博雅大数据学院
BOYA BIG DATA COLLEGE

第十次课 文本分析

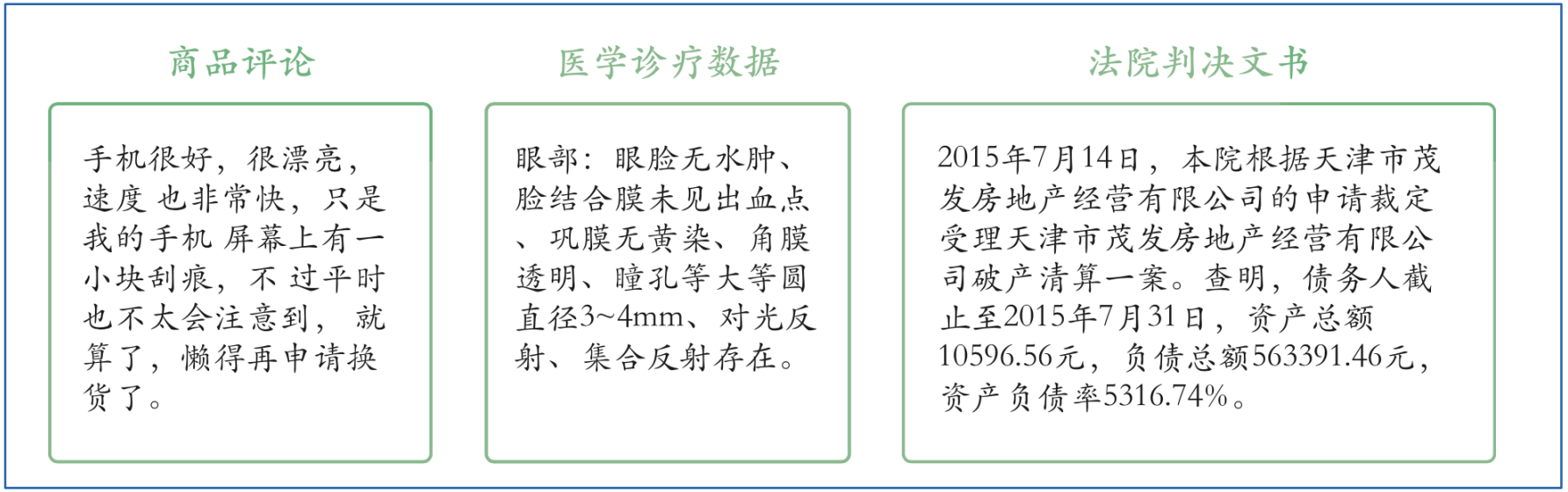
何向南
hexn@ustc.edu.cn

15 May 2020

目录

文本数据的普遍性

- 文本数据来源
  - 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
  - 法院判决文书



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Waston：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

文本分析面临的挑战

- 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

本讲内容

- 如何对非结构化的文本进行表示?
  - 文本表示模型
- 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

文本表示方法

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	1
中国	0
规律	1
描摹	1
博雅	0
个体	1
价值	1

TF模型

- TF模型（Term Frequency）
- 文本特征向量的每一个维度对应词典中的一个词，其取值为该词在文档中的出现频次
- 给定词典$W = \{w_1, w_2, \dots, w_V\}$，文档 d 可以表示为特征向量 $d = (t_1, t_2, \dots, t_V)$
  其中 V 为词典大小，w_i 表示词典中的第 i 个词，t_i 表示词 w_i 在文档 d 中出现的次数
- $tf(t, d)$表示词在文档 d 中出现的次
- $tf(t, d)$代表了词t在文档 d 中的重要程度

用数据刻画规律，以数据描摹个体，让数据创造价值。

$\boldsymbol{d} = (\mathbf{3}, \mathbf{1}, \mathbf{0}, \mathbf{1}, \mathbf{1}, \mathbf{0}, \mathbf{1}, \mathbf{1})$

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

TF模型

- TF模型的缺点
  模型假设文档中出现频次越高的词对刻画文档信息所起的作用越大，而不考虑不同词对区分不同文档的不同贡献
  - 例如在中文中，“的”、“了”和“是”等功能性词会在大部分文档中以较高的频次出现，然后这些词对于刻画文档的信息作用却不大
- 改进：TF-IDF模型

TF-IDF模型

- TF-IDF (Term Frequency-Inverse Document Frequency)
  在计算每一个词的权重时，不仅考虑词频，还考虑包含词的文档在整个文档集中的频次信息
  - 词 t 的文档频率 df(t) 是指文档集中出现了词 t 的文档数量
  - 词的逆文档频率

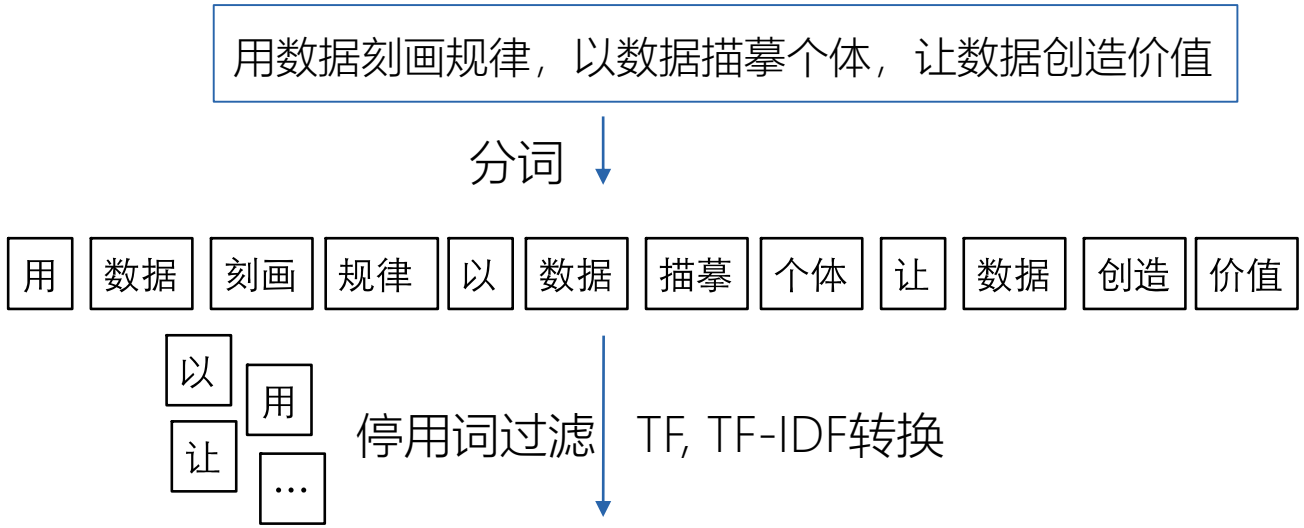
$$idf_t(t) = \log \frac{n+1}{df(t)+1} + 1$$

  - 则有

$$tf - idf(t) = tf(t,d) \cdot idf(t)$$

  - 一定程度消除常用词在文档中权重过高的问题

文本处理流程



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

N-gram模型

- N-gram: 将多个词组合起来当作单一的特征
  - 考虑次序，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

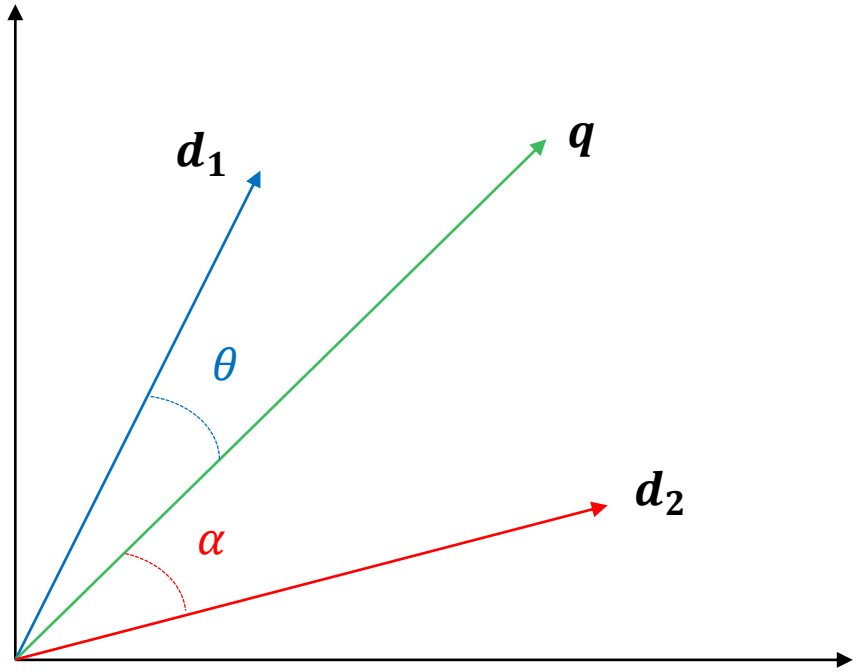
- 一定程度上考虑了词序信息
- 维度成指数级增长
- 稀疏性进一步增大

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$\boldsymbol{q}$和查询与文档$\boldsymbol{d}$的相关性

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

  - 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
  - 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
  - text1 = “发货 速度 快”, text2 = “物流 迅速”
  - $cos(text1, text2) = 0$
- 丢失词序
  - text1 = “我 不 是 很 喜欢 这件 衣服”
  - text2 = “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
  - 隐含语义分析 (Latent Semantic Analysis， LSA)
  - 概率隐含语义分析 (probabilistic Latent Semantic Analysis，pLSA)

隐含语义分析

- LSA的基本思想：将文本和词映射到一个低维表示
- 低维空间能够反映语义关系（隐语义空间）

- LSA的基本思想：将文本和词映射到一个低维表示低维空间能够反映语义关系（隐语义空间）

奇异值分解（Singular Value Decomposition）

$$\mathbf{X} = \mathbf{U} \Sigma \mathbf{V}^{\mathrm{T}}$$

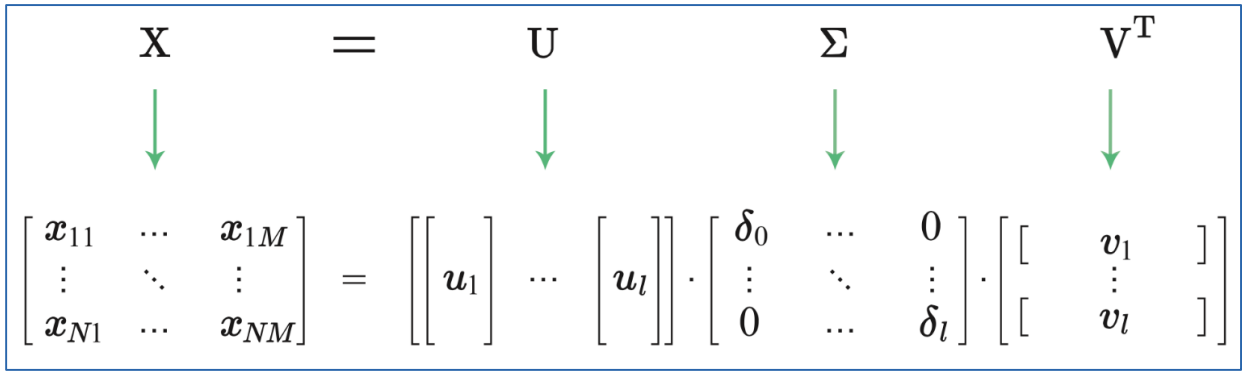
$$\begin{bmatrix} x_{11} & \cdots & x_{1M} \\ \vdots & \ddots & \vdots \\ x_{N1} & \cdots & x_{NM} \end{bmatrix} = \begin{bmatrix} \begin{bmatrix} \\ u_1 \\ \\ \end{bmatrix} & \cdots & \begin{bmatrix} \\ u_l \\ \\ \end{bmatrix} \end{bmatrix} \cdot \begin{bmatrix} \delta_0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \delta_l \end{bmatrix} \cdot \begin{bmatrix} [& &] \\ & v_1 & \\ & \vdots & \\ [& v_l &] \end{bmatrix}$$

隐含语义分析

- 在潜在语义空间计算文本之间的相似度
- 假设文本集合表示成 N 行 M 列的矩阵 **X**
  N 表示词的个数， M 表示文档数量，X的第 j 列表示文档 d_j
- 对矩阵 X 进行奇异值分解 (SVD)

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$$

（U为左奇异向量， Σ为奇异值降序排序后构成的对角矩阵，V^T为右奇异向量）



语义空间表示

- 选择 k 个最大的奇异值（k < l，l是矩阵 X 的秩），以及 U 和 V^T中对应的奇异向量，能够得到一个矩阵 X 的近似 X_k

$$X_k = U_k \Sigma_k V_k^T$$

- 得到的矩阵X_k是原始矩阵 X 在F范数下误差最小的近似
- U_k是一个 N 行 k 列的矩阵，每一行代表一个词在 k 维语义空间表示
- V_k^T是一个 k 行 M 列矩阵，每一列代表文档在k 维语义空间的表示

$$X_k = U_k \quad \Sigma_k \quad V_k^T$$

$$X_k = \begin{bmatrix} u_1 \end{bmatrix} \cdots \begin{bmatrix} u_k \end{bmatrix} \cdot \begin{bmatrix} \delta_0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \delta_k \end{bmatrix} \cdot \begin{bmatrix} & v_1 & \\ & \vdots & \\ & v_k & \end{bmatrix}$$

词和文档语义空间转换的示例

- 文档集包含4篇关于计算机视觉的文档$\{c_1, c_2, c_3, c_4\}$和4篇关于医疗大数据的文档$\{m_1, m_2, m_3, m_4\}$
- 下划线标注的词为我们关注的词

- c_1: 与此同时，计算机视觉技术仍然面临诸如信息描述模糊、图像特征检测不稳定且效率低下等问题；
- c_2: CV 技术的应用十分广泛，如数字图像像检索管理、医学影像分析、智能安检、人机交互等；
- c_3: 数字图像和视频数据蕴含了丰富的视觉资源，如何智能化地提取和分析影像中的有用信息逐渐成为近年的研究热点；
- c_4: 虽然取得了上述成就，但是和精密且灵活的人类视觉水平相比，CV 系统仍显得过于粗糙；

- m_1: 随着人类基因组学研究和高通量技术的发展，涉及蛋白质知识以及相关疾病、药物、基因的医学文献呈指数增长；
- m_2: SemRep 得到的特定疾病的 MEDLINE 文献的语义输出，通过显著信息提取算法对该语义输出进行打分排序；
- m_3: 利用文本挖掘技术从大量的生物医学文本中发现和抽取有价值的新颖的蛋白质知识已经成为可能；
- m_4: 实验结果对理解疾病的基因、蛋白质功能预测以及药物辅助设计都有重要的研究意义.

词和文档语义空间转换的示例

- 使用TF模型来计算词的权重，文档集矩阵 $\mathbf{X}$

	c_1	c_2	c_3	c_4	m_1	m_2	m_3	m_4
图像	1	1	1	0	0	0	0	0
人机交互	1	0	0	0	0	0	0	0
视觉	1	0	1	1	0	0	0	0
影像	0	1	1	0	0	0	0	0
研究	0	0	1	0	1	0	0	1
蛋白质	0	0	0	0	1	0	1	1
基因	0	0	0	0	2	0	0	1
疾病	0	0	0	0	1	1	0	1

- 利用LSA模型 $\mathbf{k} = 2$ ，得到近似矩阵 $\mathbf{X}_2$

	c_1	c_2	c_3	c_4	m_1	m_2	m_3	m_4
图像	0.886	0.67	1.165	0.354	-0.029	-0.037	-0.037	0.019
人机交互	0.291	0.22	0.376	0.116	-0.046	-0.018	-0.018	-0.021
视觉	0.783	0.592	1.031	0.313	-0.016	-0.031	-0.031	0.024
影像	0.595	0.451	0.79	0.238	0.017	-0.018	-0.018	0.041
研究	0.308	0.242	0.605	0.126	1.115	0.18	0.18	0.856
蛋白质	-0.086	-0.055	0.094	-0.031	1.174	0.205	0.205	0.882
基因	-0.114	-0.072	0.147	-0.04	1.678	0.292	0.292	1.261
疾病	-0.086	-0.055	0.094	-0.031	1.174	0.205	0.205	0.882

文档之间的相似度

- 原始表示下文档的相似度度量

	c_1	c_2	c_3	c_4	m_1	m_2	m_3	m_4
c_1	1							
c_2	0.14	1						
c_3	0.26	0.57	1					
c_4	0.49	-0.21	0.37	1				
m_1	-0.69	-0.51	-0.53	-0.33	1			
m_2	-0.29	-0.21	-0.37	-0.14	0.2	1		
m_3	-0.29	-0.21	-0.37	-0.14	0.2	-0.14	1	
m_4	-0.77	-0.57	-0.5	-0.37	0.89	0.37	0.37	1

- 近似表示下文档的相似度度量

	c_1	c_2	c_3	c_4	m_1	m_2	m_3	m_4
c_1	1							
c_2	0.99	1						
c_3	0.98	0.98	1					
c_4	1	0.99	0.85	1				
m_1	-0.85	-0.84	-0.76	-0.84	1			
m_2	-0.88	-0.87	-0.79	-0.87	0.99	1		
m_3	-0.88	-0.87	-0.79	-0.87	0.99	1	1	
m_4	-0.84	-0.84	-0.75	-0.83	0.99	0.99	0.99	1

词之间的相关性

- 图像和影像在计算机视觉领域表示的是相近的含义
- 皮尔逊相关性

Corr(图像，影像) = 0.46

	c_1	c_2	c_3	c_4	m_1	m_2	m_3	m_4
图像	1	1	1	0	0	0	0	0
人机交互	1	0	0	0	0	0	0	0
视觉	1	0	1	1	0	0	0	0
影像	0	1	1	0	0	0	0	0
研究	0	0	1	0	1	0	0	1
蛋白质	0	0	0	0	1	0	1	1
基因	0	0	0	0	2	0	0	1
疾病	0	0	0	0	1	1	0	1

Corr(图像，影像) = 0.99

	c_1	c_2	c_3	c_4	m_1	m_2	m_3	m_4
图像	0.886	0.67	1.165	0.354	-0.029	-0.037	-0.037	0.019
人机交互	0.291	0.22	0.376	0.116	-0.046	-0.018	-0.018	-0.021
视觉	0.783	0.592	1.031	0.313	-0.016	-0.031	-0.031	0.024
影像	0.595	0.451	0.79	0.238	0.017	-0.018	-0.018	0.041
研究	0.308	0.242	0.605	0.126	1.115	0.18	0.18	0.856
蛋白质	-0.086	-0.055	0.094	-0.031	1.174	0.205	0.205	0.882
基因	-0.114	-0.072	0.147	-0.04	1.678	0.292	0.292	1.261
疾病	-0.086	-0.055	0.094	-0.031	1.174	0.205	0.205	0.882

LSA小结

- 优点
  - 低维空间下的文档表示能够处理同义词现象（多词同义）
  - 通过降维,能够减小数据噪音
  - 经验结果表明，LSA比简单的向量空间模型表现更好，特别是在信息检索任务
- 缺点
  - 没有解决"一词多义"的问题
  - 需要人为选择k
  - 仍然没有考虑文档中的词序
  - 没有概率模型来对词的频次建模

概率隐含语义分析

- 概率隐含语义分析(pLSA)在LSA的 基础上增加概率模型，以克服稀疏性问题
- 其中 $p(d)$ 表示文档 d出现的概率，$p(w|d)$表示词w在文档d中生成的概率
- 假设隐含语义变量为 z，则：

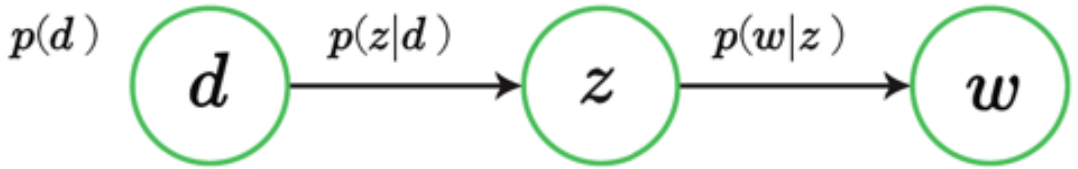
$$p(w|d) = \sum_z p(w|z)p(z|d)$$

$p(z|d)$：文档 d 中语义 z 的概率

$p(w|z)$：词 w 从语义 z 中生成的概率

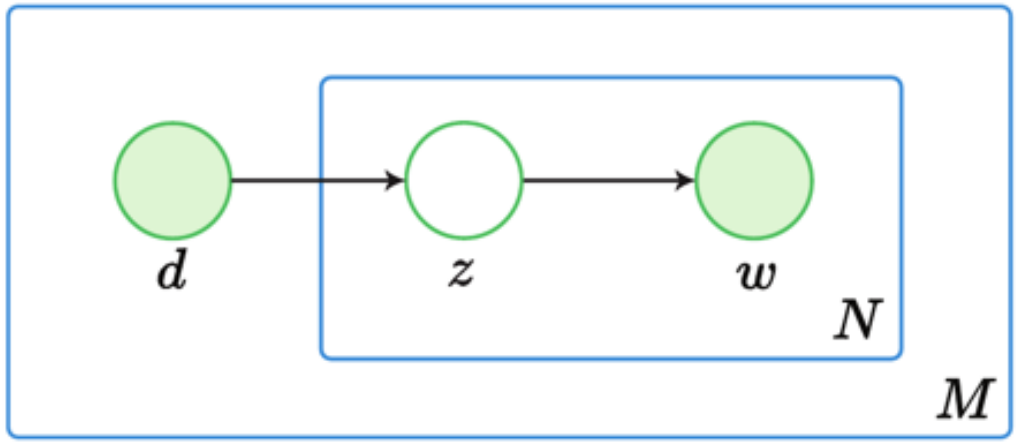
pLSA 生成过程

- pLSA属于生成模型，给定文档 d 后，以一定的概率选择 d对应的语义(主题) z，然后以一定概率选择 z 中的单词 w



- 其中 $p(d)$ 表示文档 d出现的概率，$p(z|d)$ 表示文档 d中主题 z的出现概率，$p(w|z)$ 表示给定主题 z出现单词 w的概率
- 每个主题中的词语服从多项分布，每篇文章的主题是服从多项分布

pLSA的图形化表示



- 假设有N个词，D个文档
- d为文档变量，$p(z|d)$是文档的主题分布，$p(w|z)$为主题在词上的分布
- d和z为观察变量(observable variable)，z为隐含变量(latent variable)
- 参数估计方法：最大对数似然估计和EM算法

$$p(w_j, d_i) = p(d_i) \sum_z p(w_j|z)p(z|d_i)$$

pLSA似然函数

- pLSA对数似然函数

$n(w_j, d_i)$为词w_j在文档d_i中出现的次数

$$L = \sum_{i,j} n(w_j, d_i) \log p(w_j, d_i)$$

$$= \sum_{i,j} n(w_j, d_i) \log \left(p(d_i) \sum_z p(w_j|z) p(z|d_i) \right)$$

- 约束条件：$\sum_z p(z|d_i) = 1$ 和 $\sum_j p(w_j|z) = 1$

pLSA的EM步骤

- E-步骤

$$p(z|d_i,w_j)=\frac{p(w_j,z|d_i)}{p(w_j|d_i)}=\frac{p(w_j|z,d_i)p(z|d_i)}{\sum_z p(w_j|z,d_i)p(z|d_i)}$$

- M-步骤
  - $n(w_j,d_i),n(d_i),\ \mathrm{p}(d_i)$可通过数据直接计算
  - 按照以下步骤重新估计参数$p(z|d_i)$和$p(w_j|z)$

$$p(w_j|z)=\frac{\sum_i n(w_j,d_i)p(z|d_i,w_j)}{\sum_m\sum_i n(w_m,d_i)p(z|d_i,w_m)} \qquad p(z|d_i)=\frac{\sum_j n(w_j,d_i)p(z|d_i,w_j)}{n(d_i)}$$

- E步骤和M步骤交替迭代，直至收敛

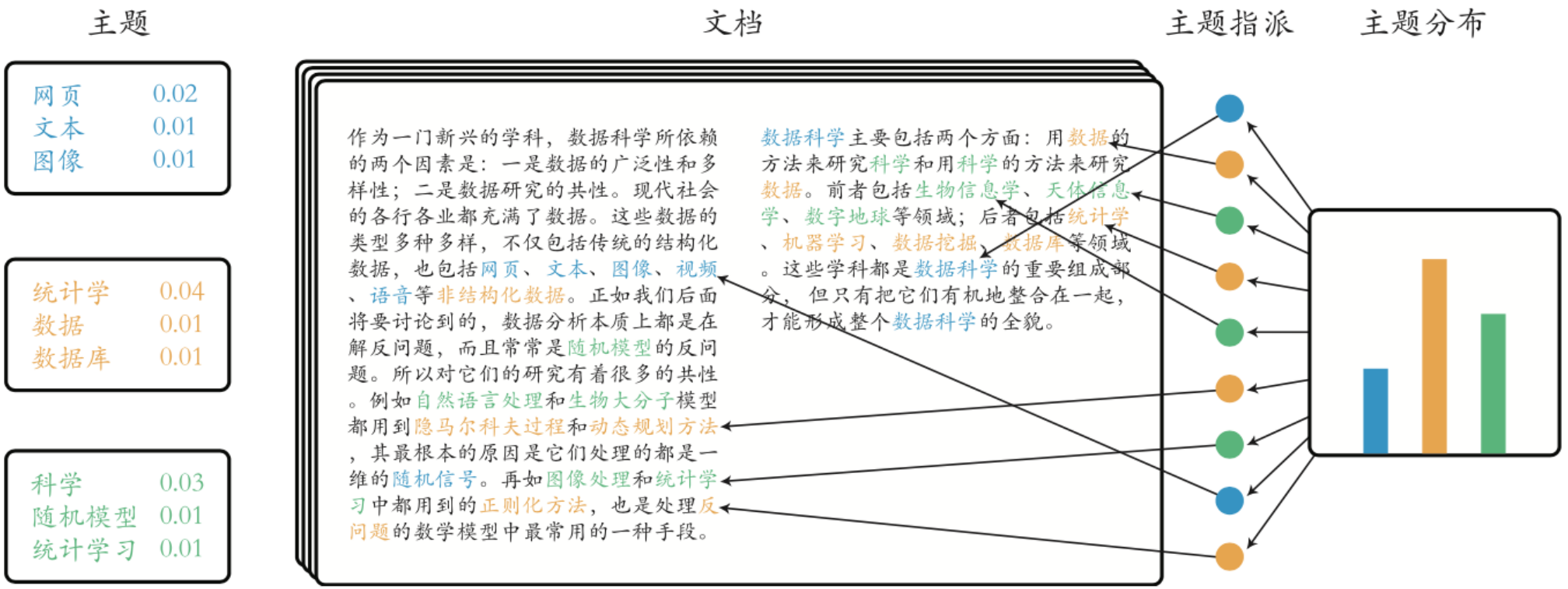
主题模型和LDA

主题模型

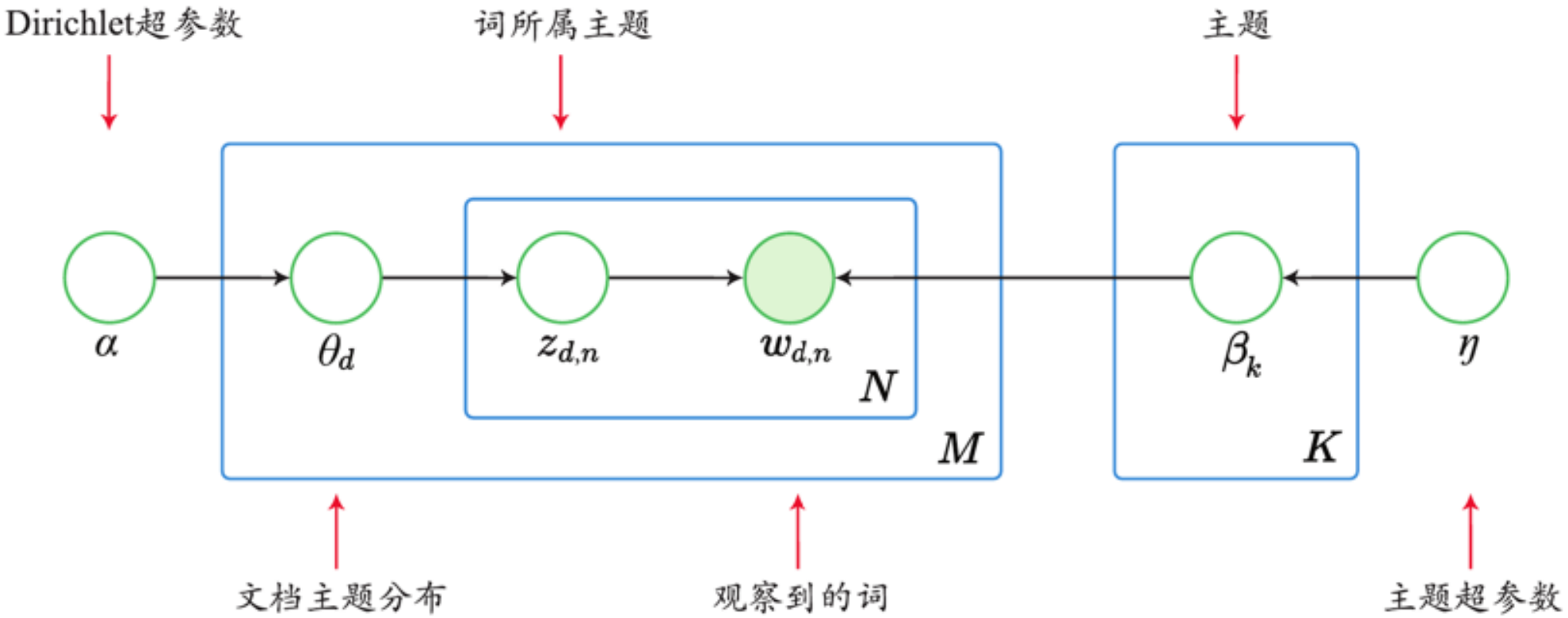
- 主题模型可以理解成一种从文档集中发现主题的方法，这些主题能够很好地代表文档中包含的信息
- 主题模型
  - 发现文档集中隐含的主题模式
  - 根据主题来对文档进行标注
  - 利用文档的主题标注来对文本数据进行组织、查询和摘要

- 主题是由一组揭示主题含义的词所构成
  - 对于``饮食''这个主题，可以由``早餐'',``咖啡'',``水果''和``鸡蛋''等词来表示
- 研究人员提出多种主题模型，最为典型的主题模型为
  - 隐含狄利克雷分配(Latent Dirichlet Allocation, LDA)模型
    - 由David Blei，Andrew Ng和Michael Jordan于2003年提出的
    - 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

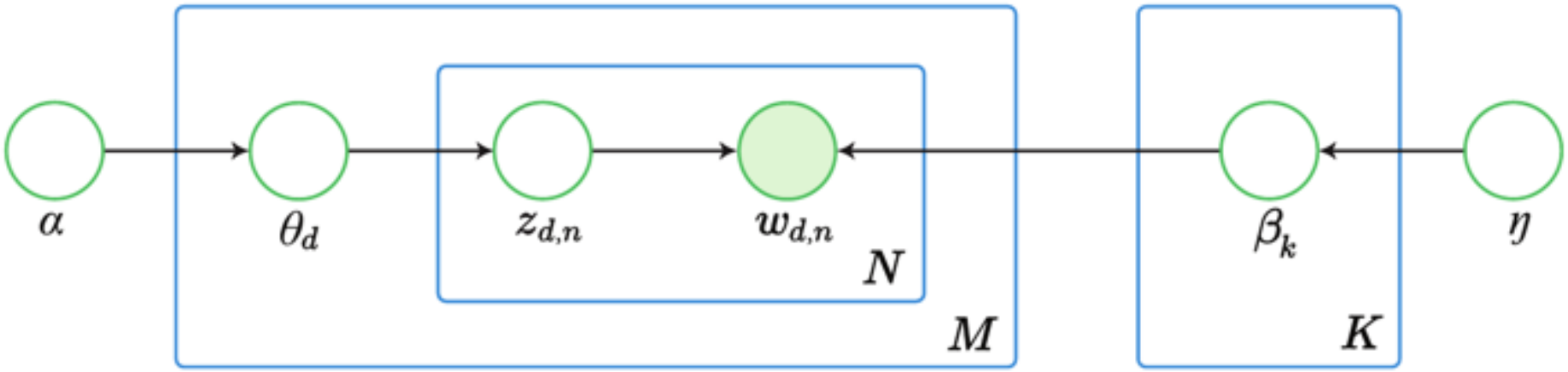

主题
网页 0.02
文本 0.01
图像 0.01
统计学 0.04
数据 0.01
数据库 0.01
科学 0.03
随机模型 0.01
统计学习 0.01
文档
作为一门新兴的学科，数据科学所依赖的两个因素是：一是数据的广泛性和多样性；二是数据研究的共性。现代社会的各行各业都充满了数据。这些数据的类型多种多样，不仅包括传统的结构化数据，也包括网页、文本、图像、视频、语音等非结构化数据。正如我们后面将要讨论到的，数据分析本质上都是在解反问题，而且常常是随机模型的反问题。所以对它们的研究有着很多的共性。例如自然语言处理和生物大分子模型都用到隐马尔科夫过程和动态规划方法，其最根本的原因是它们处理的都是一维的随机信号。再如图像处理和统计学习中都用到的正则化方法，也是处理反问题的数学模型中最常用的一种手段。
数据科学主要包括两个方面：用数据的方法来研究科学和用科学的方法来研究数据。前者包括生物信息学、天体信息学、数字地球等领域；后者包括统计学、机器学习、数据挖掘、数据库等领域。这些学科都是数据科学的重要组成部分，但只有把它们有机地整合在一起，才能形成整个数据科学的全貌。
主题指派
主题分布

LDA的图形化表示


Dirichlet超参数
词所属主题
主题
α
θ_d
$z_{d,n}$
$w_{d,n}$
N
M
β_k
K
η
文档主题分布
观察到的词
主题超参数

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θ_d
  - 整个文档集中的主题分布β_k



- 经过推导可得 Gibbs 抽样算法的更新规则为

$$p(z_{d,n}=k|\boldsymbol{z}_{-d.n},\boldsymbol{w},\alpha,\eta)=\frac{\eta_v+\Psi_{k,v}-1}{\sum_{v=1}^{V}(\eta_v+\Psi_{k,v})-1}\frac{\alpha_k+\Omega_{d,k}-1}{\sum_{k=1}^{K}(\alpha_k+\Omega_{d,k})-1}$$
$$\propto\frac{\eta_v+\Psi_{k,v}-1}{\sum_{v=1}^{V}(\eta_v+\Psi_{k,v})-1}(\alpha_k+\Omega_{d,k}-1)$$

- 主题β_k和文档主题分布θ_d的采样公式为

$$\beta_{k,v}=\frac{\Psi_{k,v}+\eta_v}{\sum_{v=1}^{V}\Psi_{k,v}+\eta_v}$$
$$\theta_{d,k}=\frac{\Omega_{d,k}+\alpha_k}{\sum_{k=1}^{K}\Omega_{d,k}+\alpha_k}$$

Gibbs采样

话题	APSM			ME-APSM		
	话题词	正向词	负向词	话题词	正向词	负向词
Staff	**staff**	staff	unhelpful	**staff**	good	**rude**
	helpful	friendly	poor	helpful	great	**unfriendly**
	friendly	courteous	bad	friendly	helpful	unhelpful
	english	helpful	noise	english	friendly	noise
	desk	attentive	cold	desk	excellent	poor
	front	clean	problem	extremely	wonderful	disappointed
	good	great	overpriced	**waiter**	staff	cheap
	extremely	recommend	disappointed	**waitress**	clean	hard
	nice	good	bother	breakfast	efficient	grumpy
	spoke	nice	arrogant	spoke	pleasant	complaints
Room	shower	nice	smell	room	rooms	small
	room	clean	dirty	bathroom	large	hard
	bathroom	great	stall	clean	good	tired
	water	spacious	cramped	bed	huge	uncomfortable
	small	comfortable	cold	shower	comfortable	noise
	bath	good	problem	small	shower	bad
	clean	modern	drain	comfortable	nice	worn
	towels	safe	broken	large	great	cold
	hot	warm	worn	size	clean	broken
	good	free	problems	spacious	room	dark
Meal	**breakfast**	breakfast	cold	**breakfast**	good	cold
	coffee	friendly	scrambled	**coffee**	great	scrambled
	buffet	fresh	problem	**fruit**	fresh	awful
	room	variety	hard	**buffet**	hot	limited
	fruit	good	bad	**eggs**	wonderful	terrible
	eggs	great	expensive	**cheese**	excellent	bad
	fresh	delicious	poor	**cereal**	nice	poor
	included	nice	die	**juice**	fantastic	disappointed
	great	clean	cheap	fresh	decent	cheap
	continental	hot	miss	pastries	loved	negative

LDA小结

- 作为经典的主题模型，LDA有不同的应用场景
  - 主题发现：发现大规模文档集中的隐含主题
  - 文本降维：文档主题分布向量作为文档的降维表示
  - 文本聚类

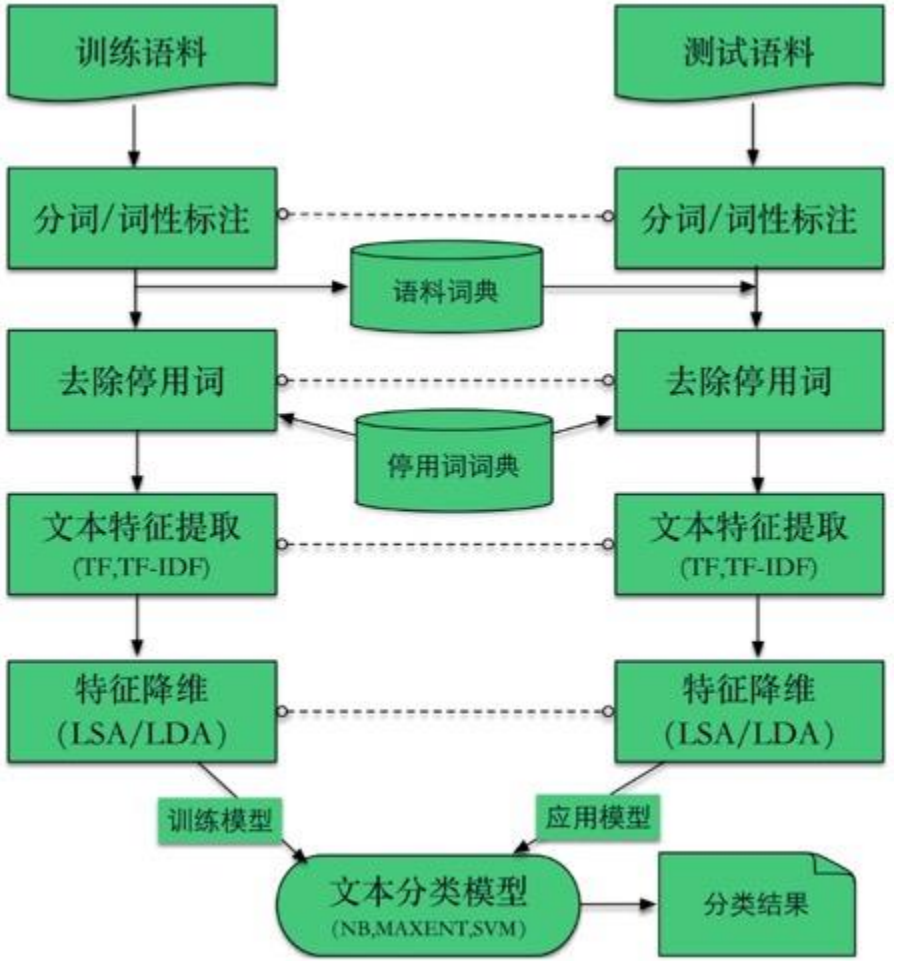
文本分类和聚类

- 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）
- 人名消歧（“刘志军”）
- 互联网舆情分析
- 情感分类（好评/中评/差评；喜/怒/哀/乐；支持/反对）

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研究院副总工程师
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
  - 朴素贝叶斯(Naïve Bayes)
  - 支持向量机(Support Vector Machines)
  - 最大熵(Maximum Entropy)



情感分析

情感分析

- 情感（观点）的含义
  - 人们的观点、情感、评价、态度、情绪等
- 情感分析（观点挖掘）
  - 分析和研究人们对实体的观点、情感、态度等
  - 实体：产品、服务、机构、个体、问题、事件、话题等

情感分类

- 情感分类的任务是根据文本对象中所包含的情感将其划分到不同的情感类别
  - 分类粒度： 文档、句子、短语、词语
  - 分类方式
    - 二类法：正向和负向
    - 三类法：正向、负向和中性
    - 多类法：1-5星或喜怒哀乐等情绪类别

情感分类

- 文档情感分类的基本假设：情感一篇文档只包含一个观点持有者针对一个实体或特性的观点：
- 对于商品评论等文本通常有效，在线讨论，辩论性文本，以及微博等往往包含不同观点持有者对不同的实体或特性的情感
- 情感分类方法：
  - 基于规则的方法
  - 基于有监督学习的方法

基于规则的方法

- 基本思路：根据文本词语的情感倾向性和一些人工定义的规则(否定处理规则和转折处理规则)等 来确定句子的情感倾向性
- 两个步骤：
  - 计算词的情感分数
  - 通过词的情感得到文本的情感
- 词情感分数的计算：
  - 利用人工编制的情感词典
  - 种子词方法（如“好”，“坏”）

- 基于规则的方法处理流程
  - 提取文档中的情感词
  - 否定和转折处理
  - 根据词语的情感确定文档的情感

词语的情感分数计算

- 在中文中，我们可以选取 好 和 坏 分别作为正向种子词和负向种子词，将其情感分别标注成 1 和 −1
- 计算其他词与正向种子词和负向种子词的语义相似度，通过语义相似度计算词的情感分数
- 点间互信息 (Pointwise Mutual Information，PMI)：词的语义相似度计算

$$PMI(w_1, w_2) = \log \frac{p(w_1, w_2)}{p(w_1)p(w_2)}$$

- 情感分数：

$$\mathrm{SO(w)} = \mathrm{PMI(w, “好”)} - \mathrm{PMI(w, “坏”)}$$

- 我们将一段文本中的所有词的情感分数进行平均，可以得到整段文本的情感分数

基于规则的方法

- 优点：
  - 可以应用在任何领域，且不需要进行数据标注（无监督）
- 缺点
  - 在特定的领域内，这种方法的分类准确率不高
  - 词的情感会随着领域变化，通用的情感词典或者使用情感分数计算的词的情感在特定领域可能不准确
    “阻碍癌细胞扩散”和“阻碍社会发展”中的阻碍代表不同的情感

基于有监督学习的方法

- 将情感倾向性分类当作一个分类问题来解决，主要不同在于特征的选择
- 使用的分类模型：朴素贝叶斯，支持向量机，最大熵等
- 特征选择（关键）
  - 词和词频特征：文档中的词及其在文档中出现的频次
  - 词法特征：包括词性标注特征和N-gram词组特征
  - 情感词特征：将情感词典中的词语作为单独的特征，例如统计正向词和负向词的数量（漂亮、喜欢、讨厌...）
  - 否定特征：包括否定词、情态动词、强化词和弱化词等（不、没有、非、否等）
  - 语法特征：语法依赖关系特征等

基于有监督学习的方法

- 电影评论数据集（1000条正向评论和1000条负向评论）

分类模型	分类准确率	使用的特征
支持向量机	82.9%	unigram
朴素贝叶斯	86.4%	unigram + 去除客观句
支持向量机 + 规则	86.2%	unigram + bigram
支持向量机	89.6%	unigram
支持向量机	90.5%	unigram, bigrams, trigrams, 语法依赖特征, 形容词情感特征

- 优点：
  - 在特定领域内，有监督的情感分类方法效果往往比无监督的情感分类方法要好
- 缺点
  - 有监督学习的方法需要使用不容易获取的标注数据，因此领域扩展性不佳

基于特性的情感分析

- 情感分类能够将文本按照情感倾向性进行分类，却不能告诉我们更细粒度的信息
  谁（情感持有者）在什么时间（时间）针对什么对象（实体）的哪种特性 （特性）发表了怎样的情感（情感）？

> *UserID:27086* 发表于 *2014* 年 *12* 月 *20* 日“手机很好，商家的服务全优，周到及时，赠品很好又实用，顺丰物流超快，千里发货不到一天就收到．手机通话是双待单通的不够方便，双通开通不了．这是机子的设计问题．对商家是很满意的，以后有需要还会再来．”

基于特性的情感分析

- 基于特性的情感分析 (aspect-based sentiment analysis)：抽取出与情感相关的诸多要素

观点编号	实体	特性	情感	观点持有者	时间
o_1	手机	手机	很好	27086	2014-12-20
o_2	商家	服务	全优、周到、及时	27086	2014-12-20
o_3	手机	赠品	很好、实用	27086	2014-12-20
o_4	手机	物流	超快	27086	2014-12-20
o_5	手机	双待单通	不够方便	27086	2014-12-20

- 侧重：实体和特性两个粒度；从实体和特性的粒度进行情感分析是许多实际应用的需求

基于特性的情感分析

- 五个子任务：实体抽取和分类、特性抽取和分类、特性情感分类、观点持有者抽取和分类以及时间抽取和标准化

观点编号	实体	特性	情感	观点持有者	时间
o_1	手机	手机	很好	27086	2014-12-20
o_2	商家	服务	全优、周到、及时	27086	2014-12-20
o_3	手机	赠品	很好、实用	27086	2014-12-20
o_4	手机	物流	超快	27086	2014-12-20
o_5	手机	双待单通	不够方便	27086	2014-12-20

特性抽取和分类

- 目标：从待分析的文本中找出所有的特性表达式，并将特性表达式聚集为特性
- 特性抽取的方法：
  - 基于名词和名词性词组频次的方法
  - 利用特性和情感之间关系的方法
  - 有监督学习的方法
  - 主题模型的方法

特性抽取和分类

- 特性抽取能够从文本中抽取出特性表达式，然而特性在具体文本中有不同的特性表达式
- 特性分类：将相同含义的特性表达式聚集在一起
- 特性分类的方法：
  - 语义距离和同义词关系
  - 字符串相似度
  - 聚类和主题模型

特性情感分类

- 目的：对提取出的每一个特性，确定其情感倾向性
- 方法：基于有监督学习的方法、基于情感词典的方法两类
- 基于有监督方法的方法存在的问题
  - 标注数据难以获取
  - 在一个领域训练的模型往往在其他领域表现不佳
  - 确定情感表达式的范围很困难，在一个句子中往往很难确定一个情感表达式是否覆盖了目标特性

特性情感分类

- 基于情感词典利用情感词典和一些规则模板来确定特性的情感倾向性
- 现有的大部分特性情感分类采用的是基于情感词典的方法

情感词典	说明
General Inquirer[5]	人工标注的情感词典，每一个词被标注为正向或负向
Opinion Lexicon[6]	Bing Liu 等提供的英语情感词典，大约 6800 个词
MPQA 主观性词典[7]	OpinionFinder 系统使用的情感词典资源
SentiWordNet[8]	利用 WordNet 扩展的情感词典，标注为正向、负向或中性
Emotion Lexicon[9]	每一个词标注为正向或负向，以及八种基本情绪之一
HowNet 知网情感词典[10]	知网中文情感词典

特性情感分类的步骤

1.情感词标记。对于那些包含特性的句子，标记句子中出现的所有情感词，对于正向情感词，赋予情感分数 +1；对于负向情感词，赋予情感分数 -1

- 例如：
  “这款手机的音质不好，但是电池寿命很长”
  经过标记后
  “这款手机的音质不好[+1]，但是电池寿命很长[+1]”

特性情感分类的步骤

2.极性偏移器应用。极性偏移器 (valence shifter)是那些能够改变情感倾向性的词语或词组（包括否定词、强化词和弱化词等）

- 例如：
  “这款手机的音质不好[+1]，但是电池寿命很长[+1]”
  经过极性偏移器应用后
  “这款手机的音质不好[-1]，但是电池寿命很长[+1]”

特性情感分类的步骤

3.转折处理。句子中转折的使用也能够影响情感倾向性

常用的转折词语包括“但是”和“而是”，转折词将句子分为两个部分

- 转折处理的方法是:如果被转折词分开的两个部分中的一个部分的情感倾向性难以确定，则取另一个部分的情感倾向性的反转

特性情感分类的步骤

4.情感聚集。应用一个情感聚集函数来将一个特性的所有句子的情感倾向性进行聚集得到特性的情感倾向性

- 对于特性集$\{a_1, a_2, \ldots, a_m\}$，句子 s 中的情感词集合$\{sw_1, sw_2, \ldots, sw_n\}$，情感词集合中的词语的情感倾向性由前三个步骤确定，则情感聚集函数为

$$\text{score}(a_i, s) = \sum_{sw_i}^{n} \frac{SO(sw_j)}{\text{dist}(sw_j, a_i)}$$

- 其中$\text{dist}(sw_j, a_i)$表示在句子 s 中情感词 sw_j 与指示特性的特性表达式 a_i 的距离，$SO(sw_j)$ 表示在句子 s 中，情感词 sw_j 的情感倾向性

小结

- 文本表示模型：TF模型，TF-IDF模型，N-gram模型
- 文本降维：LSA和pLSA
- 主题模型：LDA
- 情感分析：情感分类，基于特性的情感分析

雷文强博士

2019年于新加坡国立大学计算机学院获得博士学位，目前在新加坡国立大学从事博士后研究。他的研究兴趣主要在自然语言处理，近期重点关注对话系统，对话推荐系统。同时他也感兴趣语言学，同语言学家一起将系统功能语言学与现代计算技术结合。它在一流国际会议如ACL，IJCAI，AAAI，EMNLP等发表多篇论文。其研制的对话系统模型Sequicity被选为对话系统历史悠久的竞赛DSTC2019的baseline。其作为主要成员参与研发的机器翻译系统在新加坡本地四种语言（英语，汉语，马来语，泰米尔语）的互译中超过google translation，被新加坡多个政府部门采用为官方翻译系统。他当选新加坡自然语言处理会议 The 2020 Singapore Symposium on Natural Language Processing (SSNLP 2020)程序委员会主席。