#MLorDS

# 第三部分：你好，阳光！——隐含狄利克雷分配模型

Sunshine 公司开发了一系列产品——这包括婴儿奶嘴、吹风机和微波炉。经过一段时间后，公司发现婴儿奶嘴销售很好，但是微波炉和吹风机却遭遇了巨大的失败。公司负责人很不甘心，他爬取了用户对产品的评分星级和评论内容，试图从中找到挽回的可能。如何从浩如烟海的用户评论中找出核心的词汇、挑选出最具代表性的主题呢？——LDA 就是一种可行的方法。

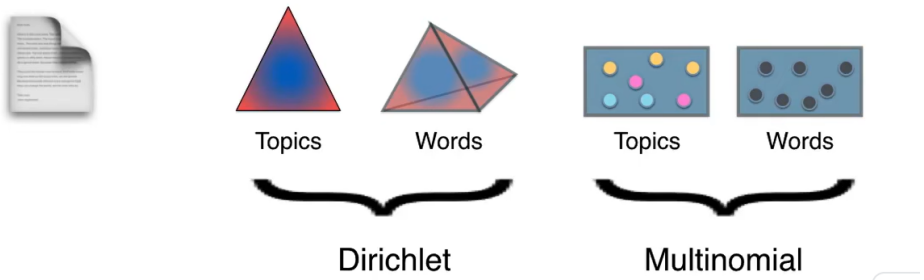
*本章的标题和导入文字均改编自 2020 MCM Problem C*

## What is LDA?——Dirichlet Distributions

LDA 是一个生成模型，我们实际是希望调整 LDA 中的参数，使得由模型生成的文章和真实的文章最为接近。

那么这个概率是由四项连乘式得到：四项连乘式中有两个 Dirichlet分布和两个多项分布，形象地来说，这四项分别代表着制造两个骰子和投掷两个骰子的过程。

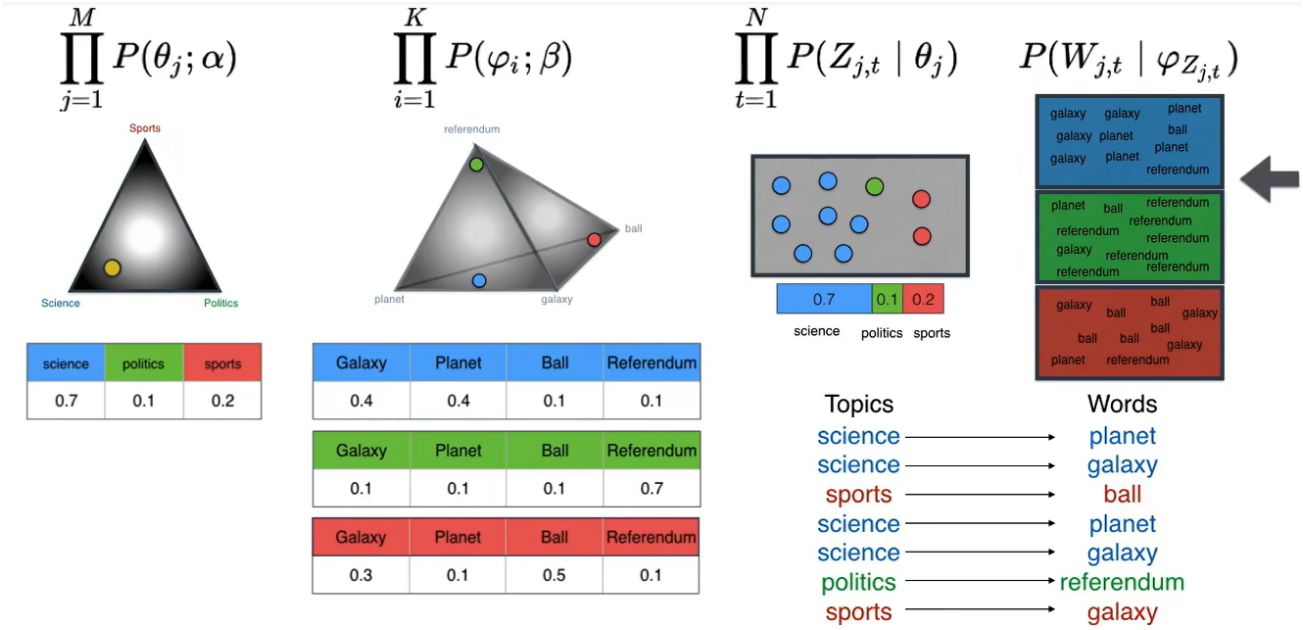
$$P(\boldsymbol{W}, \boldsymbol{Z}, \boldsymbol{\theta}, \boldsymbol{\varphi}; \alpha, \beta) = \prod_{j=1}^{M} P(\theta_j; \alpha) \prod_{i=1}^{K} P(\varphi_i; \beta) \prod_{t=1}^{N} P(Z_{j,t} \mid \theta_j)\, P(W_{j,t} \mid \varphi_{Z_{j,t}})$$




我们分别来解说这四个分布：

- 第一个 Dirichlet 分布——描述 Doc-Topics 的关系：将文章和主题联系起来。根据给定的参数 $\alpha$，这个 Dirichlet分布生成一个向量 $\theta_j$，向量中的每一个分量都是一个主题出现的概率。例如，生成的 $\theta_j = [0.7, 0.1, 0.2]$ 意味着科学、政治、体育出现的概率是 0.7，0.2，0.1。形象地来说，这一步中，我们由 Dirichlet 分布制造了一个三面骰，每个面都是一个主题，掷出三个面的概率分别就是 0.7,0.2,0.1。上面方程中的 $P(\theta_j; \alpha)$ 代表着在 $\alpha$ 参数下，Dirichlet 分布给出主题概率向量 $\theta_j$ 的概率。
- 第一个多项分布——用于确定 Topics。在这一步中，我们抛掷 $t$ 次刚才制造的骰子，从而得到 $t$ 个不同的主题，第 $t$ 次得到的主题被记作 $Z_{j,t}$。上面方程中，由概率向量 $\theta_j$ 生成主题 $Z_{j,t}$ 的概率被记作 $P(Z_{j,t}|\theta_j)$
- 第二个 Dirichlet 分布——描述 Topics-Words 的关系：将之前生成的一系列主题和单词联系起来。根据给定的参数 $\beta$，这个 Dirichlet 分布针对可能取到的 $K$ 个主题生成 $K$ 个向量 $\phi_i$，向量中的每一个分量都是指定主题下每一个单词出现的概率。形象地来说，这相当于又制造了一个多面骰，每个面都是一个单词。上面方程中，在 $\beta$ 参数下，对第 $i$ 个主题生成出单词概率向量 $\phi_i$ 的概率记为 $P(\phi_i; \beta)$

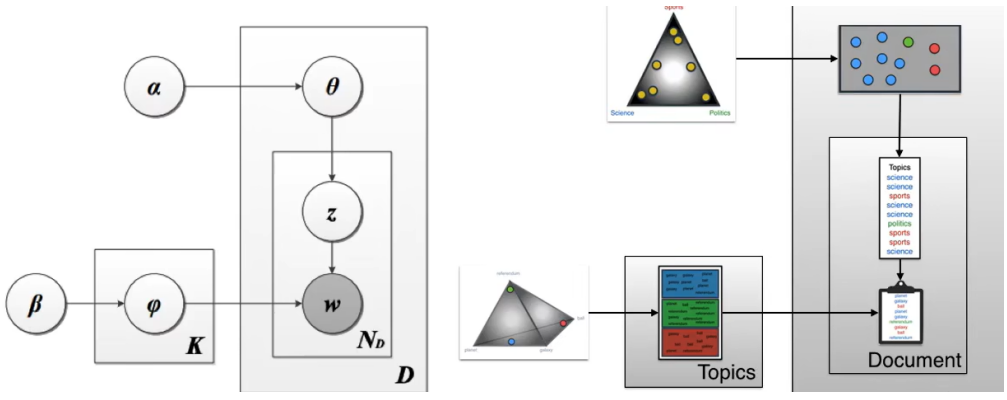
- 第二个多项分布——用于确定 Words。对第 $t$ 次得到的主题 $Z_{j,t}$，投掷一次单词的“骰子”，生成出一个单词。在概率向量 $\phi_{Z_{j,t}}$ 下生成单词 $W_{j,t}$ 的概率被记为 $P(W_{j,t}|\phi_{Z_{j,t}})$




另外，提请注意：你会发现如果按照上面的表述来写式子，和最初的方程是对不上的，这是因为，上面的四步只是按照逻辑顺序写的，但是在实际生成文章的时候，步骤和上面不太一样

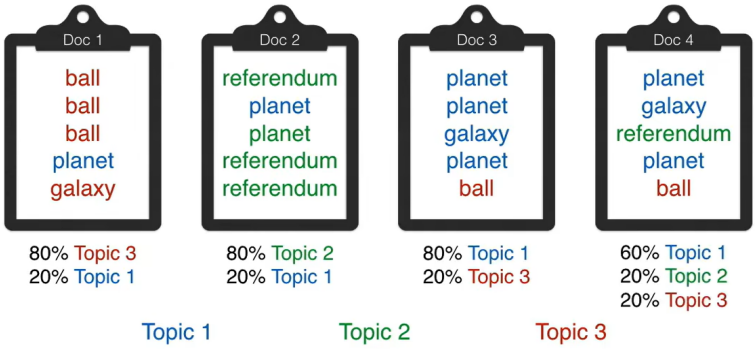
我们说一下这个概率的方程是怎么写出的，也就是说，我们要求解一个在参数 $\alpha, \beta$ 下生成了一系列主题概率分布 $\theta$、一系列单词概率分布 $\phi$、一系列主题 $Z$ 和一系列单词 $W$ 的概率。
首先，对于指定的文档 $j$，我们进入准备阶段：在准备阶段中，我们首先制造第一枚骰子——生成文章主题的概率分布 $\theta_j$，这会导致概率乘上一项 $P(\theta_j;\alpha)$；然后，对于所有可能的 $K$ 个主题，先把每个主题对应的单词概率分布生成好，这会导致概率乘上一项连乘式 $\prod_{i=1}^{K} P(\phi_i;\beta)$，至此，准备阶段结束。
下面，我们进入文章生成阶段：一篇文章中共需要生成 $N$ 个单词，对第 $t$ 个 单词首先生成主题 $Z_{j,t}$，再根据主题生成单词 $W_{j,t}$，这会导致概率乘上一项连乘式 $\prod_{t=1}^{N} P(Z_{j,t}|\theta_j)P(W_{j,t}|\phi_{Z_{j,t}})$
最后，由于一共有 $M$ 篇文章，所以在概率的最前面还要加一个连乘符号 $\prod_{j=1}^{M}$
我们写下的表达式就是：

$$P(\theta,\phi,Z,W;\alpha,\beta) = \prod_{j=1}^{M} P(\theta_j;\alpha) \prod_{i=1}^{K} P(\phi_i;\beta) \prod_{t=1}^{N} P(Z_{j,t}|\theta_j)P(W_{j,t}|\phi_{Z_{j,t}})$$




## How to train LDA?—— Gibbs Sampling

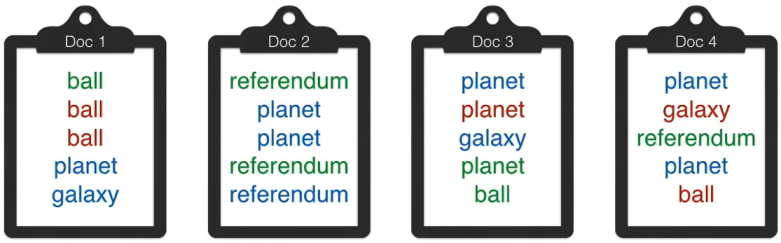
如图，对于给定的几篇文章，我们可以指定“**每个单词属于哪个主题**”，这样，我们就可以统计每篇文章中，属于每个主题的单词所占的比例：



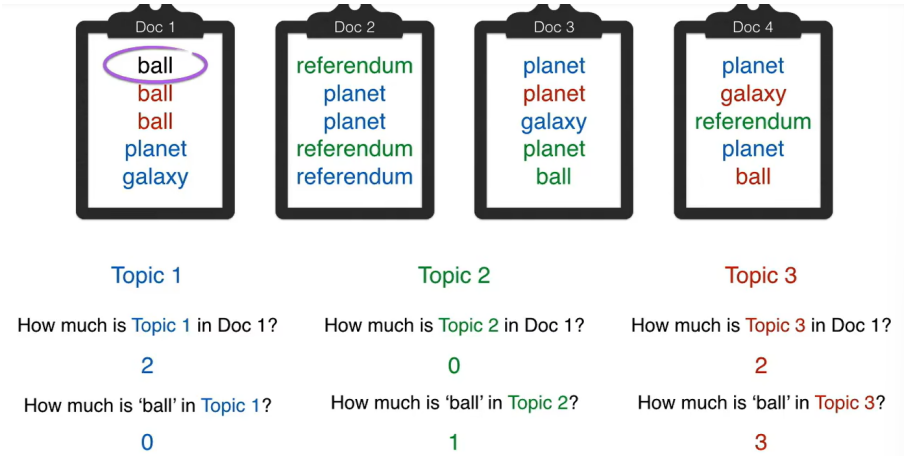

我们之前提到过，LDA 是一个生成模型，其任务是调节两个 Dirichlet 分布的参数，使得生成指定文章的概率尽可能地高。然而，这个优化目标可能太难做到，我们可以将目标变得简单一些，以使得计算机容易处理：根据常识，我们知道，一种很好的单词-主题关系应该服从以下两个条件：

- 每篇文章尽量只具有一个主题
- 每个单词尽量属于唯一的主题
  接下来，我们将使用 Gibbs 采样来达到这两个目标。

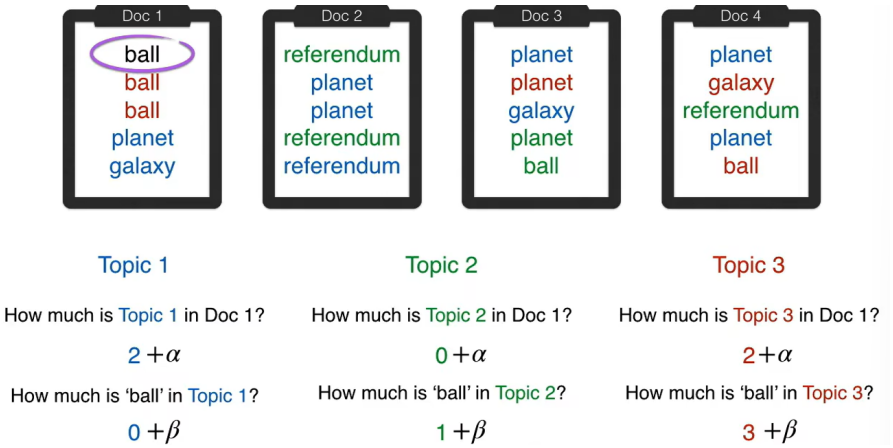
**Gibbs 采样的基本思想是：每次随机选择一个单词，为其选择一个“相对正确”的主题**
例如，我们先将所有单词随机地选择主题：




- 第一次，我们选择文档 1 中的第一个单词"ball"，并观察其他所有"ball"的主题——显然，其他"ball"中，红色（主题 3）的最多。此外，我们还要观察文档 1 中所有单词的主题——蓝色（主题 1）有两个词，红色（主题 3）也有两个词。结果是这样的：




- 那么我们怎样决定将这个"ball"赋予何种主题呢？答案是将两次统计得到的数量相乘，作为一个主题的"得分"，选择乘积最大的那个主题作为单词"ball"的主题。为什么是相乘？因为这个统计的数量实际上是**概率**，例如，$2 \times 3$ 代表着 "文档 1 出现主题 3 的概率"和“ball 这个词在主题 3 中的概率”相乘。具体相乘得到的是什么含义，我们暂时不追究，这涉及到对 Gibbs 采样方法的严格定义

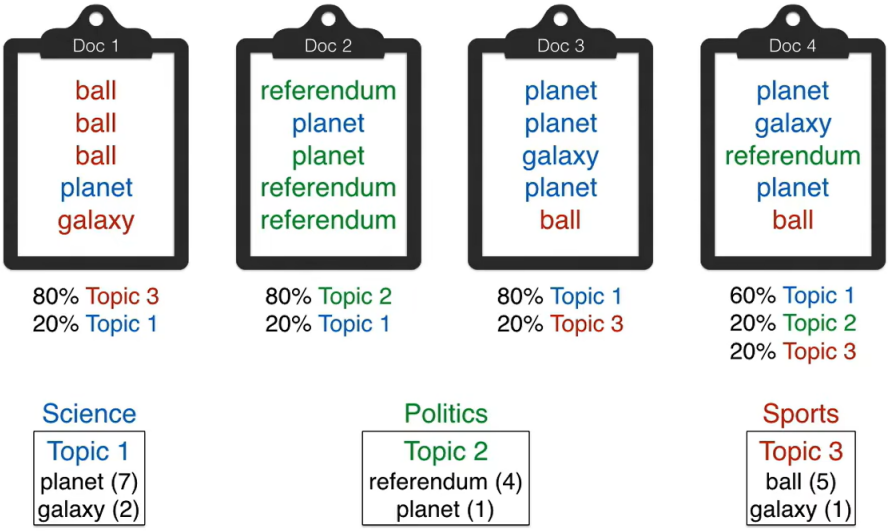



在实际的使用中，我们会对上面的方法进行一些改进，这些改进包括

- 将两次统计的数值加上一个很小的数据 $\alpha$ 和 $\beta$，这防止了“0”的出现
- 依照归一化的乘积作为给单词选择主题的概率，而不是直接选择乘积最大的主题

循环迭代很多次，直至每个单词所对应的主题不再有大的变化，我们的算法就结束

当然，最后，我们还需要得到每个文档的主题。我们只需要统计每个文档中属于每个主题的单词所占的比例即可。模型将输出每个文档属于第 $i$ 个主题，具体对模型的解释需要我们自行进行。




## 写在最后：LDA 的数学基础

在这一部分中，我们将解释 Dirichlet 分布、多项式分布、Gibbs 采样，以及 LDA 的概率表达式和训练之间到底有什么关系。