



## 文本分析

---

# 目录

- 第一部分  简介
- 第二部分  文本表示模型
- 第三部分  文本分类与聚类
- 第四部分  主题分析
- 第五部分  情感分析

# 目录

第一部分  简介

第二部分  文本表示模型

第三部分  文本分类与聚类

第四部分  主题分析

第五部分  情感分析

# 文本数据的普遍性

- 与表格型数据相比，文本数据占比越来越高
  - 论坛、新闻、博客、微博、微信
  - 商品评论、投诉文本
  - 电子邮件
  - 医学诊疗记录
  - 调查问卷
  - 法院判决书

商品评论

手机很好，很漂亮，速度也非常快，只是我的手机屏幕上有一小块刮痕，不过平时也不太会注意到，就算了，懒得再申请换货了。

手机不错的，除了一个屎黄色的手机套我不喜欢，其他一切都很完美；用了几天，感觉挺好的

医学诊疗数据

眼部：眼睑无水肿  
脸结合膜未见出血点  
巩膜无黄染 角膜透明 瞳孔等大等圆  
直径3~4mm 对光反射、集合反射存在。

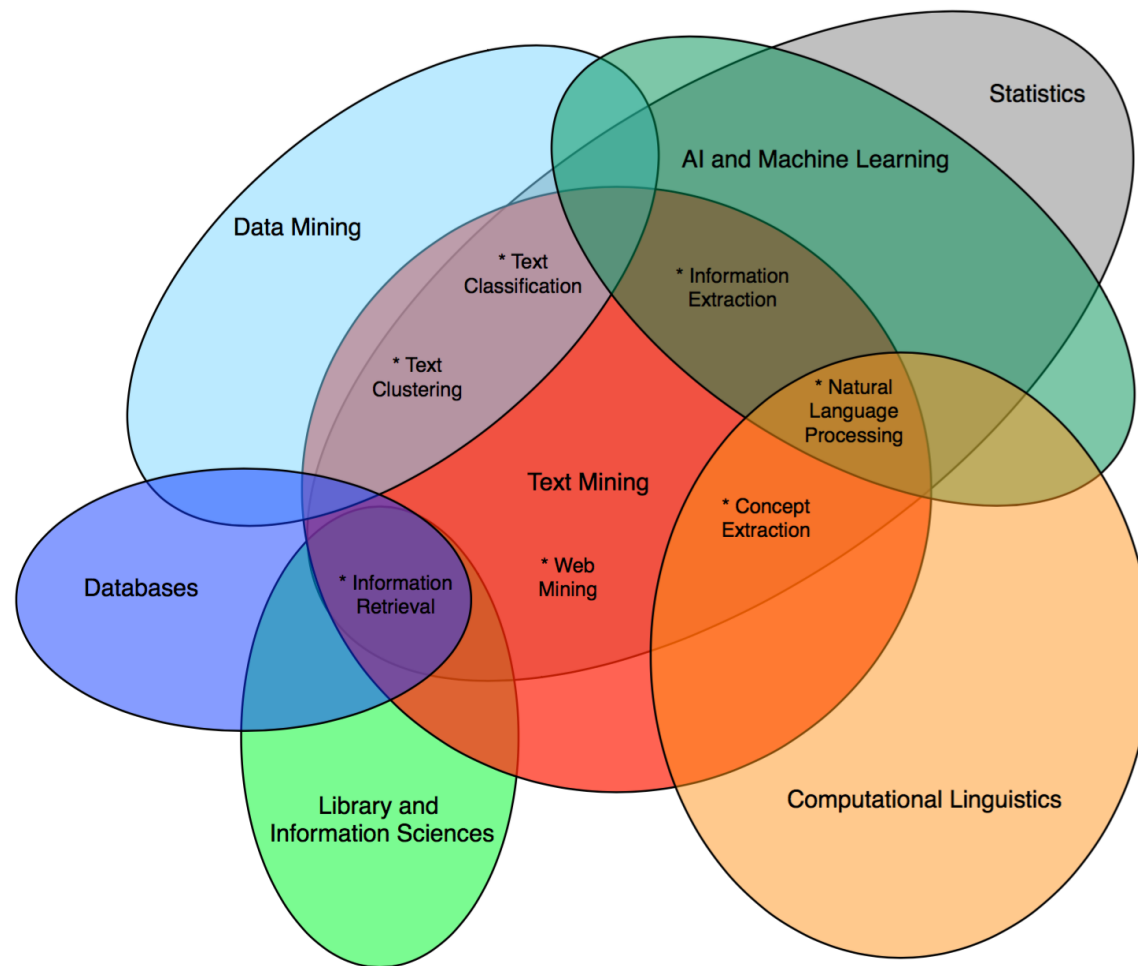
造血系统：无乏力,发  
晕,眼花 ,牙龈出血  
鼻出血 皮下出血  
骨痛。

法院判决书

2015年7月14日， 本院根据天津市茂发房地产经营有限公司的申请裁定受理天津市茂发房地产经营有限公司破产清算一案。查明， 债务人截止至2015年7月31日， 资产总额10596.56元， 负债总额 563391.46 元， 资产负债率 5316.74%。本院认为， 债务人天津市茂发房地产经营有限公司为有限责任公司， 具有破产主体资格。依照《中华人民共和国企业破产法》第二条， 本院于2015年11月5日裁定宣告天津市茂发房地产经营有限公司破产。

# 文本分析中的主要问题

- 分词和词性标注
- 实体抽取和信息抽取
- 文本分类和聚类
- 文本语义分析
- 主题分析
- 情感分析
- 信息检索
- 机器翻译



# 文本分析的重要性

- 与结构化数据结合，提升决策和预测模型的准确性
  - 结合互联网舆情和法院判决，评估企业信用状况
  - 结合交易流水文本，提高用户画像精确度
- 情感分析等技术应用广泛
  - 股票市场分析
  - 互联网舆情分析与监控
  - 商品服务质量评估
- 人工智能系统
  - IBM的Watson：NLP和文本分析是核心技术

用户ID	消费时间	消费金额	备注
20188230	1414857600	2309	杭州联华华商集团联华超市龙都连锁店
20188230	1414771200	6500	湖州天虹百货有限公司
20188230	1414512000	2096	支付宝 - 中国铁路总公司资金清算中心
20188230	1414771200	22450	湖州市星火服装有限公司
20188230	1414771200	20900	湖州市星火服装有限公司
20188230	1414771200	2340	吴兴晓华化妆品商行
22569099	1414771200	7600	财付通快捷支付 ( 客服 :0755-86013860)
22569099	1414771200	4100	北京弘泰基业 ( 大悦城 )
22569099	1414598400	2550	网银在线 ( 北京 ) 科技有限公司
22569099	1414771200	2000	北京悦府盛宴餐饮管理有限公司

# 文本分析的挑战性

- 歧义性，需结合上下文分析
  - 一词多义：“这款车的油耗很高” “这部新手机的性价比相当高”
  - 多词同义：“发货速度快” “物流迅速” “物流超快”；“计算机”，“电脑”
- 高维与稀疏性
  - 使用向量空间模型（VSM）表示文本时，维度往往较高（万-百万）
  - 只有少数维度取值不为0
- 表达的随意性
  - 网络用语，拼写错误，缩写等
  - “十动然拒”，“然并卵”，“老司机”




## 本节内容

- 如何对非结构化的文本进行表示？
  - 文本表示模型
- 如何进行文本分类和聚类？
  - 文本分类和聚类
- 如何挖掘文本中隐含的语义信息？
  - 主题分析
- 如何理解文本中蕴含的情感信息？
  - 情感分析

# 目录

第一部分  简介

第二部分  文本表示模型

第三部分  文本分类与聚类

第四部分  主题分析

第五部分  情感分析



# 向量空间模型

- VSM是20世纪60年代末期由 Salton 等人提出的，最早用在 SMART 信息检索系统中，目前已成为自然语言处理中常用的模型
- 向量空间模型(Vector Space Model)
  - 将文本表示成高维的向量，每一个维度代表一个词，取值表示词在文本中出现的频次(或其他取值)

用数据刻画规律，以数据描摹个体，用数据创造价值。



词典	TF
数据	3
刻画	0
中国	0
规律	1
描摹	1
普林	0
个体	1
价值	1

# 如何定义每一个维度上的取值？

用数据刻画规律，以数据描摹个体，让数据创造价值。

分词 ↓

用 数据 刻画 规律 数据 描摹 个体 让 数据 创造 价值

以 用 让 ...

停用词过滤 BOW, TF, TF-IDF转换

- 词袋模型 (Bag-of-words)
  - 将文本表示成语料词典大小维度的高维向量
  - 用1和0表示某个词是否出现在文本中
  - 语料词典维护词到对应维度索引号的映射
  - 使用稀疏矩阵来表示
- TF模型 (Term Frequency)
  - 与词袋模型不同在于用词在文本中的出现次数来代替是否出现
  - $tf(t,d)$ 表示词在文档d中出现的次数
  - $tf(t,d)$ 代表了词t在文档d中的重要程度
- TF-IDF
  - 使用词在整个语料中的逆文档频率(idf)来归一化tf权重
  - 一定程度消除常用词在文档中权重过高的问题
  - $idf(t) = \log(N/n(t))$ ,  $n(t)$ 为词t出现的文档数,  $N$ 为语料中所有文档数
  - $Tf-idf(t) = tf(t,d)*idf(t)$

词典	索引号	词袋模型	TF模型
数据	134	1	3
刻画	156	0	0
中国	567	0	0
规律	1076	1	1
描摹	1024	1	1
普林	2048	0	0
个体	2314	1	1
价值	2457	1	1



## 多个词特征：N-gram

- N-gram: 将多个词组合起来当作单一的特征
  - N表示考虑多少个词进行组合
  - 1-gram (unigrams): 公司
  - 2-gram (bigram,digrams) : 有限公司
  - 3-gram (trigrams) : 科技有限公司
- 一定程度上考虑了词序信息
- 维度成指数级增长
- 稀疏性进一步增大

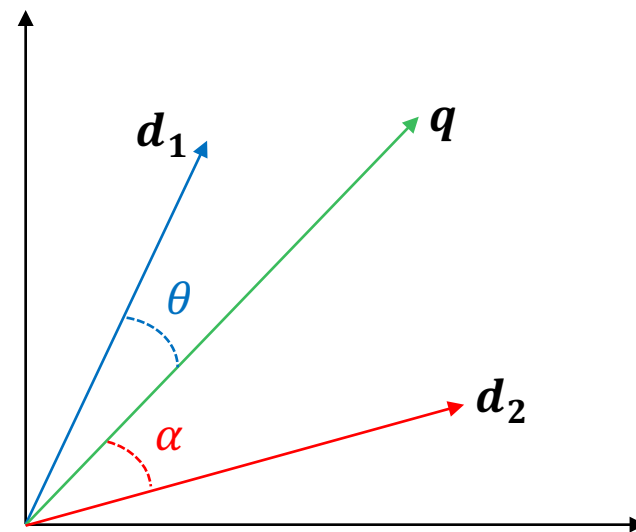


## 计算文本相似度

- 文本之间的相似度计算
  - 例如用户输入的查询 $q$ 和查询与文档 $d$ 的相关性

$$\cos(q, d) = \frac{q^T d}{\|q\| \|d\|}$$

- 按照与查询的相似度对文档进行排序
- 查询和文档都是稀疏向量, 因此计算效率高





## 向量空间模型的缺陷

- 维度灾难
  - 向量空间模型表示文本时，维度很高（万-百万）
  - 模型的学习需要考虑维度灾难问题
- 稀疏性
  - 文本向量十分稀疏
- 语义信息（例如同义词和多义词）
  - 计算相似度时，只对词进行计算，忽略词之间的语义关系
  - $\text{text1} = \text{“发货 速度 快”}$ ,  $\text{text2} = \text{“物流 迅速”}$
  - $\cos(\text{text1}, \text{text2}) = 0$
- 丢失词序
  - $\text{text1} = \text{“我 不 是 很 喜 欢 这 件 衣 服”}$
  - $\text{text2} = \text{“我 很 是 不 喜 欢 这 件 衣 服”}$



## 文本降维

- 简单的方法
  - 具体问题，选取词的子集（分析用户兴趣时，只选取跟兴趣相关的词）
  - 去除停用词和常用词（“了”，“的”，“是”）
- 模型的方法
  - LSA (Latent Semantic Analysis)
  - pLSA (probabilistic Latent Semantic Analysis)



# 目录

- 第一部分  简介
- 第二部分  文本表示模型
- 第三部分  文本分类与聚类
- 第四部分  主题分析
- 第五部分  情感分析

# 文本分类和聚类的应用

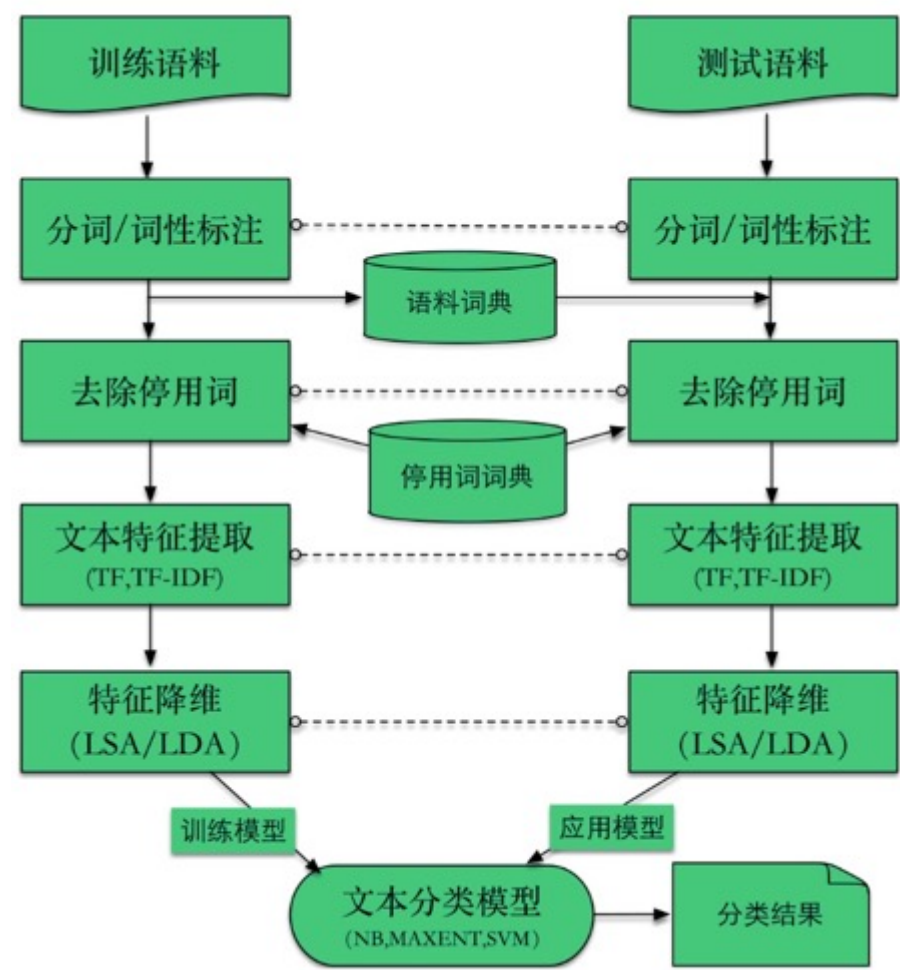
- 垃圾邮件分类（垃圾邮件/正常邮件）
- 新闻自动归类（体育/政治/财经）
- 情感分类（好评/中评/差评；喜/怒/哀/乐; 支持/反对）
- 人名消歧（“刘志军”）
- 互联网舆情分析

刘志军：铁道部原部长
刘志军：山东省汶上
刘志军：定西市政府研究室经济科科长
刘志军：刘志军试任河北省人民政府督查室督查专员
刘志军：宁夏人大内务司法工作委员会副主任
刘志军：双辽市粮食局副局长
刘志军：中铁二院贵阳勘察设计院副总工程师
刘志军：人民检察院反贪局侦查二科科长
刘志军：武警部队训练局副局长
刘志军：哈尔滨市人社局局长，党委书记
刘志军：昆明市政协副秘书长



# 文本分类和聚类流程

- 与一般的分类聚类的区别：
  - 分词、词性标注、去除停用词等预处理
  - 文本非结构数据的结构化（VSM）
  - 文本降维
- 常用的分类和聚类算法
  - Naïve Bayes
  - Support Vector Machines
  - Maximum Entropy（最大熵）



# 目录

- 第一部分  简介
- 第二部分  文本表示模型
- 第三部分  文本分类与聚类
- 第四部分  主题分析
- 第五部分  情感分析



# LDA

- 潜在狄利克雷分配 (Latent Dirichlet Allocation)
- 2003年, David Blei, Andrew Ng和Michael Jordan提出
- 在pLSA的基础上, 为文档增加一个概率模型, 使得文档的主题分布能够生成



# LDA生成过程

## Topics

gene 0.04  
dna 0.02  
genetic 0.01  
...

life 0.02  
evolve 0.01  
organism 0.01  
...

brain 0.04  
neuron 0.02  
nerve 0.01  
...

data 0.02  
number 0.02  
computer 0.01  
...

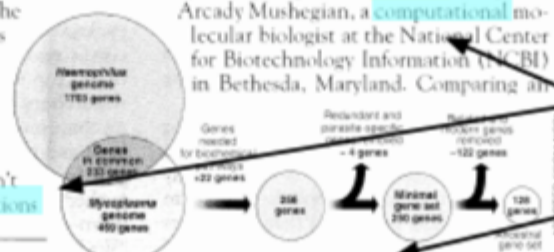
## Documents

### Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,\* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions

"are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson at Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

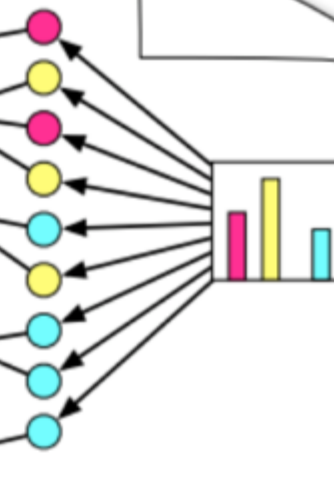


\* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

SCIENCE • VOL. 272 • 24 MAY 1996

## Topic proportions and assignments

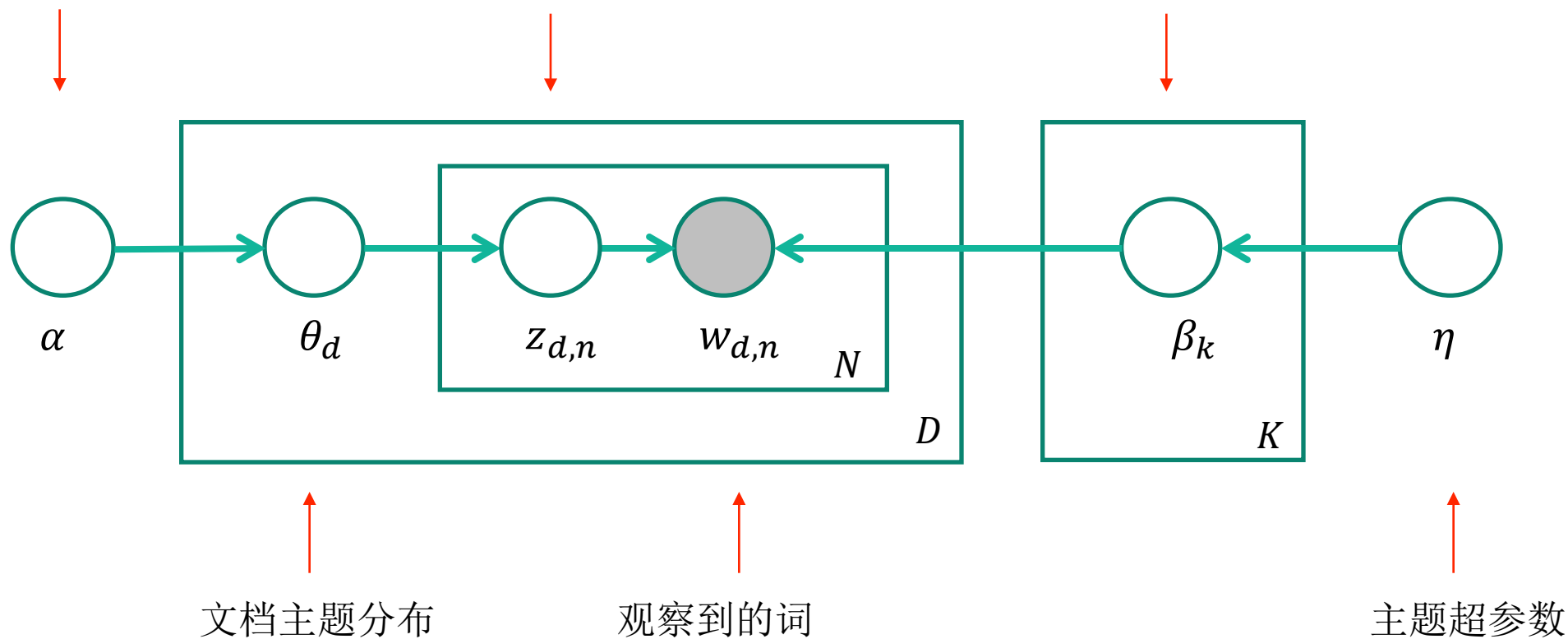


# LDA的图形化表示

Dirichlet超参数

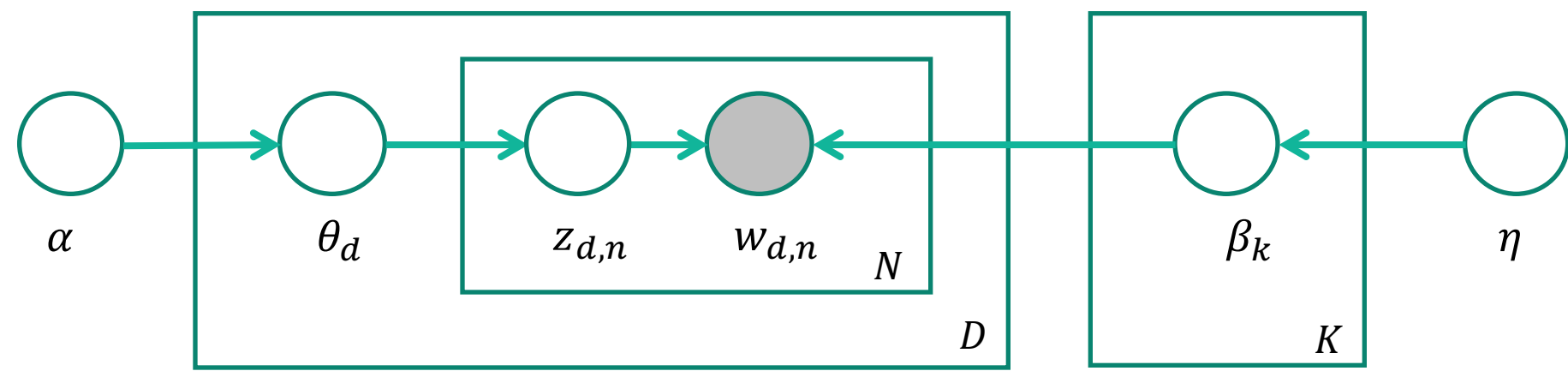
词所属主题

主题





# LDA 目标



- 对于给定的文档集，推理：
  - 每一个词属于哪一个主题  $z_{d,n}$
  - 每一篇文档的主题分布  $\theta_d$
  - 整个文档集中的主题分布  $\beta_k$





THANK YOU !