

人工智能之机器学习

隐含狄利克雷分布 (Latent Dirichlet Allocation)

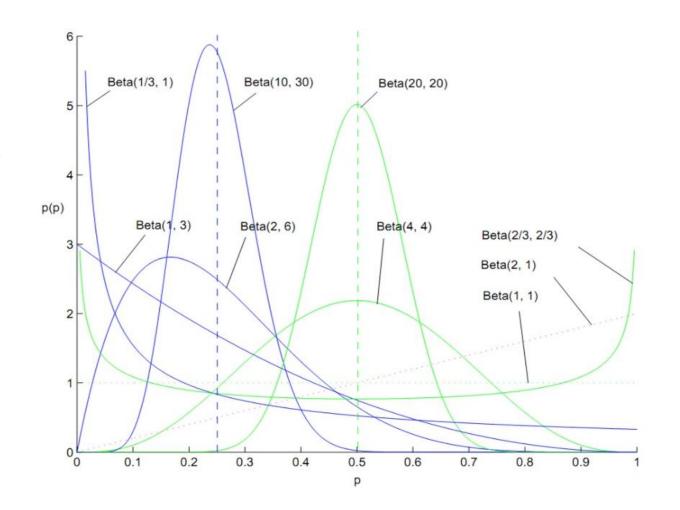
上海育创网络科技股份有限公司

主讲人: 赵翌臣

LDA基础知识——beta分布



- Beta分布:
 - Beta(a,b)
 - · 以抛硬币为例,抛了a+b次,参数a为出现了正面的次数,参数b为出现了反面的次数。
 - Beta分布是估计硬币正面向上概率的分布,横轴表示出现正面的概率,纵轴表示出现正面概率的可能性。



LDA基础知识——狄利克雷分布



- 狄利克雷分布:
 - 狄利克雷分布就是beta分布的推广情形。现在抛出的结果不是只有正和反两种了,可能有很多种,就像抛骰子可能有6种。



- 作者
 - LDA是由吴恩达等人于2003年提出的一种主题模型,是无监督学习模型
 - 论文: Latent Dirichlet Allocation
- 应用
 - 文档分类、降维
- 思想
 - 模型构建
 - 训练数据: 文档集, 目标函数稍后谈
 - 模型使用
 - 文本分类:每个文档以不同的概率分别属于每一个主题,当有新文档待预测时,LDA可以算出这个文档属于每个主题的概率,下面举例子说明一下
 - 降维: 主题是个抽象的概念,无法直接理解,但可以作为特征



•第一步准备训练数据:这里准备了两类文档(航空类、体育类),并带有类别y,但是训练LDA模型时不用y,y只是用来评估LDA文档分类的效果

0,宇航 学报 journal of astronautics vol no 空间 辐射 散热器 介质 辐射 特性 辐射 传热 阮立 发射 各向同性 散射 介质 辐射 传递 系数 计算 空间 辐射 散热器 辐射 文献 比较 表明 本文 导出 等 实验 次数 主题词 空间 辐射 散热器 电磁 散射 理论 辐射 辐射 传递 系数 radiative properties school of energy science and engineering harbin institute of technology harbin abstra the redistribution of radiative energy in the case of isotropic scattering is investi the transient radiative heat transfer of droplet filled layer in space radiator is so finally the effects of size or size distribution of droplet which are created by gene experiment can be reduced in design of space radiator key words space radiator drople 使用 散热器 mattick hertzberg 提出 可以 空间 喷洒 通过 瞬间 辐射 冷却 然后 加以 回收 文献 数值 模拟 本文 导出 两侧 透明 界面 吸收 发射 各向同性 散射 介质 辐射 传递 系数 空间 辐射 散射 方向 范围 相对 因此 方向 变化 忽略 面上 温度 均匀 ti 由于 空间 散热 冷却 因此 温度 半径 组成 发生器 接受 距离 并且 方向 两侧 辐射 边界 条件 相同 两侧 环境 辐射 散热 因此 向上 光学 厚度 方程 ti 位于 代表 环境 表面 之间 温度 分别 ts ts 厚度 方向 分成 nm 节点 控制 表示 时间 间隔 折射率 表面 反射率 波长 变化 表示 nb 表示 模型 区域 透明 界面 介质 辐射 引起 控制 辐射源 ri 介质 控制 控制 控制 表面 单色 辐射 传递 系数 求解 关键 辐射源 辐射源 关键 单色 辐射 传递 系数 计算 分别 粒子 单色 衰减 散射 吸收 系数 相应 单个 粒子 单色 衰减 散射 吸收 因子 粒子 粒子 半 粒子 尺度 参数 粒子 散射 可以 利用 简化 rayleigh 散射 电磁 理论 计算 粒子 尺度 参数 可以 利 反照 相同 因此 必须 计算 平均 散射 反照 本文 假定 散射 各向同性 辐射 特性 计算 如果 其它 吸收 吸收 系数 因此 粒子 衰减 系数 衰减 系数 分布 发生器 确定 设计 发生器 遵循 回收率 损耗 分布 🤋 半径 范围 常量 分别 下列 数据 分布 曲线 符合 modifided gamma 分布 均匀 粒子 分布 曲线 case 反照 可用 mie 散射 理论 计算 程序 单色 衰减 散射 吸收 因子 分布 函数 得到 辐射 特性 参数 符 特性 单色 辐射 传递 系数 反射 吸收 发射 介质 辐射 传递 系数 vivj visj 推导 过程 表达式 参阅 考虑 反射 并且 具有 相对性 散射 衰减 系数 此时 吸收 发射 介质 辐射 传递 系数 表示 能量 传递 散射 考虑 吸收 散射 计算 结果 验证 外层 空间 微粒 介质 只能 热辐射 方式 物理 模型 化为 透明 积分 方程 数值 本文 计算 点数 控制 nm 单位 光学 厚度 计算 结果 比较 可以 看出 结果 非常 理想 nm xo 散射 反照 散射 反照 散射 反照 本文 计算 结果 本文 计算 收敛 条件 tnm tnm eps 本文 计算 结果 比较 透明 界面 各向同性 粒子 散射 介质 发射率 eps eps 散射 反照 散射 反照 本文 计算 结果 本文 计算 结果 本文 计算 结果 nm 文献 nasa 资助 空间 辐射 散热器 辐射 进行 数值 模拟 建立 合理 数学 模型 本文 米氏 电磁 散射 数学 模型 计算 空间 辐射 散热器 辐射 文献 比较 表明 本文 导出 辐射 传递 系数 计算 方法 正确 次数 介质 折射率 前提 采用 其它 空间 辐射 散热器 进行 数值 模拟 国家 杰出 青年 科学 基金 国家 radiators for heat rejection in space journal of energy presler coles diem kirsop whi systems for spacecraft thermal control journal of spacecraft rockets siegel transient linear radiative cooling int heat mass trans 刘晓彦 mie 散射 程序 改进 哈尔滨 丁业 大学 transient radiative conductive heat transfer in flat glasses submitted to temperature conductive heat transfer in an absorbing emitting and scattering medium int heat mass 0,航空 工程 维修 aviation engineering maintenance no 飞机 工艺 装备 设计 制造 一直 飞机 问题 方法 途径 飞机 装配 智能化 设计 应用 研究 design automation of aircraft assembling 工艺 准备 周期 国内 飞机 制造厂 情况 来看 接受 新机 研制 改型 任务 工装 设计 制造 采用 分工

才能 政治 思想 道德 体质 气质 卫生 等等 主要 文化 素质 思想 素质 审美 素质 文化 素质 包括 教育 科技 卫生 体育 方面 思想 素质 包括 理想 道德 情操 纪律 方面 至于 审美 素质 美的 意识 观念 标准 观察 感觉 衡量 扩展 而且 有力 促进 科学技术 发展 现在 世界 发达国家 如果 没有 高度 发达 教育 不会 高科技 文明 昌盛 不能 带来 经济 高速 发展 思想 素质 理想 道德 激励 人们 主动性 创造性 促进 人们 内心 自律 成为 人们 前进 精神 力量 从而 空前 提高 我国 千年 从未 有的 现象 思想 觉悟 来说 拥护 共产党 共产党 自己 行动 在建 设有 中国 特色 社会主义 成绩 辉煌 正在 小康 共同 富裕 目标 迈进 祖国 明天 充满 远大 理想 文化 素质 来说 广大 人民 群众 文化 水平 提高 就是 农村 建设 学校 图书馆 文化站 俱乐部 农民 书报 电影 而且 彩电 摄影 集邮 旅游 进入 寻常 百姓家 过去 难以 想像 事情 广大 青少年 来说 大都 机会 学校 读书 发扬 爱国主义 精神 广大 人民 素质 不同 层次 有所 提高 欣欣向荣 景象 但是 另一方面 我们 应当 看到 素质 方面 存在 问题 有的 还是 严重 比如说 文化 程度 存在 识字 识字 文盲 半文盲 阻碍 素质 提高 如同 发展 感到 精神 空處 苦闷 因此 赌博 迷信 歪风 乘机 蔓延 当然 影响 人们 素质 提高 别是 我国 现在 还有 人口 贫困 地区 尚未 解决 温饱 问题 那里 人民 生活 更加 贫困 更加 影响 素质 提高 比如说 我们 党风 社会 风气 存在 风气 当然 阴碍 影响 人们 素质 提高 虽然 我们 扫黄 打非 斗争 胜利 开展 战果 累累 艰苦 斗争 雲要 持久 深入 进行 下去 并且 加强 文化 市场 管理 文化 市场 繁荣 健康 有序 发展 才能 有效 教育 引导 启发 鼓励 人们 从事 思想 政治 教育 工作 优良 传统 我国 人民 思想 政治 觉悟 我们 优势 现在 历史 时期 认真 继承 发扬 这种 光荣 传统 唱响 主旋律 增强 我们 优势 为此 精神文明 重在 建设 落实 实效 不断 精神 产品 教育 群众 人们 博物馆 俱乐部 体育场 广大 群众 报刊 电影 文化 娱乐 活动 体育 活动 场所 美的 精神 这样 才能 解決 精神 贫乏 不断 高人 文化 素质 例如 攰疆 文化 长廊 工程 年间 全国 地区 文化馆 图书馆 博物馆 农民 俱乐部 起来 文化 网点 流动 文化 服务 结合 起来 丰富 各族 人们 文化 高人 素质 作用 所以 受到 好评 热烈 欢迎 为了 高人 文化 素质 我们 一方面 加强 文化 市场 管理 进行 正确 引导 促使 文化 市场 繁荣 健康 有序 发展 另一方面 新闻 媒体 文学 艺术 负有 重大 职责 报纸 广播 电视 天天 群众 见面 覆盖面 课堂 影响力 科学 理论 武装 正确 舆论 引导 进行 舆论 监督 形成 舆论 气候 发挥 重大 作用 文学 艺术 优秀 作品 鼓舞 有的 艺术 魅力 感染 启迪 潜移默化 移风易俗 高人 素质 具有 不可 代替 力量 因此 我们 新阃 媒体 文学 艺术 寄予 般切 希望 要求 它们 发挥 重大 作用 例如 实施 五个一工程 引导 鼓舞 高人 素质 其中 加强 群众性 精神文明 建设 活动 我们 精神文明 建设 实践 创建 活动 各级 党政 组织 领导 不仅 官传 文教 部内 大力 工会 青年团 工商联 群众组织 力量 调动 起来 充分 发挥 有的 作用 分工 负责 各自 联系 群众 调动 起来 愚昧 作战 千方百计 高人 素质 奋斗 同时 还要 采取 措施 各种 力量 值得 我们 高度 警惕 不能 大意 由于 我们 党中央 领导 正确 我们 信心 能力 文明 建设 搞好 现在 我们 精神文明 建设 总结 过去 经验 发扬 成绩 接受 教训 面对 现实 矛盾 深入 调查 研究 密切 联系 群众 改进 大力 建设 同时 还要 展望 未来 胸怀 雄心壮志 远大 精神文明 建设 推向 台阶 按照 五中全会 提出 计划 远景 目标 今后 年内 我们 文明 建设 都会 生龙活虎 大踏步 前进 我国 社会 生产力 综合国力 人民 生活 台阶 社会主义 精神文明 建设 都会 取得 明显 进展 世纪 中叶 实现 战略 目标 基本 实现 现代化 奠定 坚实 基础 这样 我们 培养 四有 社会主义 新人 整个 中华民族 思想 素质 大大 提高 别是 最近 张家港 璀璨 明珠 出现 全国 各地 引起 强烈 反响 我们 启示 鼓舞 本来 知名 可是 经过 短短的 文明 设有 惊人 成就 张家港 精神 标志 那里 秦质 大大 提高 从而 成为 江南 新加坡 全国 闻名 文明 城市 我们 张家港 看到 社会主义 现代化 雏形 看到 中国 特色 社会主义 缩影 现在 地区 各部 努力 学习

缩短 生产 准备 周期 一直 飞机 制造 工艺学 重要 研究 课题 年代 中期 结合 计算机 輔助 设计 制造 一体化 技术 我国 飞机 制造业 推广 应用 工装 设计 领域 展开 装配 工装 cad cam 软件 fixcad 代表 一系列 工装 设计 软件 研究 开发 提高 我国 工装 设计 自动化 水平 百外 相关 研究 工作 年代 末期 苏联 开始 工装 设计 自动化 水平 百外 相定 工艺 学院 飞行器 制造 工艺 軟好室 提供 研究 报告 报告 首先 引入 功能 外形 概念 飞机 构件 那些 飞机 构造 基准 表面 理论 外形 庙名 正合 直接 图 大家面 骨架 大家面 骨架 水管 定价 可究 工作 年代 末期 苏联 开始 公离 坐你 定位 可以 飞机 生产 准备 各个 茅节 采用 种种 工艺 方法 首先 就是 保证 这些 形状 信息 制造 过程 准确性 结构 装配 过程 空间 位置 正确性 根据 功能 外形 生产 准备 各个 环节 作用 苏联 学者 试图 找到 量化 描述 手段 建立 飞机 制造 整个 生产 准备 工作 飞机 结构 工艺 分解 工装 协调 值至 工套 设计 各个 环节 形式化 描述 以此 形成 他们 设计 自动化 理论 他们 生产 准备 工作 飞机 结构 工艺 分解 工装 协调 值至 工套 设计 各个 环节 原因 计算数 依次 对应



- •第二步:将文本转为词频,注意这里是TF,不是TF-IDF
- •第三步:训练LDA模型

- •注:这里指定了2个主题,用户指定个数也是可以的,这里指定为2的目的是,因为已知是两类文档,我想对照y看一下LDA的效果。
- 效果评估
 - LDA给出的主题概率分布如图

```
[[9.99417172e-01 5.82828207e-04]

-[9.48272130e-01 5.17278701e-02]

[8.36948741e-01 1.63051259e-01]

...

[1.73361605e-03 9.98266384e-01]

[2.72210690e-04 9.99727789e-01]

[3.21560981e-04 9.99678439e-01]
```

第二篇文档属于第一个主题的概率



- 使用y评估后,准确率为0.97196969696969
- 打印主题 print (Ida.components_)

• 6.04629107 代表第一个主题的第一个词的非规范化概率, 0.50931329代表第一个主题的第二个词的非规范化概率,以此类 推。



- 打印每个主题下概率Top50的词语
- 主题1:分布,确定,使用,得到,具有,提高,工艺,故障,信号,要求,条件,状态,问题,不同,应用,实验,通过,性能,本文,误差,材料,工作,叶片,变化,航空,时间,由于,速度,温度,过程,参数,数据,发动机,试验,影响,研究,结果,可以,结构,模型,采用,控制,技术,设计,分析,测量,计算,进行,方法,系统——航空主题
- 主题2: 水平,方面,目标,训练,过程,国家,作用,自己,方法,可以,基础,心理,需要,运动,重要,理论,要求,建设,他们,经济,我国,能力,知识,内容,中国,学科,管理,培养,科学,改革,提高,思想,进行,问题,文化,学习,研究,我们,工作,素质,课程,教师,活动,学校,社会,教学,学生,发展,体育,教育——体育主题

LDA介绍



- 基础概念
 - •词(word): 待处理数据的基本离散单元,如"hello"、"你好"
 - 文档(document):是待处理的数据对象,由一组词组成,这些词在文档中不计顺序,例如一篇论文,这样的表示方式为"词袋"(bag-of-words)
 - 主题/话题(topic):表示一个概念,具体表示为一系列相关的词,以及它们在该概念下出现的概率。

LDA介绍



- LDA构建目标函数的思想
 - 训练数据(不携带类别信息) $W=\{w_1, w_2, ..., w_T\}$
 - LDA模型认为你这些数据是由我生成的,我有我的参数α和η,我可以量化出生成这样数据的概率

$$LL(oldsymbol{lpha}, oldsymbol{\eta}) = \sum_{t=1}^T \ln p(oldsymbol{w}_t \mid oldsymbol{lpha}, oldsymbol{\eta})$$

• 使用极大似然估计法,可以得到最优的α和η。





•一个话题就像一个箱子,里面装着在这个概念下词的出现频率。

话题









- 不妨假定数据集中一共包含K个话题和T篇文档,文档中的词来自一个包含N个词的词典。我们用T个N维向量W={ w_1 , w_2 , ..., w_T }表示数据集(即文档集合),K个N维向量 β_k (k=1,2,...,K)表示话题。
- 其中w_t ∈ R^N的第n个分量w_{t,n}表示文档t中词n的词频
- 其中 $\beta_t \in R^N$ 的第n个分量 $\beta_{t,n}$ 表示话题k中词n的词频

LDA介绍

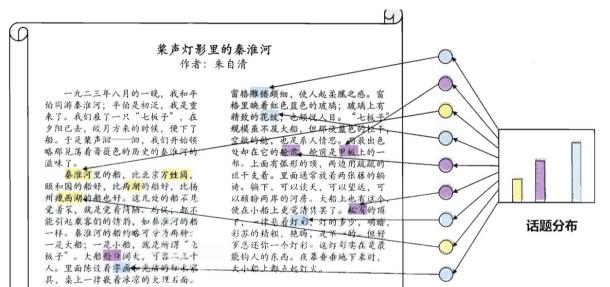
文档



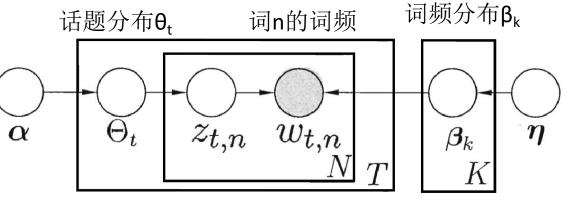
• LDA认为每篇文档有多个话题,用向量 $\theta_t \in R^k$ 表示文档t中所包含的每个话题的比例, $\theta_{t,k}$ 表示文档t中包含话题k的比例,进而通过下面的步骤由话题生成文档t:

• (1) 根据参数 α 得到一个话题分布 θ_t ~Dirichlet(α);

- (2) 按照如下步骤生成文档中的N个词:
 - (a) 根据 θ_t 进行话题指派,得到文档t中的词n的话题 $z_{t,n}$;
 - (b) 根据指派的话题所对应的词频分布β_k随机采样生成词



话题指派



$$p(\mathbf{W}, \mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\Theta} \mid \boldsymbol{\alpha}, \boldsymbol{\eta}) = \prod_{t=1}^{T} p(\boldsymbol{\Theta}_{t} \mid \boldsymbol{\alpha}) \prod_{i=1}^{K} p(\boldsymbol{\beta}_{k} \mid \boldsymbol{\eta}) \left(\prod_{n=1}^{N} P(w_{t,n} \mid z_{t,n}, \boldsymbol{\beta}_{k}) P(z_{t,n} \mid \boldsymbol{\Theta}_{t}) \right)$$

$$LL(oldsymbol{lpha}, oldsymbol{\eta}) = \sum_{t=1}^T \ln p(oldsymbol{w}_t \mid oldsymbol{lpha}, oldsymbol{\eta})$$

常采用吉布斯采样或变分法进行近似推断

