PLSA模型

覃雄派

PLSA模型

提纲

- PLSA模型及EM算法
- PLSA模型实例
- PLSA模型实例：Python实现
- PLSA模型的优缺点
- PLSA模型的公式推导

PLSA模型

- PLSA模型及EM算法

Probabilistic Latent Semantic Analysis

Thomas Hofmann

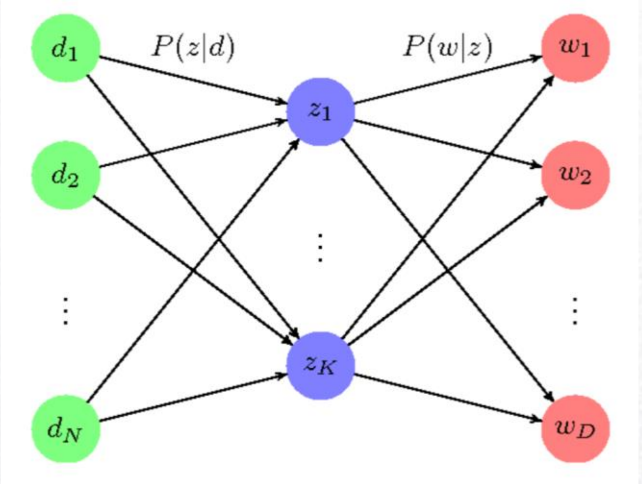
Probabilistic Latent Semantic Analysis is a novel statistical technique for the analysis of two-mode and co-occu
language processing, machine learning from text, and in related areas. Compared to standard Latent Semanti
Decomposition of co-occurrence tables, the proposed method is based on a mixture decomposition derived fro
foundation in statistics. In order to avoid overfitting, we propose a widely applicable generalization of maximun
consistent improvements over Latent Semantic Analysis in a number of experiments.

Comments: Appears in Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence (UAI1999)

Thomas Hofmann
EECS Department, Computer Science Division, University of California, Berkeley &
International Computer Science Institute, Berkeley, CA
hofmann@cs.berkeley.edu

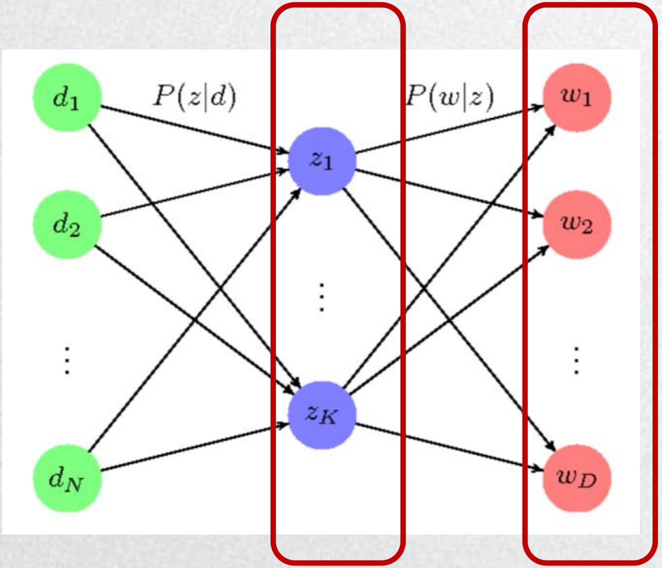
PLSA模型

- PLSA模型及EM算法
  - 一种话题建模的方法
    - 假设有一批文档
    - 每个文档使用了字典中的若干单词进行撰写
    - 这些文档描述了一些主题，如图所示
      - 我们了解到的是n(di, wj)
      - 即文档和单词的共现次数
    - 请自动找出这些主题
      - 具体是每个文档在主题上的分布
        - » P(zk | di)
      - 每个主题在单词上的分布
        - » P(wj| zk)

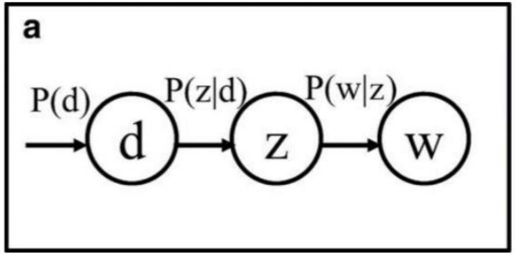


PLSA模型

- PLSA模型及EM算法
  - 在这里单词是可观察的
  - 而主题是不可观察的，是隐变量



PLSA模型



- PLSA模型及EM算法
  - 文档生成过程的建模Basic Generative Model
    - 以 P(d) 选择文档 d
    - 在此基础上，以P(z|d) 的概率选择某个隐藏的(latent)主题
    - 在此基础上，以P(w|z) 的概率生成一个单词w
  - 于是有词项-文档的联合分布模型(Joint Probability Model)如下

$$P(d,w) = P(d)P(w|d)$$

其中 $$P(w|d) = \sum_{z \in \mathbb{Z}} P(w|z)P(z|d)$$

PLSA模型

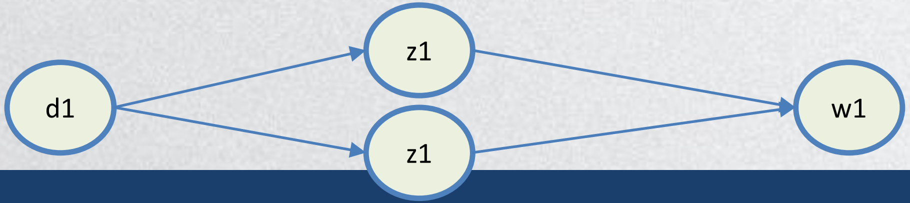
- PLSA模型及EM算法
  - 文档，表示为在各个话题上的概率分布
    - 比如θ_{ik}
    - di = (z1, . . . , zn)
    - d1 = (0.5, 0.3, 0.2)
    - 这个话题空间，是一个概率分布表示的隐藏的语义空间(Probabilistic Latent Semantic Space)
  - 话题，则表示为在各个单词撒谎那个的概率分布
    - 比如ϕ_{kj}
    - zk = (w1, . . . , wm)
    - z1 = (0.3, 0.1, 0.2, 0.3, 0.1)

PLSA模型

具体推导
本PPT末尾给出

- PLSA模型及EM算法
  - 现在已知文档-词项矩阵，如何对上述两个分布进行求解，寻找话题?
  - EM算法
  - E-Step：估计概率$P(z_k|d_i,w_j)$，具体为
    - $P(z_k|d_i,w_j) = \frac{P(z_k|d_i)P(w_j|z_k)}{\sum_{k=1}^{K} P(z_k|d_i)P(w_j|z_k)} = \frac{\theta_{ik}\phi_{kj}}{\sum_{k=1}^{K} \theta_{ik}\phi_{kj}}$
  - 在这个步骤中，假设所有的$P(z_k|d_i)$和$P(w_j|z_k)$都是已知的
    - 刚开始时可以随机地对其赋值
      - 后面的迭代过程中，每轮都能够从M步骤得到这些参数值

该公式可以理解为d_i和w_j的共现中，在不同话题上的概率分布



PLSA模型

- PLSA模型及EM算法
  - 现在已知文档-词项矩阵，如何对上述两个分布进行求解，寻找话题?
  - EM算法
  - M-Step：M-Step更新参数θ_{ik}和ϕ_{kj}，具体为
    - $\theta_{ik} = P(z_k|d_i) = \frac{\sum_{j=1}^{M} n(d_i,w_j)P(z_k|d_i,w_j)}{n(d_i)}$
    - 其中，$n(d_i,w_j)$，表示词项w_j在文档d_i中的词频，$n(d_i)$表示文档d_i中词项的总数，显然有
      - $n(d_i) = \sum_{j=1}^{M} n(d_i,w_j)$
    - 直观地理解，该公式表示，在给定d_i的情况下，z_k的条件概率是多少，即文档d_i在各个z_k上的分配比例是多少

这里暂且以直观方式理解，后面再推导公式

- 可以查看d_i的词项总数，看看里面有多大比例是和z_k相关的
- 换句话说，这个公式的分子，表示d_i和z_k都指定的情况下，所有的数量分配里面，上述式子的分子(只能)通过对词项w_j进行汇总，找出和z_k相关的比例
- 每个数量分配项的形式为$n(d_i,w_j)P(z_k|d_i,w_j)$，表示d_i和w_j关联的数量里，给z_k分配的部分是多少

PLSA模型

- PLSA模型及EM算法
  - 现在已知文档-词项矩阵，如何对上述两个分布进行求解，寻找话题?
  - EM算法
  - M-Step：M-Step更新参数 θ_{ik} 和 ϕ_{kj} ，具体为
    - $\phi_{kj} = P(w_j|z_k) = \frac{\sum_{i=1}^{N} n(d_i,w_j)P(z_k|d_i,w_j)}{\sum_{m=1}^{M}\sum_{i=1}^{N} n(d_i,w_m)P(z_k|d_i,w_m)}$
    - 该公式表示，在给定z_k的情况下，w_j的条件概率是多少，即主题z_k在各个w_j上的分配比例是多少

这里暂且以直观方式理解，后面再推导公式

- 分子表示，z_k和w_j都指定了，于是所有的数量分配里面，所有的数量分配项只能通过对文档d_i进行汇总
- 而分母里面，只有z_k指定了，那么所有的数量分配项，可以通过d_i和w_m进行汇总

PLSA模型

PLSA模型

- PLSA模型实例
  - 假设有如下文档
  - (为了说明问题人工造的文档集)
  - 文档数量为N=5个

1.apple apple apple apple apple apple apple apple apple banana banana grape
2.apple apple apple apple apple apple apple apple banana banana grape grape car
3.grape grape grape car car car truck truck truck truck train train train train train train train train
4.banana banana car car truck truck truck truck train train train train train train train
5.apple apple grape grape train train train

PLSA模型

- PLSA模型实例
  - 假设有如下文档
  - (为了说明问题人工造的文档集)

1.apple apple apple apple apple apple apple apple apple banana banana grape
2.apple apple apple apple apple apple apple apple banana banana grape grape car
3.grape grape grape car car car truck truck truck truck train train train train train train train train
4.banana banana car car truck truck truck truck train train train train train train train
5.apple apple grape grape train train train

  - 单词数量(即字典大小为M=6个)
  - Apple banana grape car truck train

PLSA模型

- PLSA模型实例
  - 假设有如下文档
  - (为了说明问题人工造的文档集)

1.apple apple apple apple apple apple apple apple apple banana banana grape
2.apple apple apple apple apple apple apple apple banana banana grape grape car
3.grape grape grape car car car truck truck truck truck train train train train train train train train
4.banana banana car car truck truck truck truck train train train train train train train
5.apple apple grape grape train train train

  - 单词数量(即字典大小为M=6个)
    - Apple banana grape car truck train

通过观察，我们人工发现两个主题：水果和车辆

PLSA模型

- PLSA模型实例
  - 假设有如下文档
  - (为了说明问题人工造的文档集)

1.apple apple apple apple apple apple apple apple apple banana banana grape
2.apple apple apple apple apple apple apple apple banana banana grape grape car
3.grape grape grape car car car truck truck truck truck train train train train train train train train
4.banana banana car car truck truck truck truck train train train train train train train
5.apple apple grape grape train train train

  - 现在设定Topic数量K=2，用算法发现
    - 每个文档的话题分布
    - 以及每个话题的单词分布

PLSA模型

- PLSA模型实例
  - 假设有如下文档
    - 建立文档-词项矩阵

1.apple apple apple apple apple apple apple apple apple banana banana grape
2.apple apple apple apple apple apple apple apple banana banana grape grape car
3.grape grape grape car car car truck truck truck truck train train train train train train train train
4.banana banana car car truck truck truck truck train train train train train train train
5.apple apple grape grape train train train

词项

文档

	Apple	banana	grape	car	truck	train
1	9.000	2.000	1.000	0.000	0.000	0.000
2	8.000	3.000	2.000	1.000	0.000	0.000
3	0.000	0.000	3.000	3.000	4.000	8.000
4	0.000	2.000	0.000	2.000	4.000	7.000
5	2.000	0.000	1.000	1.000	0.000	3.000

PLSA模型

- PLSA模型实例
  - 假设有如下文档
  - 初始化sita，即各个文档的话题分布

主题

文档

sita	z1	z2
d1	0.60	0.40
d2	0.40	0.60
d3	0.40	0.60
d4	0.60	0.40
d5	0.50	0.50

PLSA模型

- PLSA模型实例
  - 假设有如下文档
  - 初始化phi，即各个话题的单词分布

单词

主题

phi	w1	w2	w3	w4	w5	w6
z1	0.40	0.40	0.60	0.60	0.70	0.70
z2	0.60	0.60	0.40	0.40	0.30	0.30

每一行应该规范化一下

PLSA模型

- PLSA模型实例
  - 假设有如下文档
  - 更新p(zk|di,wj)

$$P(z_k|d_i,w_j)=\frac{P(z_k|d_i)P(w_j|z_k)}{\sum_{k=1}^{K}P(z_k|d_i)P(w_j|z_k)}=\frac{\theta_{ik}\phi_{kj}}{\sum_{k=1}^{K}\theta_{ik}\phi_{kj}}$$

比如$P(z_1|d_1,w_1)=\frac{\theta_{11}\phi_{11}}{\theta_{11}\phi_{11}+\theta_{12}\phi_{21}}$ $P(z_2|d_1,w_1)=\frac{\theta_{12}\phi_{21}}{\theta_{11}\phi_{11}+\theta_{12}\phi_{21}}$

sita	z1	z2
d1	0.60	0.40
d2	0.40	0.60
d3	0.40	0.60
d4	0.60	0.40
d5	0.50	0.50

phi	w1	w2	w3	w4	w5	w6
z1	0.40	0.40	0.60	0.60	0.70	0.70
z2	0.60	0.60	0.40	0.40	0.30	0.30

$P(z1\|d1,w1)$	$P(z1\|d1,w2)$	$P(z1\|d1,w3)$	$P(z1\|d1,w4)$	$P(z1\|d1,w5)$	$P(z1\|d1,w6)$
$P(z1\|d2,w1)$	$P(z1\|d2,w2)$	$P(z1\|d2,w3)$	$P(z1\|d2,w4)$	$P(z1\|d2,w5)$	$P(z1\|d2,w6)$
$P(z1\|d3,w1)$	$P(z1\|d3,w3)$	$P(z1\|d3,w3)$	$P(z1\|d3,w4)$	$P(z1\|d3,w5)$	$P(z1\|d3,w6)$
$P(z1\|d4,w1)$	$P(z1\|d4,w4)$	$P(z1\|d4,w3)$	$P(z1\|d4,w4)$	$P(z1\|d4,w5)$	$P(z1\|d4,w6)$
$P(z1\|d5,w1)$	$P(z1\|d5,w5)$	$P(z1\|d5,w3)$	$P(z1\|d5,w4)$	$P(z1\|d5,w5)$	$P(z1\|d5,w6)$
$P(z2\|d1,w1)$	$P(z2\|d1,w2)$	$P(z2\|d1,w3)$	$P(z2\|d1,w4)$	$P(z2\|d1,w5)$	$P(z2\|d1,w6)$
$P(z2\|d2,w1)$	$P(z2\|d2,w2)$	$P(z2\|d2,w3)$	$P(z2\|d2,w4)$	$P(z2\|d2,w5)$	$P(z2\|d2,w6)$
$P(z2\|d3,w1)$	$P(z2\|d3,w3)$	$P(z2\|d3,w3)$	$P(z2\|d3,w4)$	$P(z2\|d3,w5)$	$P(z2\|d3,w6)$
$P(z2\|d4,w1)$	$P(z2\|d4,w4)$	$P(z2\|d4,w3)$	$P(z2\|d4,w4)$	$P(z2\|d4,w5)$	$P(z2\|d4,w6)$
$P(z2\|d5,w1)$	$P(z2\|d5,w5)$	$P(z2\|d5,w3)$	$P(z2\|d5,w4)$	$P(z2\|d5,w5)$	$P(z2\|d5,w6)$

0.500	0.500	0.692	0.692	0.778	0.778
0.308	0.308	0.500	0.500	0.609	0.609
0.308	0.308	0.500	0.500	0.609	0.609
0.500	0.500	0.692	0.692	0.778	0.778
0.400	0.400	0.600	0.600	0.700	0.700
0.500	0.500	0.308	0.308	0.222	0.222
0.692	0.692	0.500	0.500	0.391	0.391
0.692	0.692	0.500	0.500	0.391	0.391
0.500	0.500	0.308	0.308	0.222	0.222
0.600	0.600	0.400	0.400	0.300	0.300

PLSA模型

$$\theta_{ik} = P(z_k|d_i) = \frac{\sum_{j=1}^{M} n(d_i,w_j)P(z_k|d_i,w_j)}{n(d_i)}$$

比如$\theta_{12} =$

$$\frac{n(d_1,w_1)P(z_2|d_1,w_1)+n(d_1,w_2)P(z_2|d_1,w_2)+n(d_1,w_3)P(z_2|d_1,w_3)+n(d_1,w_4)P(z_2|d_1,w_4)+n(d_1,w_5)P(z_2|d_1,w_5)+n(d_1,w_6)P(z_2|d_1,w_6)}{n(d_1)}$$

– 更新sita

12.000	9.000	2.000	1.000	0.000	0.000	0.000
14.000	8.000	3.000	2.000	1.000	0.000	0.000
18.000	0.000	0.000	3.000	3.000	4.000	8.000
15.000	0.000	2.000	0.000	2.000	4.000	7.000
7.000	2.000	0.000	1.000	1.000	0.000	3.000

P(z1|d1,w1)	P(z1|d1,w2)	P(z1|d1,w3)	P(z1|d1,w4)	P(z1|d1,w5)	P(z1|d1,w6)
P(z1|d2,w1)	P(z1|d2,w2)	P(z1|d2,w3)	P(z1|d2,w4)	P(z1|d2,w5)	P(z1|d2,w6)
P(z1|d3,w1)	P(z1|d3,w3)	P(z1|d3,w3)	P(z1|d3,w4)	P(z1|d3,w5)	P(z1|d3,w6)
P(z1|d4,w1)	P(z1|d4,w4)	P(z1|d4,w3)	P(z1|d4,w4)	P(z1|d4,w5)	P(z1|d4,w6)
P(z1|d5,w1)	P(z1|d5,w5)	P(z1|d5,w3)	P(z1|d5,w4)	P(z1|d5,w5)	P(z1|d5,w6)
P(z2|d1,w1)	P(z2|d1,w2)	P(z2|d1,w3)	P(z2|d1,w4)	P(z2|d1,w5)	P(z2|d1,w6)
P(z2|d2,w1)	P(z2|d2,w2)	P(z2|d2,w3)	P(z2|d2,w4)	P(z2|d2,w5)	P(z2|d2,w6)
P(z2|d3,w1)	P(z2|d3,w3)	P(z2|d3,w3)	P(z2|d3,w4)	P(z2|d3,w5)	P(z2|d3,w6)
P(z2|d4,w1)	P(z2|d4,w4)	P(z2|d4,w3)	P(z2|d4,w4)	P(z2|d4,w5)	P(z2|d4,w6)
P(z2|d5,w1)	P(z2|d5,w5)	P(z2|d5,w3)	P(z2|d5,w4)	P(z2|d5,w5)	P(z2|d5,w6)

0.500	0.500	0.692	0.692	0.778	0.778
0.308	0.308	0.500	0.500	0.609	0.609
0.308	0.308	0.500	0.500	0.609	0.609
0.500	0.500	0.692	0.692	0.778	0.778
0.400	0.400	0.600	0.600	0.700	0.700
0.500	0.500	0.308	0.308	0.222	0.222
0.692	0.692	0.500	0.500	0.391	0.391
0.692	0.692	0.500	0.500	0.391	0.391
0.500	0.500	0.308	0.308	0.222	0.222
0.600	0.600	0.400	0.400	0.300	0.300

sita	z1	z2
d1	0.516	0.484
d2	0.349	0.651
d3	0.572	0.428
d4	0.729	0.271
d5	0.586	0.414

PLSA模型

$$\phi_{kj} = P(w_j|z_k) = \frac{\sum_{i=1}^{N} n(d_i,w_j)P(z_k|d_i,w_j)}{\sum_{m=1}^{M}\sum_{i=1}^{N} n(d_i,w_m)P(z_k|d_i,w_m)}$$

比如$\phi_{23} = \frac{n(d_1,w_3)P(z_2|d_1,w_3)+n(d_2,w_3)P(z_2|d_2,w_3)+n(d_3,w_3)P(z_2|d_3,w_3)+n(d_4,w_3)P(z_2|d_4,w_3)+n(d_5,w_3)P(z_2|d_5,w_3)}{W_1+W_2+W_3+W_4+W_5+W_6}$

W_1=$n(d_1,w_1)P(z_2|d_1,w_1) + n(d_2,w_1)P(z_2|d_2,w_1) + n(d_3,w_1)P(z_2|d_3,w_1) + n(d_4,w_1)P(z_2|d_4,w_1) + n(d_5,w_1)P(z_2|d_5,w_1)$

W_2=$n(d_1,w_2)P(z_2|d_1,w_2) + n(d_2,w_2)P(z_2|d_2,w_2) + n(d_3,w_2)P(z_2|d_3,w_2) + n(d_4,w_2)P(z_2|d_4,w_2) + n(d_5,w_2)P(z_2|d_5,w_2)$ ...

12.000	9.000	2.000	1.000	0.000	0.000	0.000
14.000	3.000	3.000	2.000	1.000	0.000	0.000
18.000	0.000	0.000	3.000	3.000	4.000	8.000
15.000	0.000	2.000	0.000	2.000	4.000	7.000
7.000	2.000	0.000	1.000	1.000	0.000	3.000

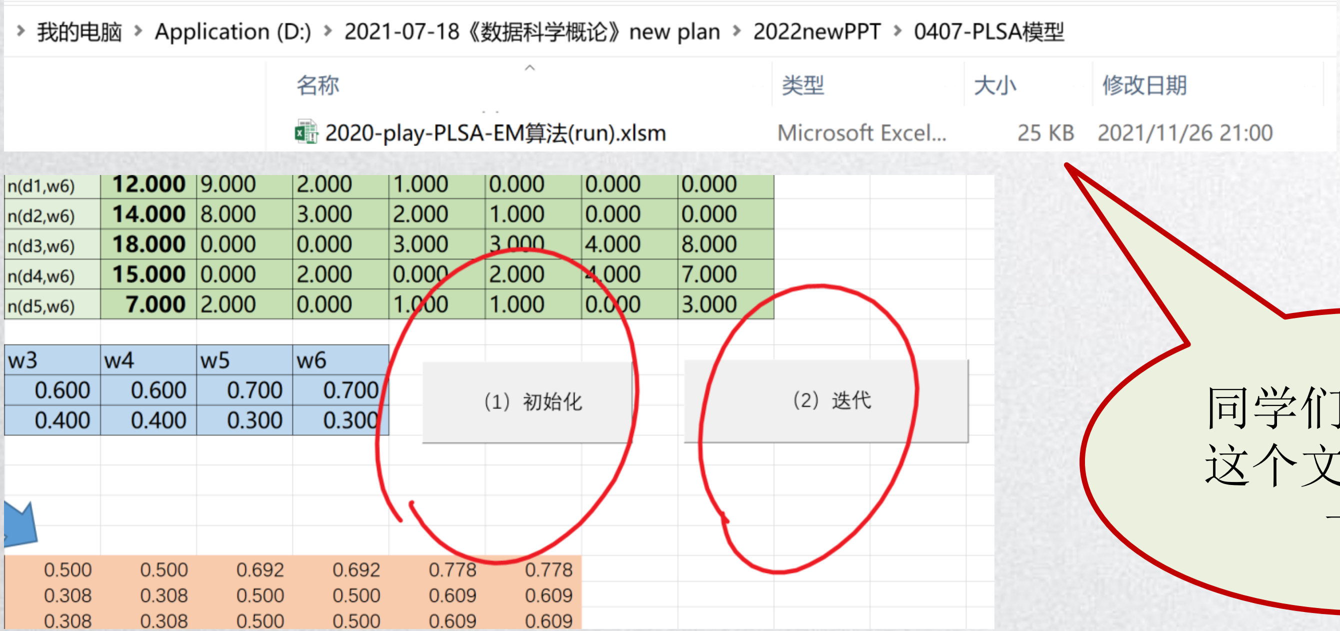
P(z1|d1,w1)	P(z1|d1,w2)	P(z1|d1,w3)	P(z1|d1,w4)	P(z1|d1,w5)	P(z1|d1,w6)
P(z1|d2,w1)	P(z1|d2,w2)	P(z1|d2,w3)	P(z1|d2,w4)	P(z1|d2,w5)	P(z1|d2,w6)
P(z1|d3,w1)	P(z1|d3,w3)	P(z1|d3,w3)	P(z1|d3,w4)	P(z1|d3,w5)	P(z1|d3,w6)
P(z1|d4,w1)	P(z1|d4,w4)	P(z1|d4,w3)	P(z1|d4,w4)	P(z1|d4,w5)	P(z1|d4,w6)
P(z1|d5,w1)	P(z1|d5,w5)	P(z1|d5,w3)	P(z1|d5,w4)	P(z1|d5,w5)	P(z1|d5,w6)
P(z2|d1,w1)	P(z2|d1,w2)	P(z2|d1,w3)	P(z2|d1,w4)	P(z2|d1,w5)	P(z2|d1,w6)
P(z2|d2,w1)	P(z2|d2,w2)	P(z2|d2,w3)	P(z2|d2,w4)	P(z2|d2,w5)	P(z2|d2,w6)
P(z2|d3,w1)	P(z2|d3,w3)	P(z2|d3,w3)	P(z2|d3,w4)	P(z2|d3,w5)	P(z2|d3,w6)
P(z2|d4,w1)	P(z2|d4,w4)	P(z2|d4,w3)	P(z2|d4,w4)	P(z2|d4,w5)	P(z2|d4,w6)
P(z2|d5,w1)	P(z2|d5,w5)	P(z2|d5,w3)	P(z2|d5,w4)	P(z2|d5,w5)	P(z2|d5,w6)

0.500	0.500	0.692	0.692	0.778	0.778
0.308	0.308	0.500	0.500	0.609	0.609
0.308	0.308	0.500	0.500	0.609	0.609
0.500	0.500	0.692	0.692	0.778	0.778
0.400	0.400	0.600	0.600	0.700	0.700
0.500	0.500	0.308	0.308	0.222	0.222
0.692	0.692	0.500	0.500	0.391	0.391
0.692	0.692	0.500	0.500	0.391	0.391
0.500	0.500	0.308	0.308	0.222	0.222
0.600	0.600	0.400	0.400	0.300	0.300

phi	w1	w2	w3	w4	w5	w6
z1	0.213	0.080	0.104	0.109	0.152	0.341
z2	0.380	0.138	0.108	0.102	0.083	0.189

PLSA模型

- PLSA模型实例
  - 打开excel文件，进行实验



n(d1,w6)	**12.000**	9.000	2.000	1.000	0.000	0.000	0.000
n(d2,w6)	**14.000**	8.000	3.000	2.000	1.000	0.000	0.000
n(d3,w6)	**18.000**	0.000	0.000	3.000	3.000	4.000	8.000
n(d4,w6)	**15.000**	0.000	2.000	0.000	2.000	4.000	7.000
n(d5,w6)	**7.000**	2.000	0.000	1.000	1.000	0.000	3.000

w3	w4	w5	w6
0.600	0.600	0.700	0.700
0.400	0.400	0.300	0.300

0.500	0.500	0.692	0.692	0.778	0.778
0.308	0.308	0.500	0.500	0.609	0.609
0.308	0.308	0.500	0.500	0.609	0.609

同学们一起打开这个文件，运行一下

PLSA模型

- PLSA模型实例
  - 打开excel文件，进行实验
  - 迭代结果与解读
    - Sita，文档在主题上的分布

1.apple apple apple apple apple apple apple apple apple banana banana grape
2.apple apple apple apple apple apple apple apple banana banana grape grape car
3.grape grape grape car car car truck truck truck truck train train train train train train train train
4.banana banana car car truck truck truck truck train train train train train train train
5.apple apple grape grape train train train

最后的结果

sita	z1	z2
d1	0.000	1.000
d2	0.000	1.000
d3	1.000	0.000
d4	1.000	0.000
d5	0.637	0.363

单一主题(水果)

单一主题(车辆)

混合主题(车辆+水果)

PLSA模型

- PLSA模型实例
  - 打开excel文件，进行实验
  - 迭代结果与解读
    - phi，主题在单词上的分布

phi	W1 apple	W2 banana	W3 grape	W4 car	W5 truck	W6 train
z1	0.000	0.053	0.096	0.157	0.214	0.481
z2	0.666	0.175	0.120	0.039	0.000	0.000

最后的结果

PLSA模型

PLSA模型

- PLSA模型实例：Python实现

我的电脑 › Application (D:) › 2021-07-18《数据科学概论》new plan › 2022newPPT › 0407-PLSA模型

名称	类型	大小	修改日期
01plsa_demo.py	Python File	7 KB	2021/11/27 17:22

PLSA模型

- PLSA模型实例：Python实现

1.apple apple apple apple apple apple apple apple apple banana banana grape
2.apple apple apple apple apple apple apple apple banana banana grape grape car
3.grape grape grape car car car truck truck truck truck train train train train train train train train
4.banana banana car car truck truck truck truck train train train train train train train
5.apple apple grape grape train train train

	apple	banana	grape	car	truck	train
D1						
D2						
D3						
D4						
d5						

```
N 5
M 6
word2id {'apple': 0, 'banana': 1, 'grape': 2, 'car': 3, 'truck': 4, 'train': 5}
id2word {0: 'apple', 1: 'banana', 2: 'grape', 3: 'car', 4: 'truck', 5: 'train'}
X [[9 2 1 0 0 0]
 [8 2 2 1 0 0]
 [0 0 3 3 4 8]
 [0 2 0 2 4 7]
 [2 0 2 0 0 3]]
```

PLSA模型

- PLSA模型实例：Python实现

```
topic 0 apple banana grape car
topic 1 train truck car grape
```

初始值

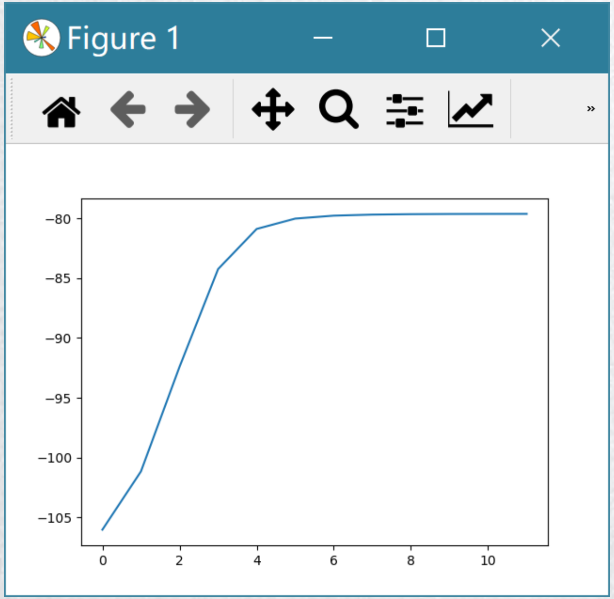
```
lamda [[0.40224784 0.59775216]
 [0.60273076 0.39726924]
 [0.14418129 0.85581871]
 [0.33180906 0.66819094]
 [0.26343995 0.73656005]]
theta [[0.23571251 0.1142369  0.03171455 0.21535756 0.19880225 0.20417624]
 [0.23456676 0.15806406 0.01805589 0.16702434 0.22780818 0.19448078]]
```

迭代后的取值

```
lamda [[ 1.000  0.000]
 [ 1.000  0.000]
 [ 0.000  1.000]
 [ 0.000  1.000]
 [ 0.426  0.574]]
theta [[ 0.679  0.143  0.142  0.036  0.000  0.000]
 [ 0.000  0.054  0.109  0.135  0.216  0.486]]
```

PLSA模型

- PLSA模型实例：Python实现
  - 目标函数的变化(最大化LogLikelihood)
  - 请参考后文的公式推导



PLSA模型

PLSA模型

- PLSA模型的优缺点
- 优点
  - Results have a clear probabilistic interpretation
  - Allows for model combination
  - Problem of polysemy (一词多义)is better addressed
    - PLSA can address synonymy and polysemy problems by exploring underlying semantic relations beneath the actual occurrences of words

PLSA模型

- PLSA模型的优缺点
- 缺点
  - Potentially higher computational complexity
  - EM algorithm gives local maximum
  - Prone to overfitting
    - Solution: Tempered EM
  - Not a well defined generative model for new documents
    - Solution: Latent Dirichlet Allocation

PLSA模型

PLSA模型

- PLSA模型的公式推导
  - 寻找文档的隐藏的话题分布
  - 以使如下目标函数最大化

最大化

Maximize:

$$\mathcal{L} = \sum_{d \epsilon D} \sum_{w \epsilon W} n(d,w) \log P(d,w)$$

Documents →latent topics → words

$$P(d,w) = P(d)P(w|d)$$

其中 $P(w|d) = \sum_{z \,\epsilon\, \mathbb{Z}} P(w|z)P(z|d)$

相当于最小化(1)和(2)之间的交叉熵
(1)单词的经验分布empirical distribution of words n(d,w)
(2)模型给出的分布p(d,w)

最小化

$$\mathrm{CrossEntropy(p,q)} = -\sum_{\mathrm{x \in X}} \mathrm{p(x)\log q(x)}$$

PLSA模型

- PLSA模型的公式推导

我的电脑 › Application (D:) › 2021-07-18《数据科学概论》new plan › 2022newPPT › 0407-PLSA模型

名称	类型	大小	修改日期
2020-new-PLSA EM算法推导.pdf	Adobe Acrobat...	475 KB	2021/11/29 18:29

该文件给出PLSA模型的EM算法的
推导过程