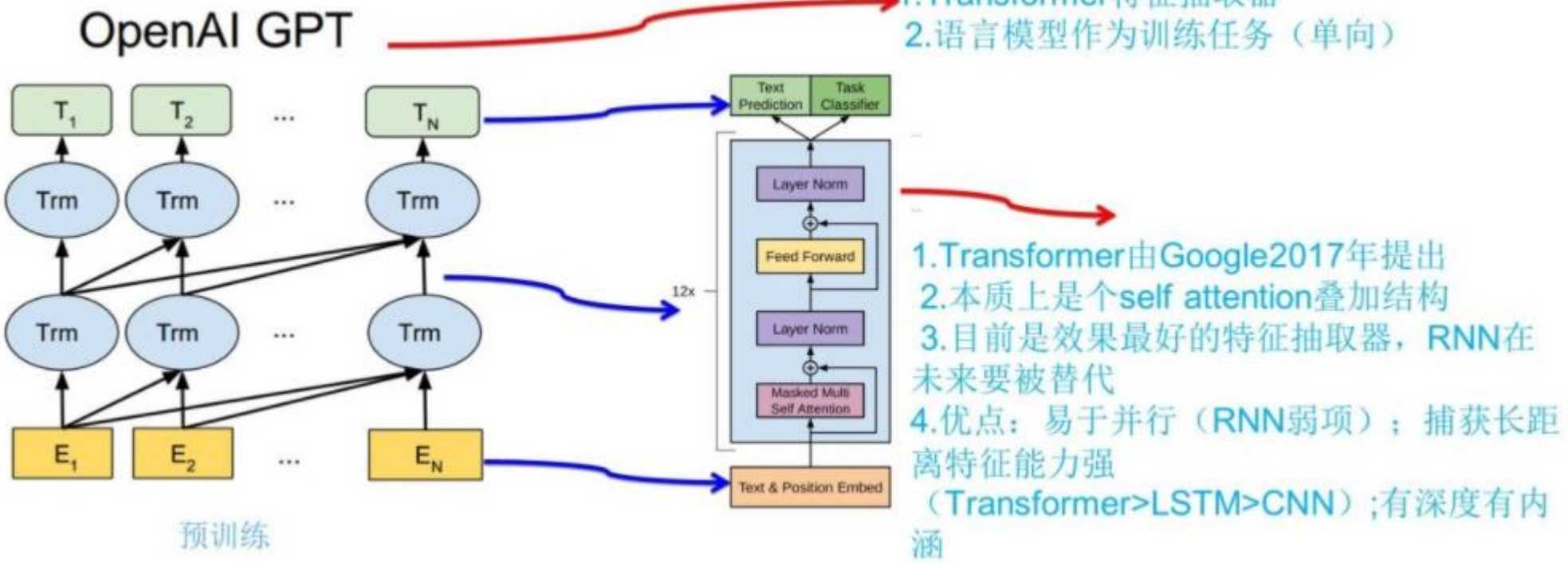
主要思想：
根据当前上下文对Word Embedding进行动态调整



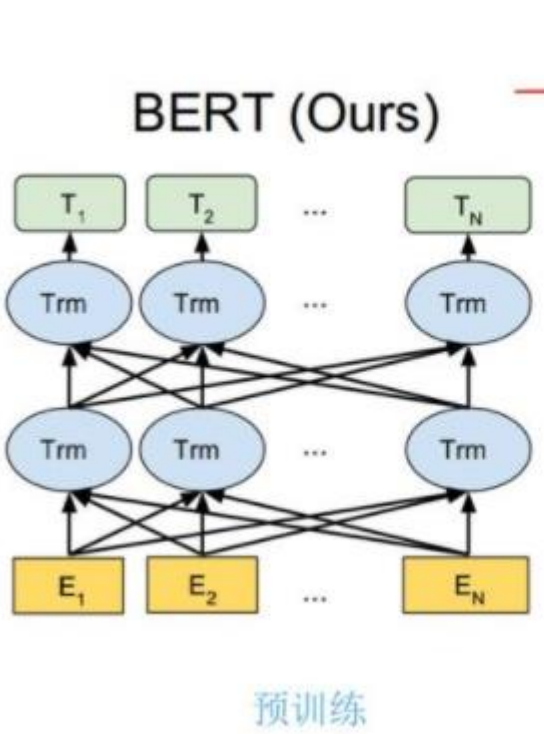
北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

GPT

使用transformer进行提取特征



BERT 利用双向语言模型+ transformer进行预训练

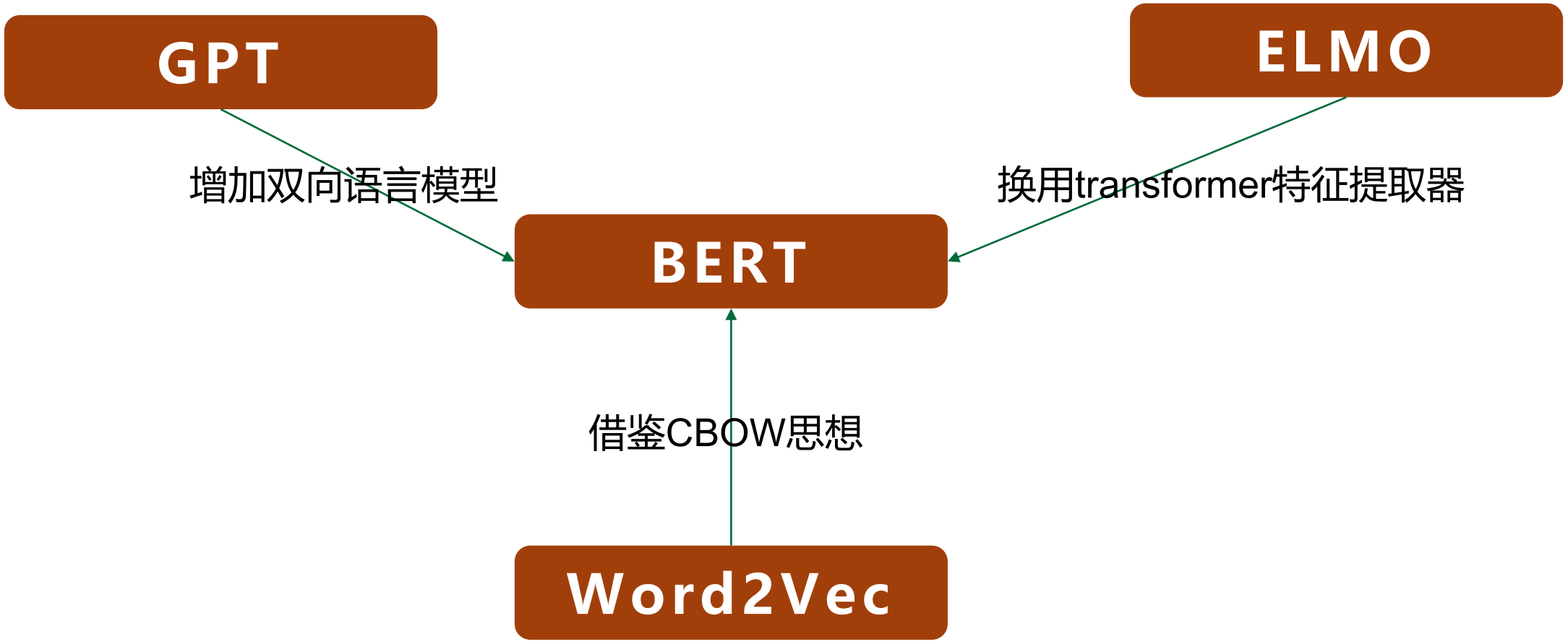


1.Transformer特征抽取器
2.语言模型作为训练任务（双向）

Masked 语言模型：
即随机选择一部分单词进行mask，然后预测这些单词，其方式和CBOW类似

Next Sentence Prediction：
考虑到很多NLP任务是句子关系判断任务，单词预测粒度的训练到不了句子关系这个层级

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[J]. 2018.


GPT
ELMO
增加双向语言模型
换用transformer特征提取器
BERT
借鉴CBOW思想
Word2Vec

分布式方法对比

向量表示方法	优点	缺点
Word2Vec	最早的预训练模型	无法解决一词多义问题
ELMO	解决一词多义问题	1.LSTM特征提取较弱 2.向量拼接方式融合上下文特征融合能力较弱
GPT	使用transformer提取特征	使用单向的语言模型
BERT	使用双向语言模型, Masked 语言模型, Next Sentence Prediction	Fine-tuning阶段看不到预训练阶段的MASK标记

北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

3

文本主题模型

讲解：刘伟杰

文本主题模型

潜在语义分析模型
Scott Deerwester, 1990

概率潜在语义分析模型
Hofmann, 1999

潜在狄利克雷分配模型
Blei, D M.、Ng A Y.、
Jordan M I.， 2003

潜在语义分析（Latent Semantic Analysis, LSA）模型是最经典的文本主题模型之一。其原理是统计各文档中出现的词语，构造一个“单词—文档”矩阵，然后对该矩阵进行奇异值分解（SVD）并进行降维，最后使用得到的矩阵构建潜在语义空间。

	$文本_1$	$文本_2$	…	$文本_n$
$词_1$	0.0017	0.0075	…	0.0029
$词_2$	0	0.0014	…	0
⋮	⋮	⋮	…	⋮
$词_m$	0.008	0.0046	…	0.0056

其中，第 i 行第 j 列的元素a_{ij}，是字典中第 i 个词在第 j 篇文档中的TF-IDF值。

优点

- 降维可去除部分噪声
- 无监督
- 与语言无关

缺点

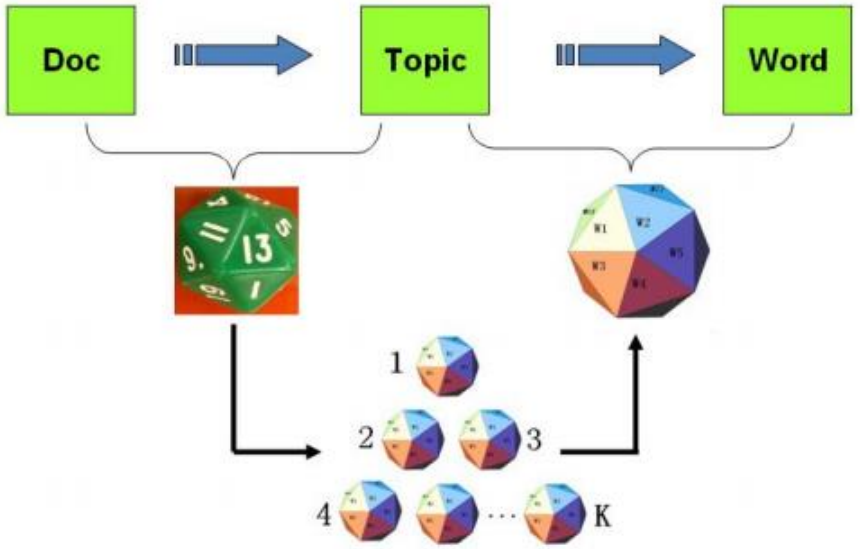
- SVD的计算复杂度相对较高
- 新数据来时要重新训练模型
- 得到的不是一个概率模型，缺乏统计基础，结果难以直观地解释

针对LSA模型缺乏概率统计基础的问题，概率潜在语义分析模型（Probabilistic Latent Semantic Analysis, PLSA）被提出。

1. 确定文章的K个主题
2. 重复选择K个主题之一，按照主题-词语概率生成词语
3. 所有词语组成文章

关键：

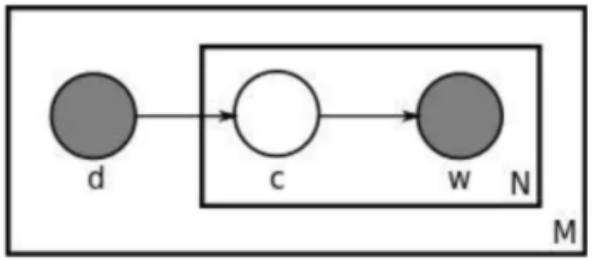
文章->主题，主题->词语



假设语料库中共有 M 篇文档，考虑第 m 篇文档 d_m 的生成过程：

1. 随机选择一个“文档-主题”骰子（共有 K 个面，每一个面代表一个主题）
2. 投掷“文档-主题”骰子得到一个主题（编号 z ）
3. 投掷编号为 z 的“主题-词语”骰子 z ，得到词语 w

重复以上过程 n 次，得到一篇词语量为 n 的文档



词 w 生成的概率：$p(w|d_m) = \sum_{z=1}^{K} p(w|z)p(z|d_m) = \sum_{z=1}^{K} \varphi_{zw}\theta_{mz}$

↓ n次

整篇文档n个词生成的概率：$p(\vec{w}|d_m) = \prod_{i=1}^{n}\sum_{z=1}^{K} p(w_i|z)p(z|d_m) = \prod_{i=1}^{n}\sum_{z=1}^{K} \varphi_{zw_i}\theta_{dz}$

PLSA模型最优化包含两个参数φ_{zw}和θ_{mz}求解，可以使用期望最大化算法 (EM) 计算

优点

- 解决了同义词和多义词的问题
- 利用 EM 算法训练潜在参数
- 相对LSA，有了坚实的统计学基础

缺点

- PLSA训练的参数会随着文档和词语数目增多而增多
- 只能生成其所在数据集的文档模型，无法生成新文档的模型
- 忽视了文档主题分布的先验知识

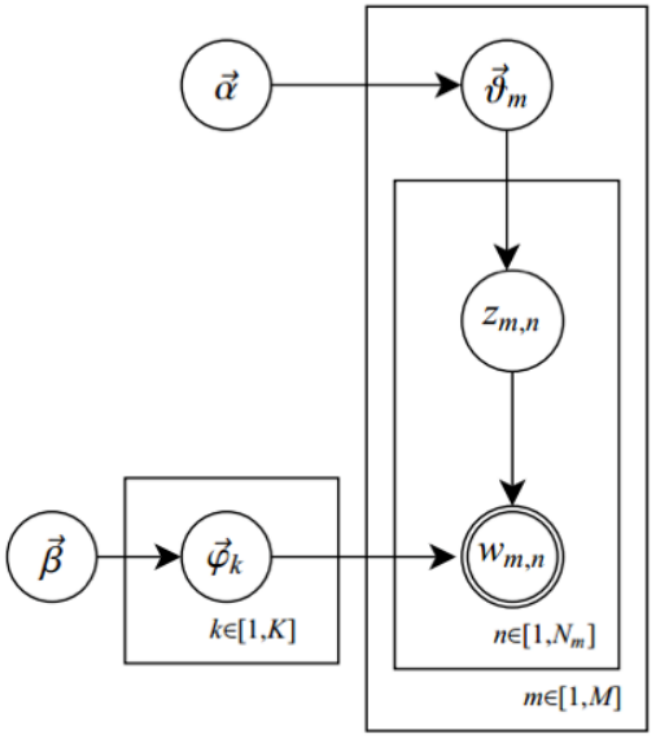
为了解决 PLSA 中出现的主要问题，潜在狄利克雷分配模型（LDA）被提出，这个模型也成为了主题模型这个研究领域内应用最广泛的模型，从根本上来讲，LDA 模型是在 PLSA 模型上引入了**参数先验分布**的概念。

在LDA模型中，每个文档关于话题的概率分布和每个话题关于词语的概率分布都被赋予了一个稀疏形式的**狄利克雷先验**，可以看成是编码了人类的先验知识：

- 一篇文章的主题更有可能是集中于少数几个话题上，很少在很多话题上都有涉猎且没有重点。
- 多数情况下，一个话题中只有少部分与这个话题高度相关的词语的出现频率会很高，而其他词出现的频率则明显偏低。

LDA生成文档的过程：

1. 按照先验概率 $p(d_i)$ 选择一篇文档 d_i
2. 从 Dirichlet 分布 α 中取样生成文档 d_i 的主题分布 θ_i
3. 从主题的多项式分布 θ_i 中取样生成文档 d_i 第 j 个词的主题
4. 从 Dirichlet 分布 β 中取样生成主题 $z_{i,j}$ 对应的词语分布 $\emptyset_{zi,j}$
5. 从词语的多项式分布 $\emptyset_{zi,j}$ 中采样最终生成词语 $w_{i,j}$



LDA训练的目标：

- 估计模型中的参数 φ 和 θ
- 对于新来的一篇文档，能够计算这篇文档的主题分布 θ

LDA训练的过程：

1. 对语料库中每篇文档的每个词语 w，随机赋予一个主题编号 z
2. 重新扫描语料库，对每个词 w，使用 Gibbs Sampling 公式进行采样，求出它的主题，然后在语料库中更新
3. 重复步骤 2，直到 Gibbs Sampling 收敛
4. 统计语料库的“主题-词语”共现频率矩阵，该矩阵就是 LDA 的模型

ATM模型（author-topic model）也是“概率主题模型”家族的一员，是LDA主题模型（Latent Dirichlet Allocation ）的拓展。

它能对数据集中作者的写作主题进行分析，并找出某个作家的写作主题倾向，以及找到具有同样写作倾向的作家，是一种新颖的主题探索方式。

	Author	Score	Size
297	YannLeCun	1.000000	11
203	PatriceSimard	0.999977	8
76	EduardSackinger	0.999712	3
114	J.S.Denker	0.999598	3
111	I.Guyon	0.997464	5
204	PatriceY.Simard	0.916426	4
51	D.Henderson	0.899918	4
156	L.D.Jackel	0.851938	4
139	JohnS.Denker	0.842672	6
163	LeonBottou	0.832937	3

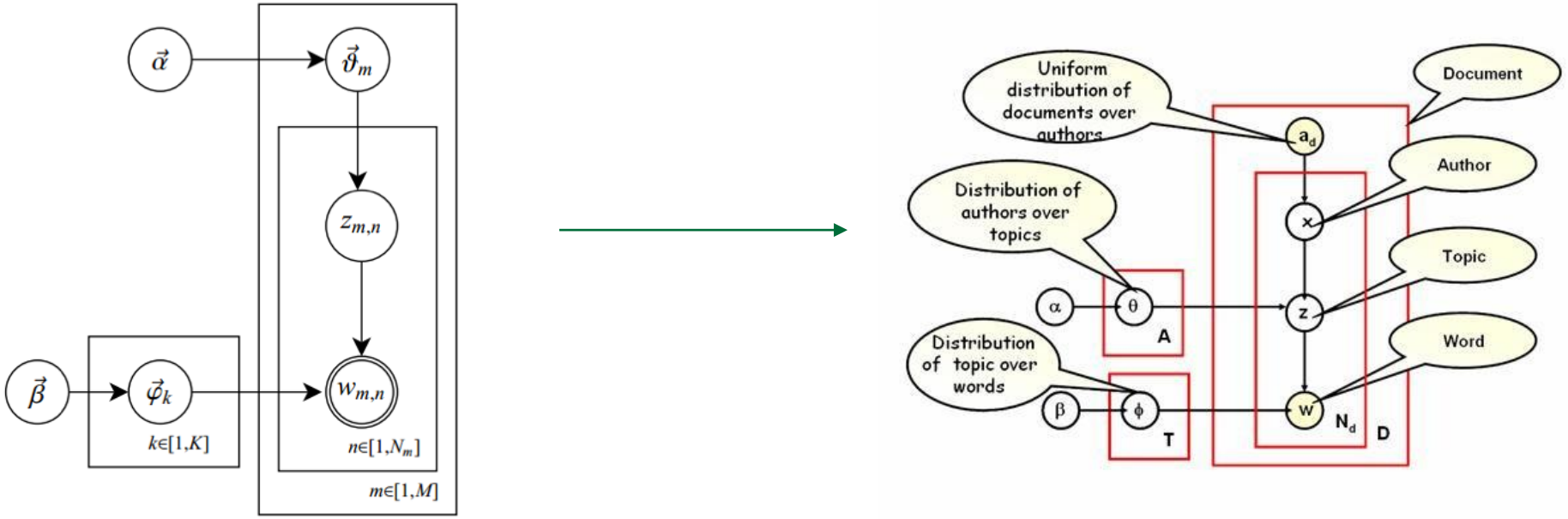
ATM模型思想

在作者主题模型中，设作者的主题概率分布为θ，主题的词汇概率分布为Φ。那么由该模型生成文档可分为3步：

①计算作者的主题多项式概率分布θ；

②计算主题的词汇多项式概率分布Φ；

③对于一个作者，按照概率分布θ、Φ抽取主题和词汇，计算过程重复N次，形成了该作者的一组文档。

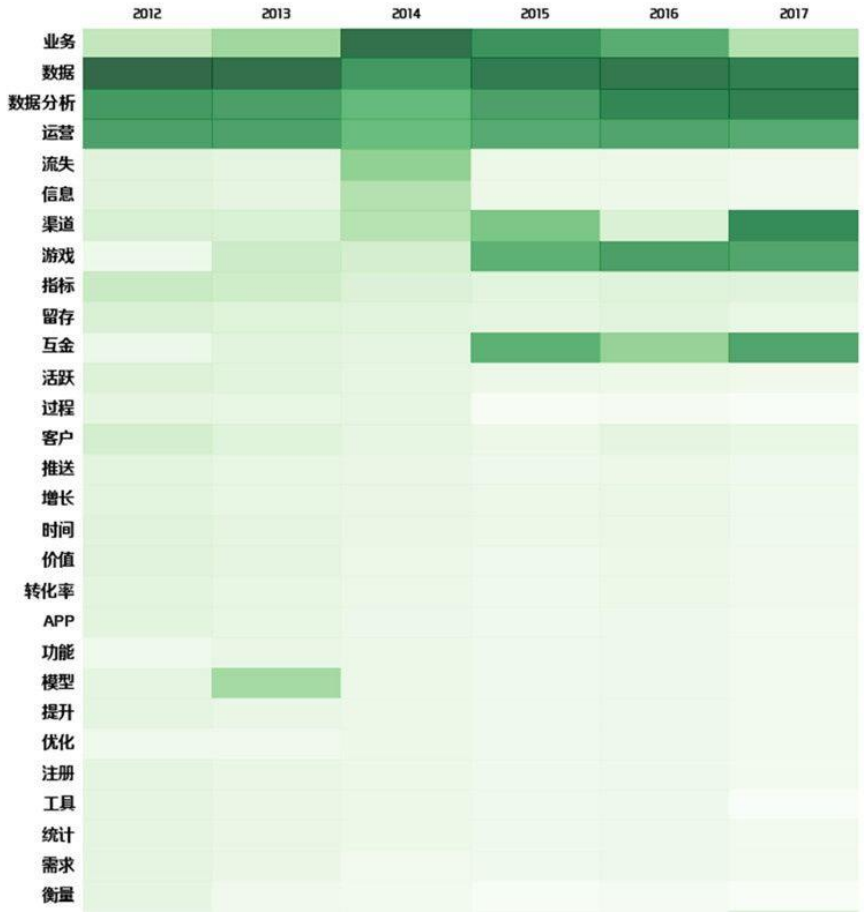


Lda到Atm的变化

DTM模型

动态主题模型（Dynamic Topic Models），即蕴含时间因素的主题模型，其使用背景有些类似于“忒修斯之船”，在时间尺度之下，时间开始和结尾文档的主题词可能完全不相同，尽管如此，这两篇文档仍然属于同一主题。

右图为某一博主对主题“流量运营&数据分析”所整理的热力图，从图中可以得知，随着年份的增长，“互金”，“指标”等词在热度上有了明显的上升。

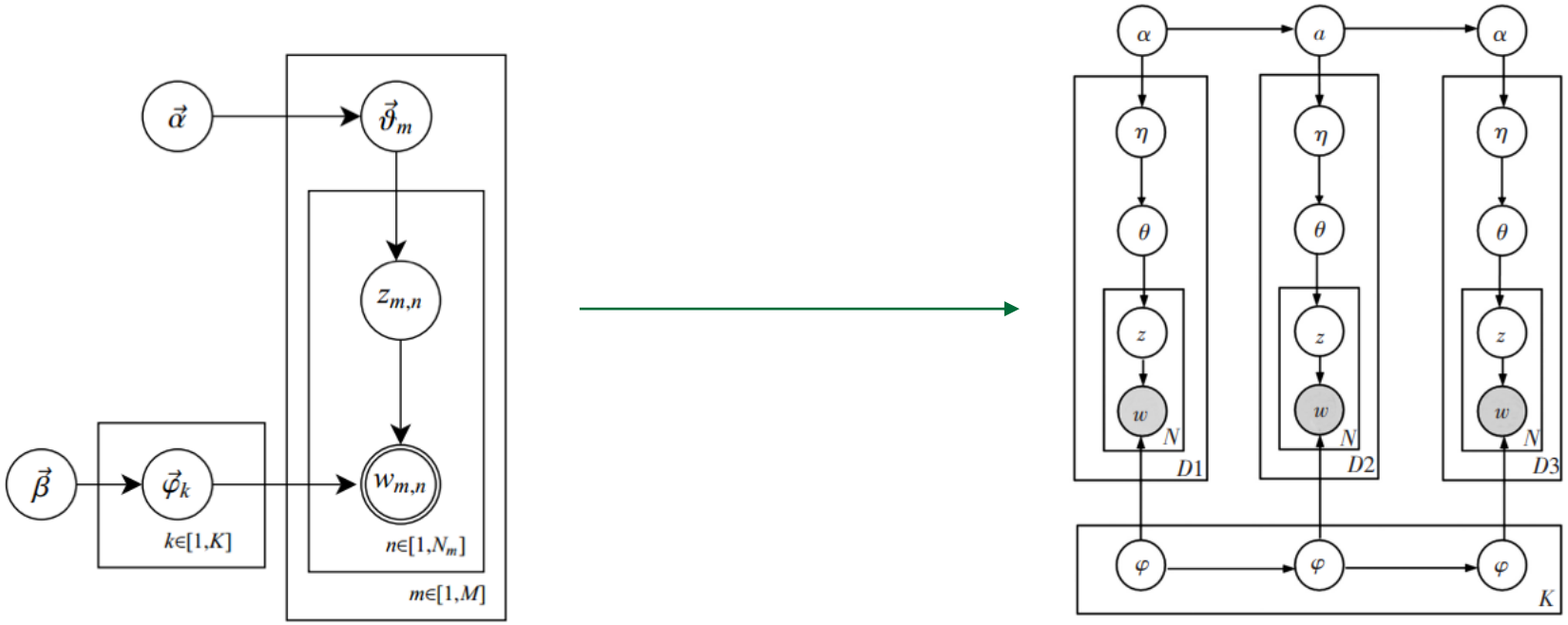


DTM模型思想

DTM 模型是一种无监督的动态时序主题模型。其基本思想分为两个部分。

首先，获得所获取的数据源的时间信息，将时间按照一定的时间段大小进行划分，然后将文档集合中的文档根据其内在的时间戳信息划分到相应的时间片中。

其次，对每一个时间片中的文档子集通过LDA进行主题挖掘得到主题随时间动态演化的情况。每一个时间片上的分布结果根据之前一个时间片的主题训练结果进行动态变化。



北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

4

基于短文本的话题发现

讲解：刘淏伟

短文本主题

在短文本（例如微博评论与即时消息）中发现主题已经成为许多内容分析应用程序的一项重要任务。

然而，直接将传统的主题模型(如LDA和PLSA)应用于这类短文可能效果不佳。其根本原因在于，传统的主题模型隐式地捕获文档级的单词共现模式来揭示主题，因此在短文档中存在严重的数据稀疏问题。

本ppt讲述两个短文本话题发现思想：
1、基于外部数据集的短文本话题发现
2、BTM主题模型

北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

4.1

基于外部数据集的短文本话题发现

基于外部数据集的短文本话题发现

由于许多与文本和Web的短片段相关的分类任务，如搜索片段、论坛和聊天消息、博客和新闻提要、产品评论、书籍和电影摘要，由于数据稀疏，无法实现较高的准确性。因此，在获取话题前，先获取外部知识，使数据更加相关，同时扩大分类器的覆盖范围，更好地处理未来的数据。

对于每个分类任务，先收集一个称为“通用数据集”的大规模外部数据集，然后在一个(小)标记训练数据集和从该数据集发现的丰富的隐藏主题集上构建一个分类器。

思想步骤

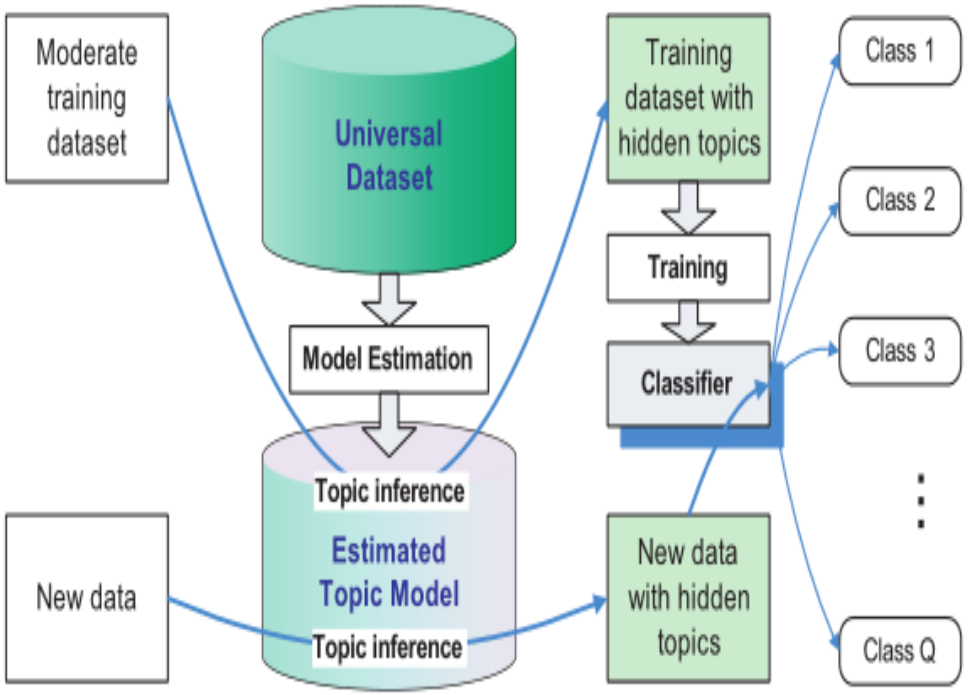
(a)选择合适的“通用数据集”。

(b)对通用数据集进行主题分析。

(c)构建中等大小的标记训练数据集。

(d)对训练和未来数据进行主题推理。

(e)构建分类器。



优缺点

优点

1、减少数据稀疏性
2、扩大分类器的覆盖范围
3、灵活的半监督学习
4、容易实现

缺点

1、过于依赖外部数据集的质量
2、对非专业性文本而言效果可能不佳

北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

4.2

BTM（Biterm Topic Model）主题模型

BTM主题模型

传统的主题模型基于文档级词共现模式来学习主题，在文本场景中，当每个文档中的词共现模式变得非常稀疏时，该模型的有效性将受到很大影响。为了解决这一问题，BTM提出了一种新的比特词（Biterm）主题模型，该模型通过直接建模整个语料库中所有比特词(即词共现模式)的生成来学习短文本上的主题。

Bitterm表示在短上下文中共现的无序词对(即词共现模式的实例)。这里的短上下文指的是包含有意义的词共同出现的适当的文本窗口。在简短的文本中，由于文档通常是简短和特定的，我们只是将每个文档作为单独的上下文单元。我们从一个简短的文本文档中提取任意两个不同的单词作为术语。

“I VISIT APPLE STORE”

↓

[‘VISIT APPLE’，
’ VISIT STORE’，
’ APPLE STORE’]

BTM主题模型

BTM中语料库的具体生成过程可以描述如下:

1、对每个主题z：得出特定主题单词分布：

$$\phi_z \sim \mathrm{Dir}(\beta)$$

2、为全部集合得出一个主题分布

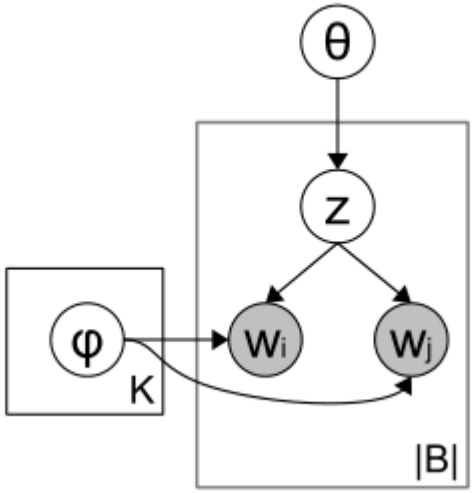
$$\theta \sim \mathrm{Dir}(\alpha)$$

3、对集合B中的每个biterm：

3.1 ： 得出一个主题分配： $z \sim \mathrm{Multi}(\theta)$

3.2 ： 得出两个词： $w_i, w_j \sim \mathrm{Mulit}(\phi_z)$

通过上述步骤得到biterm词对概率与语料库概率：



LDA-U 与词嵌入相结合

北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

5

技术前沿

讲解：刘溟伟 王铭远

北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

5.1

前沿进展

讲解：刘溧伟

我们搜寻了近几年ACL，EMNLP中与主题模型相关的论文，部分结果如下：

TAN-NTM: Topic Attention Networks for Neural Topic Modeling

Tree-Structured Topic Modeling with Nonparametric Neural Variational Inference

CluHTM - Semantic Hierarchical Topic Modeling based on CluWords

Neural Topic Modeling with Bidirectional Adversarial Training

tBERT: Topic Models and BERT Joining Forces for Semantic Similarity Detection

Tree-Structured Neural Topic Model

Topic balancing with additive regularization of topic models

Rethinking Topic Modelling: From Document-Space to Term-Space.

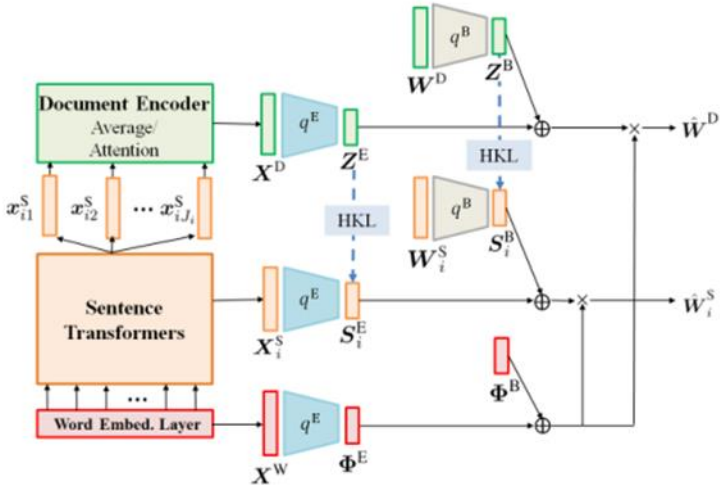
Unsupervised Few-Bits Semantic Hashing with Implicit Topics Modeling.

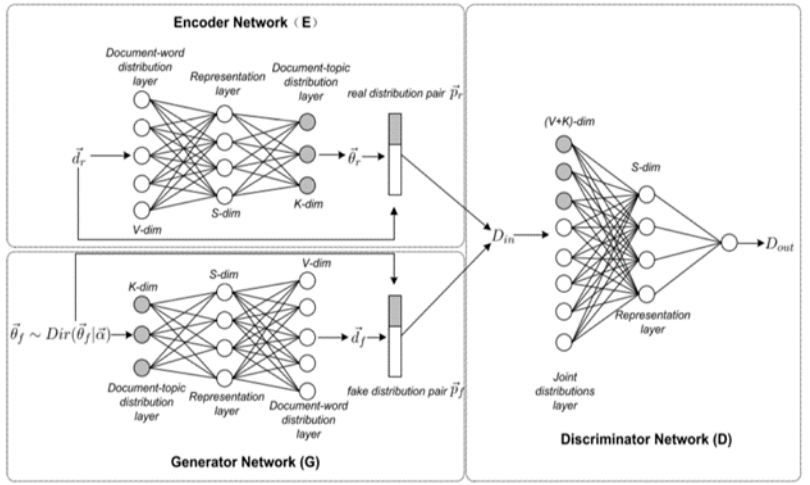
Neural Attention-Aware Hierarchical Topic Model.

Monitoring geometrical properties of word embeddings for detecting the emergence of new topics

总结

根据所搜寻的论文成果，我们得出了以下结论：神经主题模型（NTM）是近几年讨论热度较高的话题，一些论文以NTM为基础提出了新的框架（TAN-NTM: Topic Attention Networks for Neural Topic Modeling），也有一些论文将提出了NTM主题模型的现有问题并提供了解决方法（Neural Attention-Aware Hierarchical Topic Model），或是通过对抗学习的方式来学习文本主题（ Neural Topic Modeling with Bidirectional Adversarial Training ）。





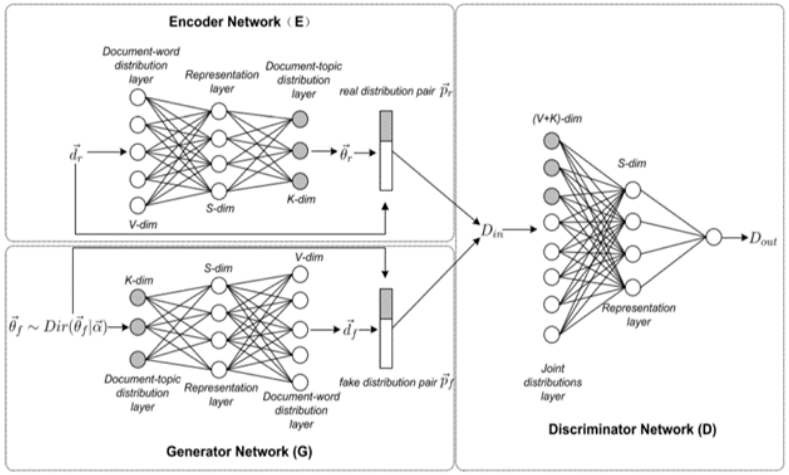
ETM

主题建模分析文档以学习有意义的单词模式。然而，现有的主题模型在使用大而重尾的词汇时无法学习可解释的主题。为此，作者开发了嵌入式主题模型(ETM)，这是一种将传统主题模型与单词嵌入相结合的文档生成模型。具体地说，它用分类分布来建模每个词，其自然参数是词嵌入与其指定主题的嵌入之间的内积。为了拟合etm，作者提出了一种有效的平摊变分推理算法。etm发现可解释的主题，即使有大量的词汇，包括罕见的词和停止词。它在主题质量和预测性能方面都优于现有的文档模型，如潜在的Dirichlet分配模型。

^ Adji B. Dieng, Francisco J. R. Ruiz https://dblp.org/search/pid/api?q=author:Francisco_J._R._Ruiz: , David M. Blei. Topic Modeling in Embedding Spaces. CoRR abs/1907.04907. https://arxiv.org/abs/1907.04907

BATM

近年来，使用神经主题模型从文本中自动抽取主题的兴趣大增，因为它们避免了传统主题模型(如Latent Dirichlet Allocation (LDA))中用于模型推理的复杂数学推导。然而，这些模型要么通常假设潜在主题空间的不恰当先验，要么无法推断给定文档的主题分布。为了解决这些局限性，作者提出了一种神经主题建模方法，称为双向对抗主题(BATM)模型，这是将双向对抗训练应用于神经主题建模的首次尝试。该方法在文档-主题分布和文档-词分布之间建立了双向投影。它使用生成器从文本中捕获语义模式，并使用编码器进行主题推断。在此基础上，进一步扩展了带高斯的双向对抗性主题模型(gauss -BATM)，引入了词的相关性信息。



^ Wang, Rui & Hu, Xuemeng & Zhou, Deyu & Xiong, Yuxuan & Ye, Chenchen & Xu, Haiyang. (2020). Neural Topic Modeling with Bidirectional Adversarial Training. https://www.aclweb.org/anthology/2020.acl-main.32.pdf

北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

5.2

基于多模态的话题发现

讲解：王铭远

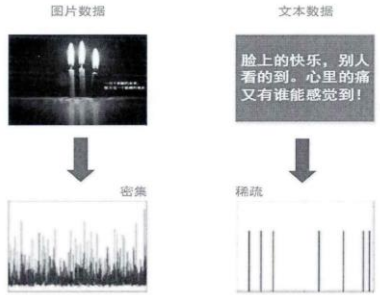
- 每一种信息的来源或者形式，都可以称为一种模态。例如，信息的媒介，有语音、视频、文字等；随着互联网技术的发展，人们获取信息的途径更广，信息载体越来越丰富，图片和视频数据量急剧增长。
- 多模态数据的海量增长给传统的话题发现技术带来了挑战。

多模态融合需要解决的难题及解决思路

多模态信息底层表示的异构性

（1）不同模态数据的表示存在差异。例如：语言通常是符号表示，而语音通常是信号表示。

（2）所承载信息的稀疏程度存在差异



融合过程中的问题

（1）发表的图片、视频与文字主题无关，给话题发现带来很大干扰。

（2）在进行多模态融合时，可能存在信息丢失的问题。

传统解决思路

利用传统方法分别从各类模态信息中提取尽可能多且有效的特征信息，然后利用一些先验规则和连接技术将这些特征信息进行融合，接着提取语义信息。最后再通过某种表达形式将语义信息表达出来，得到最终多模态信息的融合结果。

基于图模型的多模态话题发现

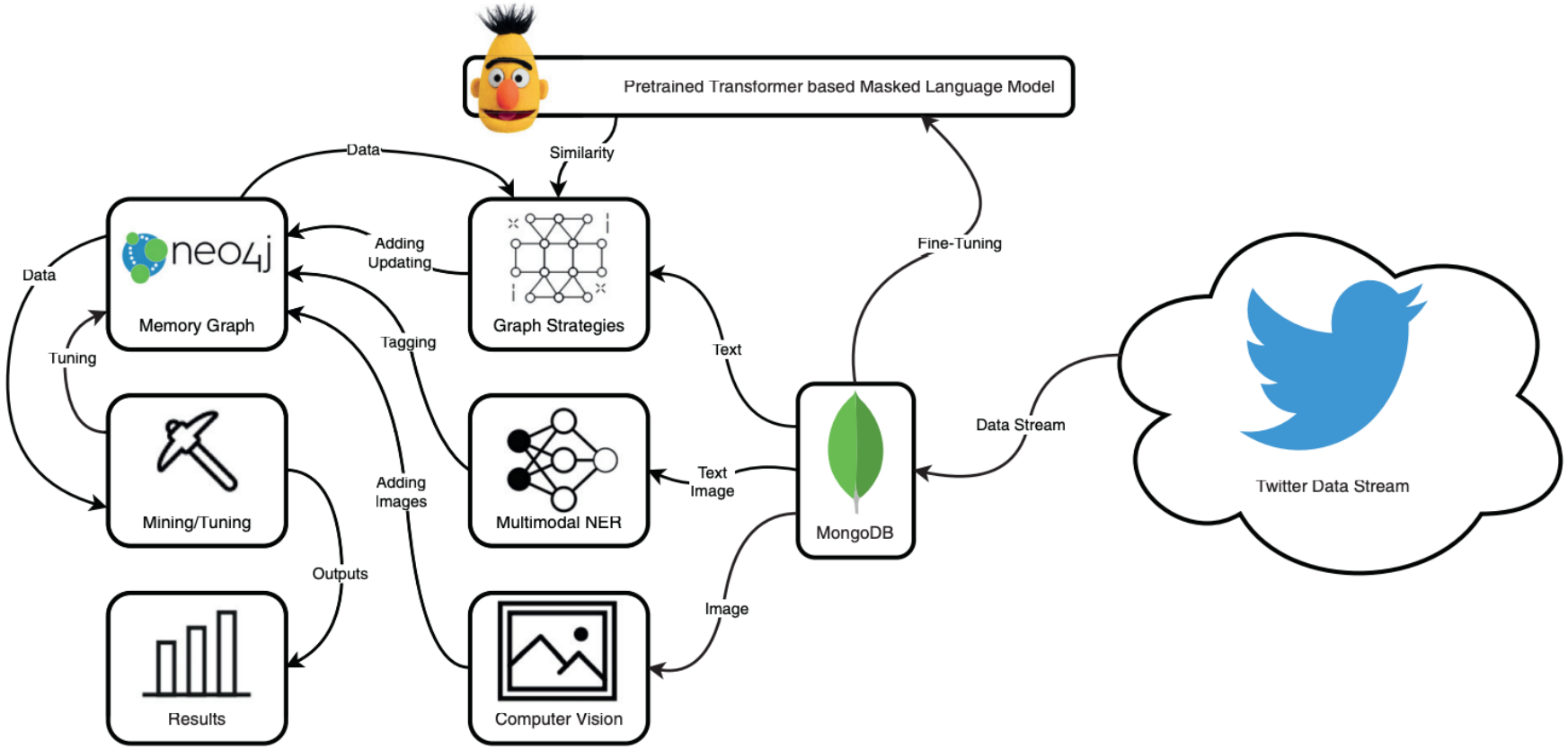


Fig. 1. Proposed System.

基于记忆的图模型主要由多个模块构成（除了数据库部分）：

（1）多模态命名实体识别模型

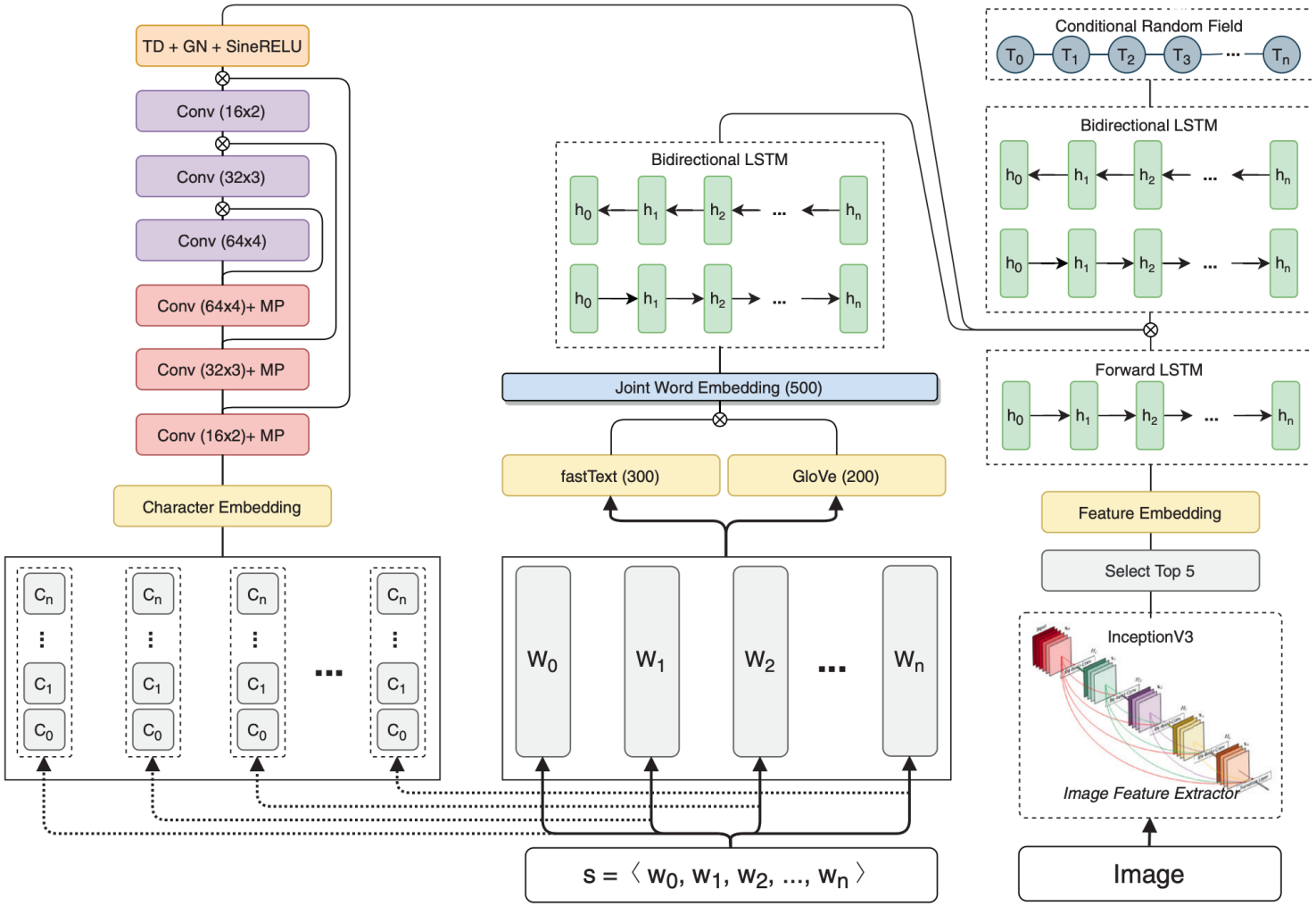
（2）基于记忆的图以及图更新策略

（3）基于图的话题发现

A multimodal deep learning approach for named entity recognition from social media[J]. arXiv preprint arXiv:2001.06888, 2020.

A cognitive approach for topic detection from multimodal post stream using BERT and memory–graph[J].Chaos, Solitons & Fractals, 2021, 151: 111274.

多模态命名实体识别模型（1）

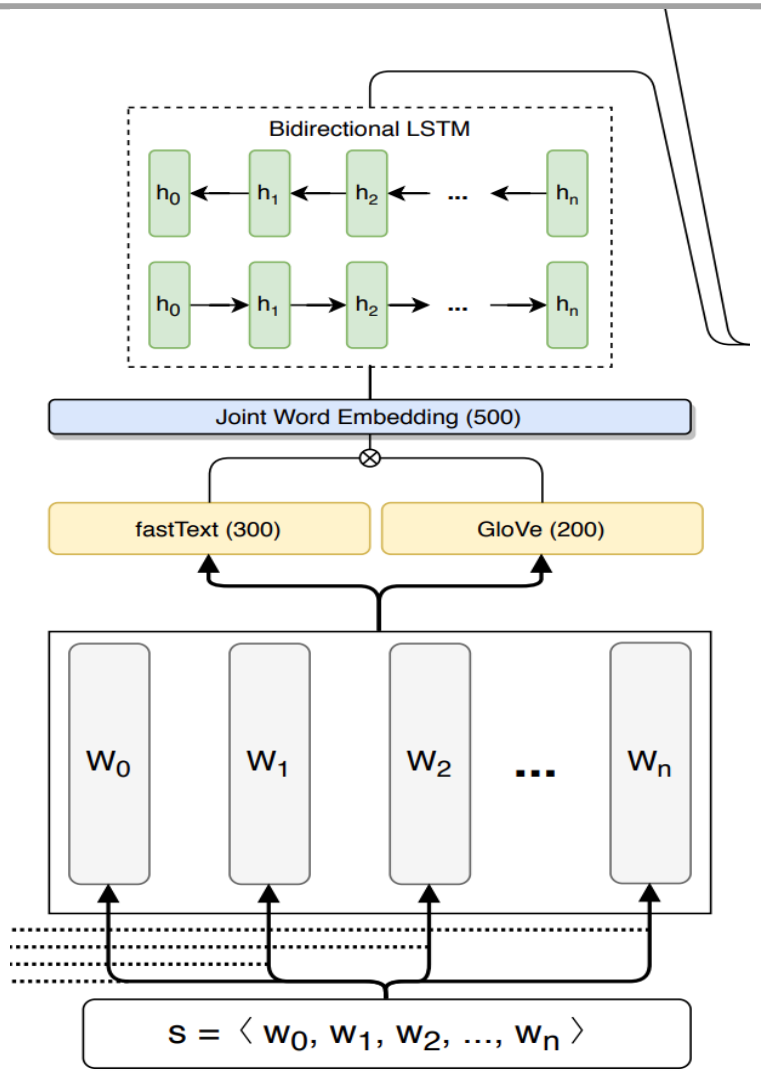


左图显示了多模态命名实体识别器模型，该模型利用推特中的图像和文本数据。并在解决多模态噪声问题中发挥了重要作用。

该方法能够通过从字符、单词和图像三种模态共同学习语义来处理噪声，发现实体。

该模型由三部分组成，卷积字符特征提取、基于fastText、GloVe的词语特征提取和基于InceptionV3的图像特征提取。

多模态命名实体识别模型（2）——词语特征提取

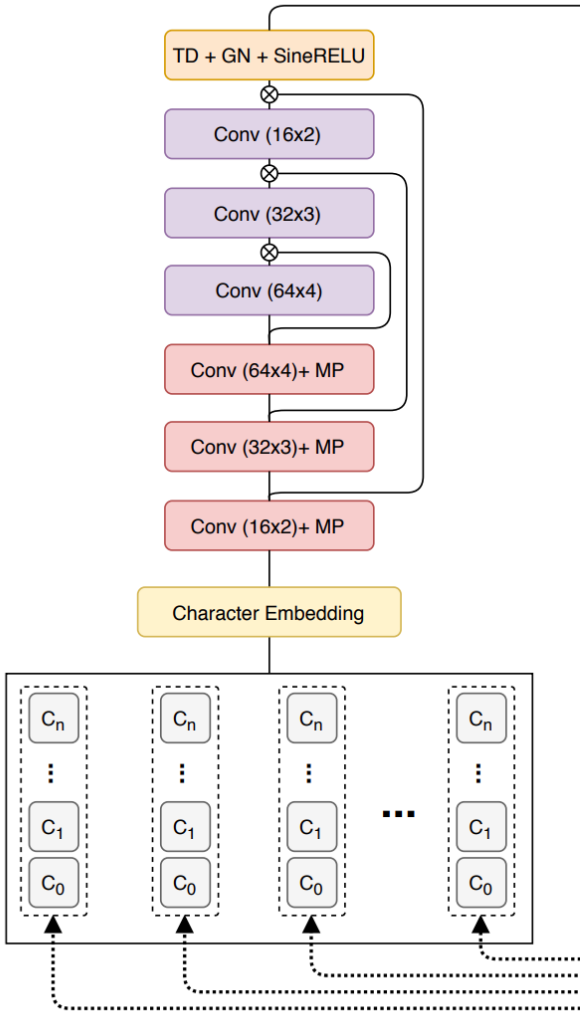


使用GloVe4和fastText5的预训练模型，将单词向量联合嵌入，可实现500维单词嵌入。

使用了双向长短时记忆，获得每个隐藏层的前向和后向信息。

当Glove失效时，FastText提供了更好的嵌入，它能够使用子词嵌入捕获语义。

多模态命名实体识别模型（3）——字符特征提取



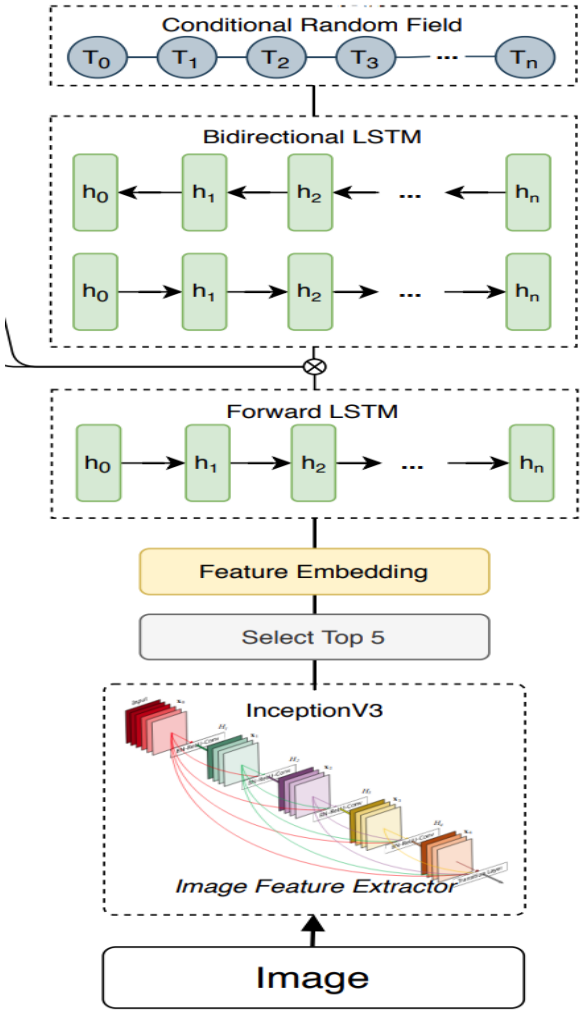
使用GloVe及fastText可以得到大多数单词的向量表示，但仍会遇到在字典中不存在的单词。研究单词中的字符构成可以来寻找单词的数字表示，因此需要用字符嵌入解决“字典中不存在的单词”的问题。

来自单条推特的每个单词序列[w1，w2，…，w]被转换为字符表示序列[[c（0,0），c（0,1），…，c（0,k）]，，[c（n，0），c（n，1），…，c（n，k）]。

首先是第一部分的三个卷积层，在每一层中，内核大小从2递增到4（每次增长1），而内核数量从16开始增加一倍直到64。在前三个卷积层之后均有一个一维池化层。第二部分与第一部分类似，但略有更改。内核大小从4个减少到2个，内核数量从64个依次减半。

最后通过 target dropout、组群归一化和SineRelu函数。防止过度拟合。

多模态命名实体识别模型（4）——图像特征提取



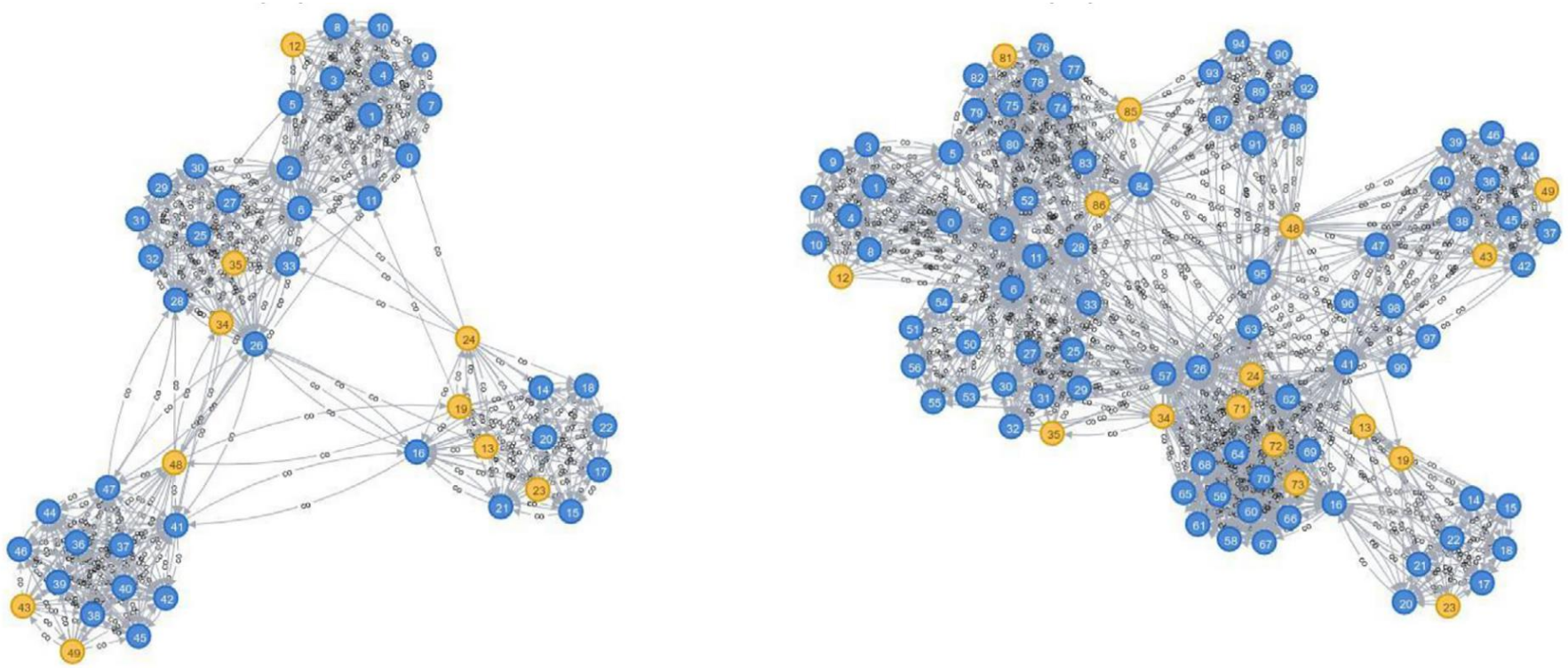
对于图像特征提取模块，使用了在ImageNet上预训练的InceptionV3网络，使用了完整1000余类ImageNet 数据集中数据来提取更好的特征。

另外选择提取前5个概率最高的特征表示与图像相关的文本关键词，这五个词的组合提供有关图像中各个对象的有用信息。

长短时记忆单元用于输出最终图像特征。最后，将单词和图像特征提取器中的LSTM单元进行叠加。

条件随机场为最后一层，形成最终输出。

基于记忆的图模型（1）



将多模态命名实体识别模型、图信息、文本信息输入，多模态命名实体识别模型通过图信息以及文本信息识别文本信息中的实体（上图黄色结点）。按照时间顺序以及更新图策略来更新图。

基于记忆的图模型（2）——图更新策略

四种插入情形

Table 2
Categorization of incoming tweets as subgraphs.

Category	Symbol	Description
Unique	$\mathcal{U}$	*All* words in tweet are **new** and did not appear in any θ_i^t
Incessant	$\mathcal{I}$	*All* words in tweet are **previously merged into a single** θ_i^t
Multiple	$\mathcal{M}$	*All* words in tweet are **previously merged into more than one** θ_i^t
Subset	$\mathcal{S}$	*Some* words in tweet are **previously merged into one or more that one** θ_i^t and *some* are **new**

遗忘曲线

$$R = e^{-\frac{t}{\varrho}}$$

作为遗忘的超参数，使用了一种动态方法来调查有多少内存可用，并据此找到新的 ϱ 。如果内存使用率超过阈值（90%），开始将 ϱ 减少1，直到在自动消除节点后达到正确的RAM使用率。如果低于阈值（80%），ϱ 则增加1，直到达到所需值。初始值设置为103。

基于图的话题发现（1）

对于结点W的评分： $\Gamma^t(W)$：

$$S_{i,j}^t = S_{i,j}^{t-1} + \cos(M_i^t, M_j^t), \text{ if } W_i \text{ and } W_j \in tweet_t \quad (3)$$

$$M_i^t = \alpha M_i^{t-1} + (1-\alpha) M_i^{tweet_t}, \text{ if } W_i \in tweet_t \quad (4)$$

$$\Gamma^t(W) = E \times \left(\Upsilon^t(W) + \sum_{W' \in N^t(W)} \Upsilon^t(W') \times S_{W,W'}^t \right) \quad (6)$$

$$\Upsilon^t(W) = \delta F^t \times \left(F^t(W) + \sum_{i=0}^{t} (L_i + R_i) \right) \quad (7)$$

$$E = \begin{cases} 1.2, & \text{if } W \text{ is tagged as named entity} \\ 1, & \text{otherwise} \end{cases} \quad (8)$$

转换为概率

$$P(W|\theta_i^t) = \frac{\Gamma^t(W)}{\sum_{W' \in \theta_i^t} \Gamma^t(W')}$$

$$P(\theta_i^t|\theta^t) = \frac{\sum_{W \in \theta_i^t} \Gamma^t(W)}{\sum_{W \in \theta^t} \Gamma^t(W)}$$

$$P(W|\theta^t) = P(W|\theta_i^t) \times P(\theta_i^t|\theta^t)$$

基于图的话题发现（2）

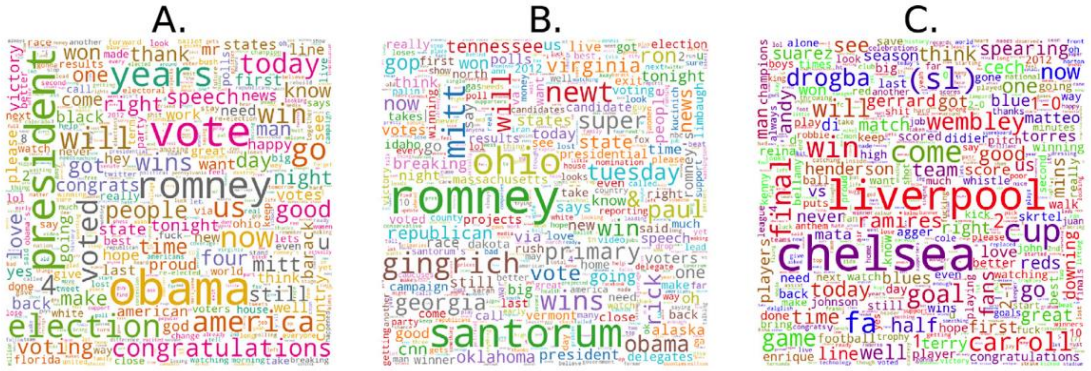


Fig. 5. WordCloud representation of FA Cup dataset.

北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

6

demo展示

讲解：严睿逸

TDT话题检测的应用场合

与一般的信息检索或者信息过滤不同，TDT所关心的话题不是一个大的领域（如美国的对华政策）或者某一类事件（如恐怖活动），而是一个很具体的“事件（Event）”TDT是一项综合的技术，需要比较多的自然语言处理理论和技术作为支撑，因此这些测评对其进行了细化。根据不同的应用需求，TDT评测会议把话题检测和跟踪分成五个子任务。

- 报道切分（Story Segmentation）
- 话题跟踪（Story Tracking）
- 话题检测（Story Detection）
- 首次报道检测（New Event Detection）
- 关联检测（Link Detection）

TDT话题检测的应用场合

检测场景与目标

场景：微博新闻评论区

新闻条目：K条新闻

话题数：$M(M \leq K)$类话题

方法与策略

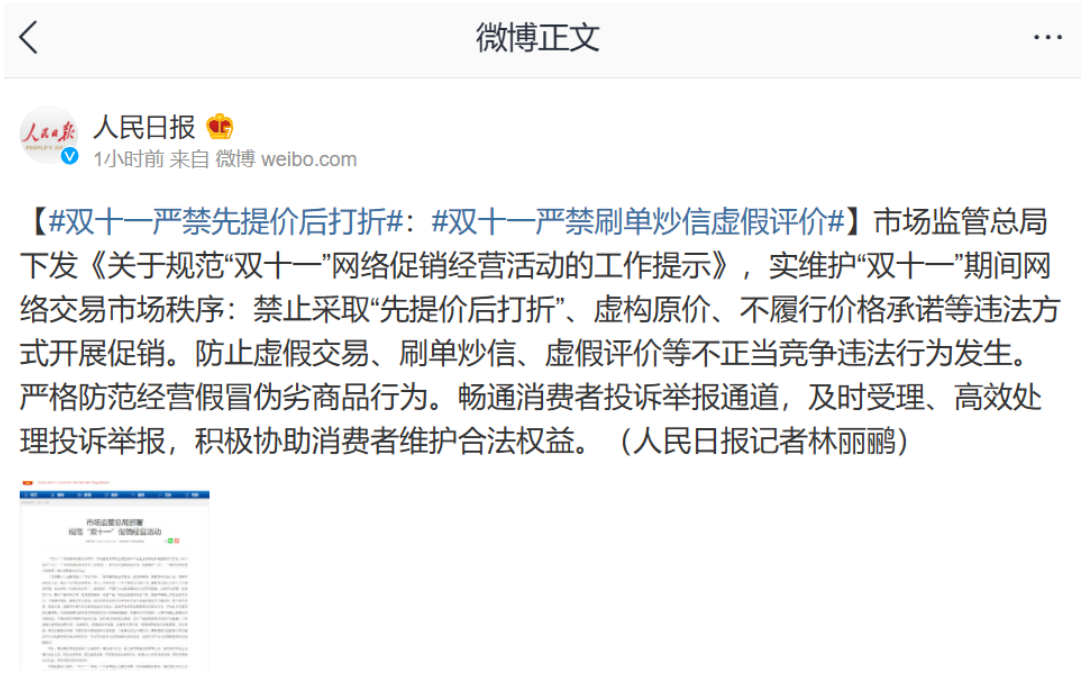
预训练方法 $\begin{cases} BOW \\ TF-IDF \end{cases}$

主题模型：LDA

2 爬取微博评论信息

本样例中的爬取目标选取的是（PC访问）微博手机端http://m.weibo.cn某一微博链接的评论区内容，对评论区内容逐条统计并加入到待分析的数据集。

在爬取内容之前需要通过Fiddler抓包软件对该页面的headers信息进行提取，以便于无差错、不遗漏地得到网页json数据内容，并进行按照评论条目顺序的分割处理。



爬取微博评论信息

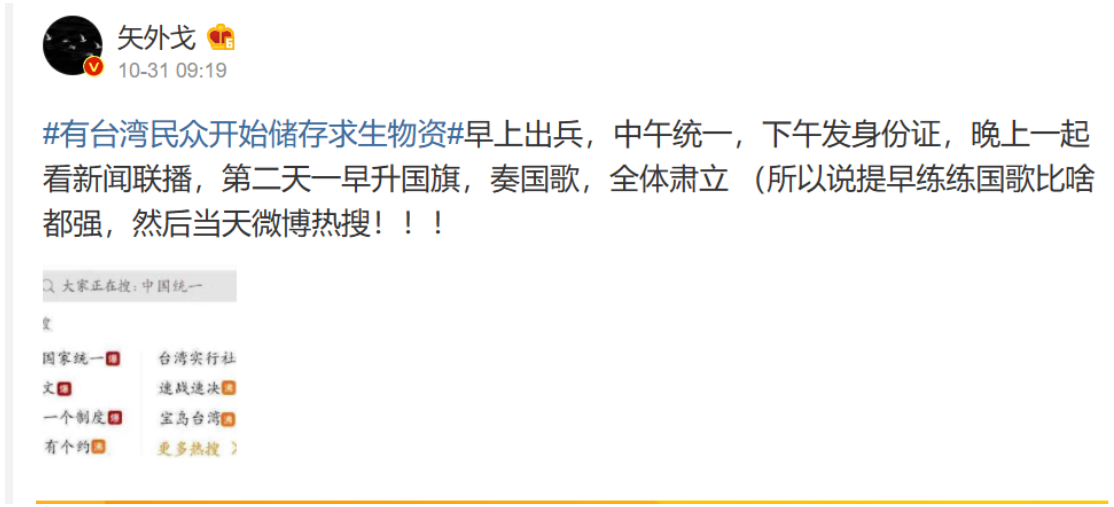
“人民日报：双十一严禁先提价后打折”评论区的爬虫运行样例如下：

```
网友馨漪已上线 狠狠支持了 Sat Nov 06 08:51:25 +0800 2021
最棒的小臻臻ing 折后比10月更贵了 Sat Nov 06 08:53:03 +0800 2021
天涯凌希1031 就是先提价后打折 Sat Nov 06 08:52:00 +0800 2021
dadudaduda 说的是斐乐吗 Sat Nov 06 09:56:25 +0800 2021
木易景三 我11.1号前买的衣服69两件，准备活动再买一件，发现11.1号卖69一件，卖家说之前活动买一送一，现在活动结束了，这算涨价吗？ Sat Nov 06 08:54:50 +0800 2021
桃花岛有桃花 现在的双十一不就是纯粹的割韭菜吗 Sat Nov 06 09:43:12 +0800 2021
```

```
乖兔乖玉 晒单送赠品是什么行为 Sat Nov 06 10:04:13 +0800 2021
三A王 能管管淘宝吗？真的是 Sat Nov 06 09:43:13 +0800 2021
Rocky2011姐姐 今年双十一没啥意思了，没有多大优惠，包括去年的款式，今年都比去年贵 Sat Nov 06 09:39:42 +0800 2021
人不炒股枉中年 用我们做企业的做法，只喊口号不实际行动，也没有实际成效，早被骂的狗血淋头了 Sat Nov 06 09:38:29 +0800 2021
素麻米分啦 我都买好了您才说[开学季] Sat Nov 06 09:38:23 +0800 2021
人不炒股枉中年 年年提，有用吗 Sat Nov 06 09:34:41 +0800 2021
```

其他微博资讯

除了“人民日报：双十一严禁先提价后打折”评论区，本样例对其他微博资讯进行爬取并拼接，以便于后续话题检测功能以及关联检测功能（事实上，不同的资讯谈的是一个话题）的测试。



3 其他微博资讯

央视新闻
16小时前 来自 微博 weibo.com 已编辑

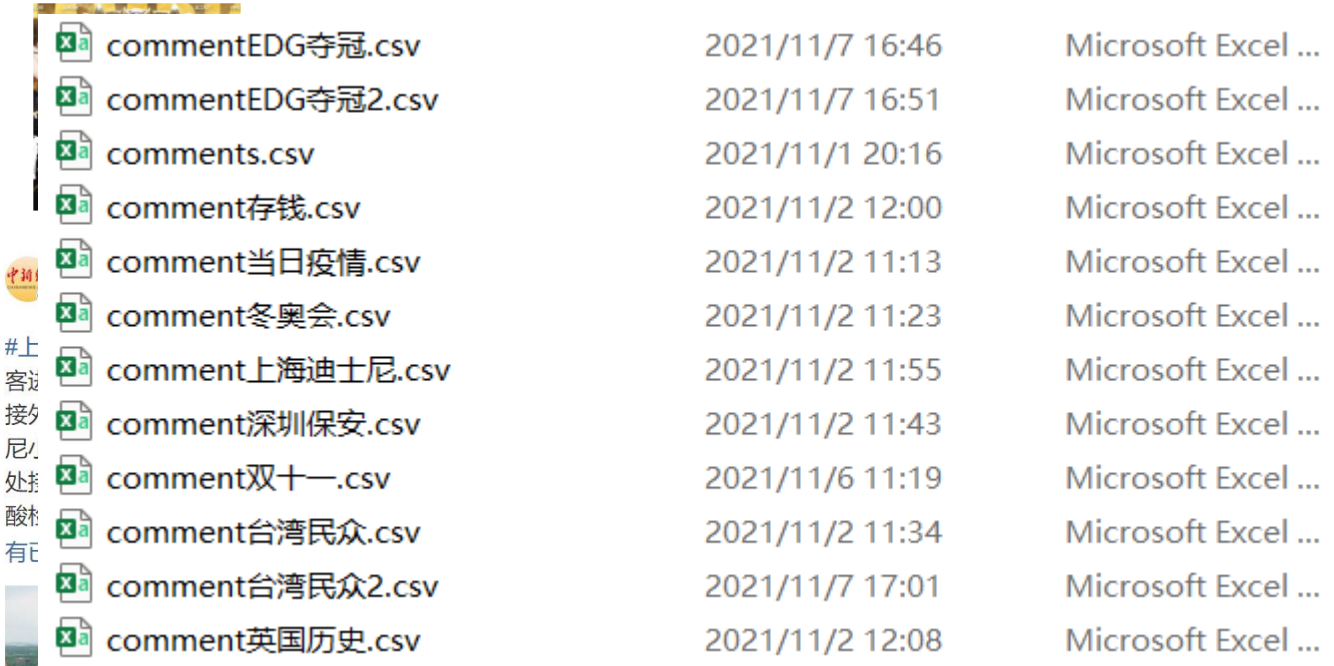
【祝贺！#EDG夺冠#】刚刚，英雄联盟S11总决赛，中国LPL赛区战队@EDG电子竞技俱乐部 以3：2战胜韩国LCK赛区战队DK，获得2021年英雄联盟全球总决赛冠军！恭喜！

央视新闻
10-27 13:08 来自 微博视频号

【#冬梦飞扬MV# 转起！分享这首歌，我们北京见！】“圣火又一次点亮古城夜空，相约北京这激情的城。冰雪中我们相知相融，全世界共赴一场冬天的梦！”#北京冬奥会倒计时100天#，听听这首《冬梦飞扬》↓为出发呐喊，将青春燃动，一起为中国冰雪加油！#北京2022年冬奥会#！@UNIQ-王一博 央视新闻的微博视频

EDG电子竞技俱乐部
16小时前 来自 EDG超话

我们是冠军！！！！！！！！！！！！！！！！！！！



北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

7个评论区拼接

↓

jieba分词系统

↓

停词表

↓

文本特

↓

LDA话题模型

↓

发现话题

上述7个微博资讯事实上只有5个话题种类（其中EDG夺冠新闻2条，台湾民众存储求生物资2条）。在18000多条评论内容中，该测试希望通过整个系统来将话题分为5个类别。

```
comments1 = pd.read_csv(r'clean_dataset\commentEDG夺冠.csv', encoding='utf-8',index_col=0)
comments2 = pd.read_csv(r'clean_dataset\commentEDG夺冠2.csv', encoding='utf-8',index_col=0)
comments3 = pd.read_csv(r'clean_dataset\comment台湾民众.csv', encoding='utf-8',index_col=0)
comments4 = pd.read_csv(r'clean_dataset\comment台湾民众2.csv', encoding='utf-8',index_col=0)
comments5 = pd.read_csv(r'clean_dataset\comment冬奥会.csv', encoding='utf-8',index_col=0)
comments6 = pd.read_csv(r'clean_dataset\comment上海迪士尼.csv', encoding='utf-8',index_col=0)
comments7 = pd.read_csv(r'clean_dataset\comment双十一.csv', encoding='utf-8',index_col=0)
alldata = pd.concat((comments1,comments2,comments3,comments4,comments5,comments6,comments7) ,igno
```

...	
18615	应该把双十一禁了，618，双十二，
18616	噢？南极人就是噢，双十一内裤加入购物车是39.9的，双十一的时候去49.5
18617	你知道谁是刷的？某猫超市和N只松鼠也刷单怎么不管？
18618	转发微博
18619	仅仅双十一吗

发现话题

↓

阈值设置和词数限制

↓

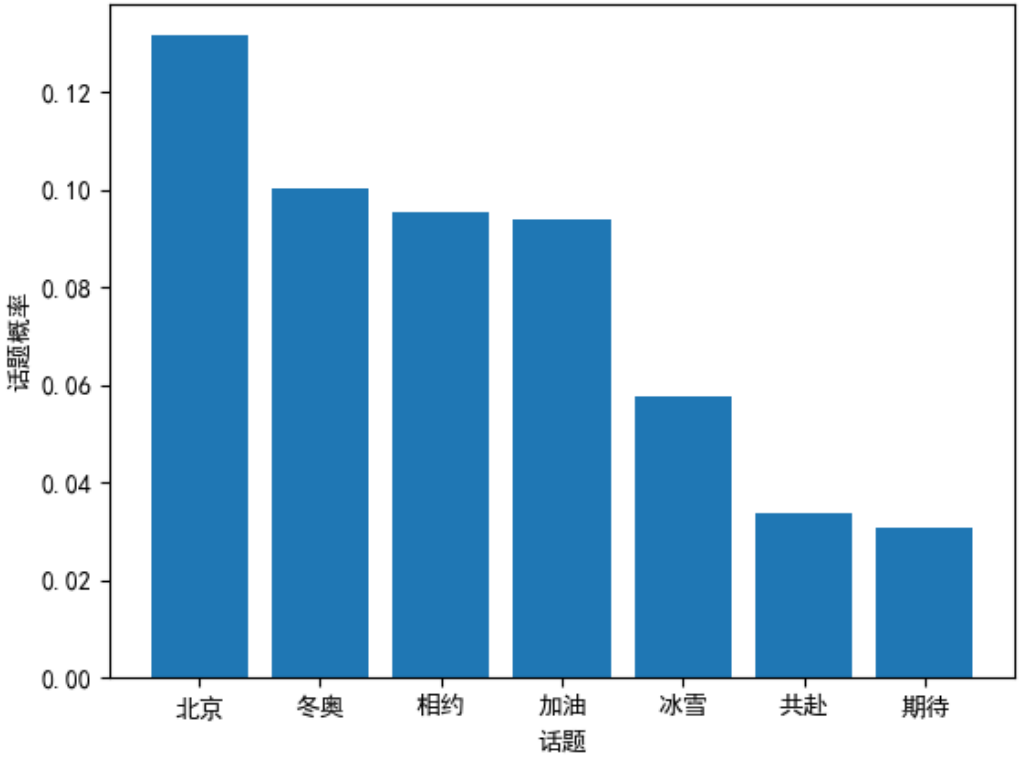
话题内容以及话题概率

↓

计算困惑度

下图是通过词袋法预处理之后的其中一个话题（显然是冬奥会话题）的话题检测结果，可见该话题的主要词条为“北京，冬奥，相约”等等。

冬奥会话题



5 实验组评价

实验效果通过困惑

但作为评价LDA模型

式中$p(w)$表示每

指标的变量集中在指

$p(w)$值越小，即熵越

```
def print_top_words(model, feature_names, n_top_words):
    P=0
    W=0
    for topic_idx, topic in enumerate(model.components_):
        print("Topic #%d:" % topic_idx)
        #topic=standardize(topic)
        #(topic)
        #print(feature_names)
        plot_x=[]
        plot_y=[]
        W+=len(topic)
        topic/=sum(topic)
        for i in topic.argsort()[::-1]:
            P += math.log(topic[i])
            if  feature_names[i]!='nan' and feature_names[i]!='ahref' and len(plot
                print(feature_names[i],topic[i], end=' ')
                plot_x.append(feature_names[i])
                plot_y.append(topic[i])

        plt.bar(plot_x,plot_y)
        plt.xlabel('话题')
        plt.ylabel('话题概率')
        plt.savefig("./topic_image/Topic #%d.png"% topic_idx)
        plt.close()
        #print(" ".join([feature_names[i]for i in topic.argsort()[:-n_top_words -
        #print(" ".join([topic[i] for i in topic.argsort()[:-n_top_words - 1:-1]])
        print()
    print(P,W)
    perplexity=math.exp(-P/W)
    print('困惑度: ',perplexity)
```

实验组评价指标设计

	BOW词袋模型	TD-IDF
#实验1	1203.37	828.26
#实验2	1440.89	836.16
#实验3	1406.21	876.16
#实验4	1154.56	884.92
#实验5	1277.16	872.91
困惑度$perlexity$平均值	1296.44	859.68

可见通过TD-IDF方法预训练之后的话题分布更集中，熵更小，模型更适应该语料库。

直观的话题分类结果

	BOW词袋模型	TD-IDF
#话题1	北京 冬奥 加油 相约 期待	加油 冬奥 北京 相约 希望
#话题2	恭喜 中国 edg 王一博 疫情	恭喜 冠军 edg 苦涩 消费者
#话题3	双十 迪士尼 上海 确诊 消费者	管管 乐吧 欺骗 双十 消费者
#话题4	台湾 冠军 统一 中国 祖国	台湾 身份证 统一 战争 祖国
#话题5	疫情 上海 战争 迪士尼 确诊	迪士尼 确诊 管管 支持 开学

从受到主题干扰的程度（红色标注）可以看出与困惑度相对应的模型优劣程度。

德以明理　　学以精工