LDA

读书笔记

一、基本介绍

1.1 什么是主题模型？

- 所有主题模型都基于这样的基本假设：
  1. 每个文档包含多个主题
  2. 每个主题包含多个词语

- 文档的语义实际上是由一些我们所忽视的隐变量来管理的。因此，主题模型的主要目标就是解释这些潜在变量——**主题**。
- 因此，主题模型是文本挖掘领域中的一种技术。主题模型具有优秀的降维能力，能够从一个文本对象中自动识别它的主题，利用挖掘出的主题能帮助人们理解海量文本背后隐藏的语义。不同于文本摘要技术，主题模型更多聚焦在文本的主题和概念，不仅仅是文本摘要。主题建模可用于：文本分类、话题检测、文本自动摘要、关联判断等文本挖掘任务。

1.2 什么是LDA？

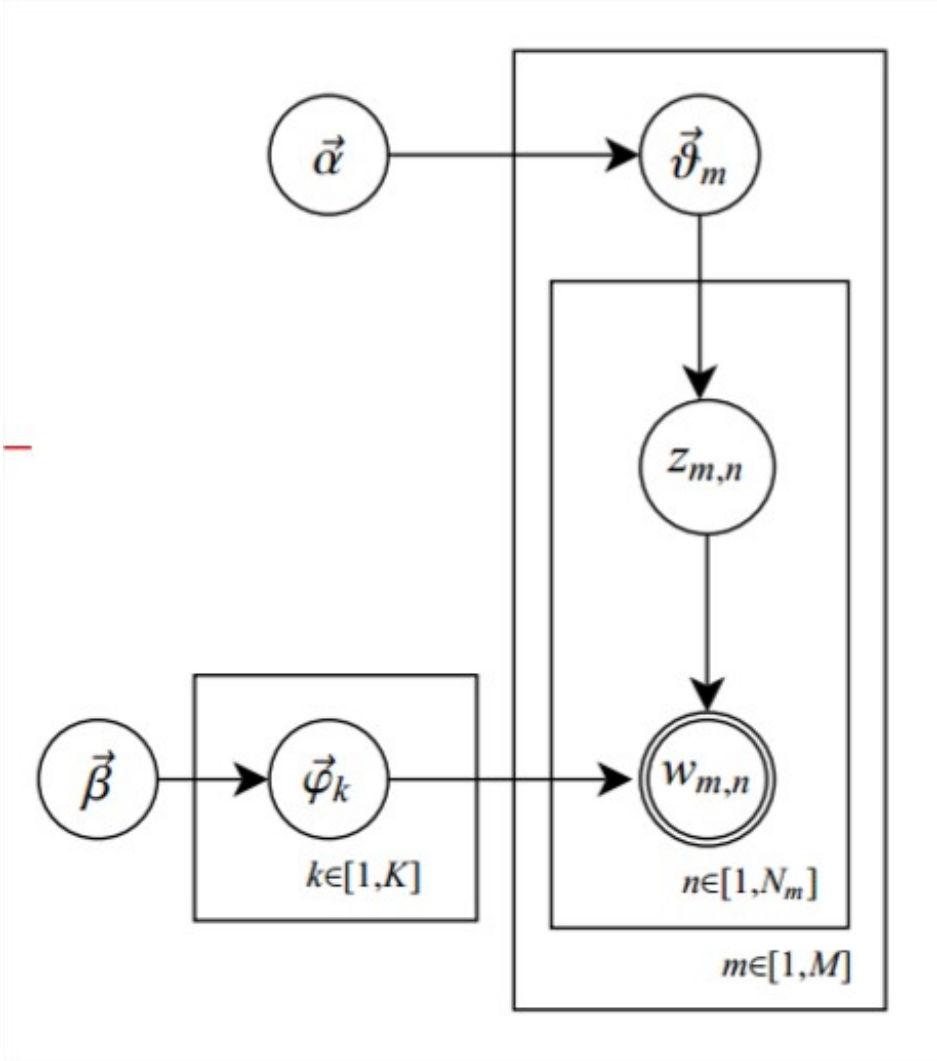
- LDA是一种文档主题生成模型，在上文提到的主题模型基本假设之上，它认为一篇文章的每个词都是以一定概率选择了某个主题，这个主题以一定概率选择了某个词语。
- LDA也称为一个三层贝叶斯概率模型，包含文档、主题和词语三层结构。每篇文档能够由若干主题的概率分布所表示，每个主题也可由若干词语的概率分布所表示。主题分布和词分布都是多项分布且服从Dirichlet分布，主题分布的Dirichlet分布参数为 α，词分布的Dirichlet分布参数为 β。
- LDA：Latent Dirichlet Allocation

> **Dirichlet分布**：二项分布的共轭先验分布是Beta分布，多项分布的共轭先验分布是Dirichlet分布。
>
> > **A. 共轭先验**：如果 $p(\theta|x)$ 和 $p(\theta)$ 满足同样的分布律，那么先验分布和后验分布就叫共轭分布，先验分布叫似然函数的共轭先验分布。
> >
> > - $p(\theta|x)=\frac{p(x|\theta)p(\theta)}{p(x)}\approx p(x|\theta)p(\theta)$
> >
> > 1. 去分母：如果不求 $p(\theta|x)$ 具体值，只看 θ 取何值时，$p(\theta|x)$ 最大，可省去分母。
> > 2. $p(\theta|x)$：后验分布
> > 3. $p(x|\theta)$：似然概率。给定某个参数情况下，x的概率分布。
> > 4. $p(\theta)$ ：先验概率。未看到样本x之前的分布。
> >
> > **B. 二项分布**：$P(x|\theta)=\theta^x\cdot(1-\theta)^{1-x}$
> >
> > **C. Beta分布**：
> > 1. 概率密度：$f(x)=\frac{1}{B(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1}$ ， 其中
> > $x\in[0,1],B(\alpha,\beta)=\int_0^1x^{\alpha-1}(1-x)^{\beta-1}dx$
> > 2. 期望：$E(x)=\int_0^1xf(x)dx=\ldots=\frac{\alpha}{\alpha+\beta}$
> >
> > **D. Dirichlet分布**
> > 1. 概率密度：$f(\vec{p}|\vec{\alpha})=\frac{1}{\Delta(\vec{\alpha})}\prod_1^Kp_k^{\alpha_k-1}$ ，其中
> > $p_k\in[0,1]$, $\Delta(\vec{\alpha})=\frac{\prod_1^K\Gamma(\alpha_k)}{\Gamma(\sum_1^K\alpha_k)}$
> > 2. 期望：$E(p_i)=\frac{\alpha_i}{\sum\alpha_k}$

> ps: 对比Beta分布的概率密度和期望。

1.3 LDA 实现步骤

符号标记	含义
D	文档集合，$D=\{d_1,d_2,\ldots,d_m\}$
m	文档个数
N_m	第N_i篇文档有多少个词
V	词典，所有词的集合
v	词的个数
K	每篇文档的主题个数



1. $\vec{\alpha} \rightarrow \vec{\theta}_m$：采样。分别采出 m 篇文档的主题分布 $(p_1, p_2, \ldots, p_K)$。(K：主题个数；服从参数为 α 的Dirichlet分布)
2. $\vec{\beta} \rightarrow \vec{\phi}_k$：采样。分别采出 k 个主题的词分布 $(p_1, p_2, \ldots, p_v)$。（v：词的个数；服从参数为 β 的Dirichlet分布）
3. $\vec{\theta}_m \rightarrow z_{m,n}$：采主题。在第 m 篇文档中遍历，分别得到第 n 个词的主题编号。
4. $w_{m,n}$：根据S3的主题编号，从S2的词分布中找到对应的主题并采样出一个词。
5. 不断重复随机生成过程，直到 m 篇文档全部遍历。

二、代码实现

2.1 lda包

2.2 Gensim

1. 是什么？

2. 怎么用？

步骤？？