

自然语言处理

• 第1章 绪论

- 自然语言处理基本概念
- 应用
- 基本问题
 - 语音学问题
 - 形态学问题
 - 语法学问题
 - 语义学问题
 - 语用学问题
- 面临困难
 - 结构歧义
 - 语义歧义
 - 未知语言现象
 - 新的词汇、新的含义、新的用法和语句结构

• 第2章 数学基础

- 概率论相关
 - 概率、最大似然估计、条件概率、贝叶斯公式、二项分布、期望、方差
- 信息论相关
 - 熵、联合熵、条件熵、互信息、相对熵、交叉熵、迷惑度
 - 熵的连锁机制 ppt8-5
 - 噪声信道模型

• 第4章 语料库与语言知识库


- 语料库基本概念
- 分类
 - 按语言种类
 - 按标注类型
- 典型语料库
- 词汇知识库
 - wordNet
 - 按照语义关系组织
 - 四种语义关系
 - HowNet

- 多种语义关系

• 第5章 语言模型

- 语言模型的定义：判断一个语言序列是否是正常语句
- 主要统计语言模型：N-gram
 - 构造方法
 - 应用
 - 音字转换、汉语分词等
 - 模型参数估计（计算题 )
 - 训练语料
 - 最大似然估计
 - 数据稀疏
 - zipf统计定律
 - 数据平滑技术
 - 加法平滑
 - Good-Touring平滑
 - Katz回退平滑
 - 线性插值
- 评价标准
 - 交叉熵或迷惑度
- 统计语言模型的作用

• 第6章 概率图模型

- 常见概率图模型 P104
- 生成式模型与判别式模型 P105
- 贝叶斯网络
 - 有向无环图
 - 三方面问题：表示、推断、学习
- 马尔可夫状态
- 隐马尔科夫模型
 - 三个经典问题
 - 求某观测序列的联合概率
 - 前向算法 ppt22
 - 后向算法 ppt23
 - 求最可能的隐含序列 
 - viterbi搜索算法 ppt24
 - 求最可能的模型参数

- 极大似然法：观测序列、隐序列都已知 ppt25
 - EM(期望最大)算法：观测序列已知、隐序列未知 ppt26
 - Baum-Welch算法（前向后向算法）
 - HMM应用：综合应用考点 ppt26.8
- 条件随机场
 - 三个基本问题
 - 特征的选取、参数训练、解码
 - 条件概率的表示 p128
 - CRF vs HMM ppt30
- 第7章 自动分词与命名实体识别**
 - 汉语自动分词概要
 - 主要难点
 - 歧义切分问题 ppt31, p130 **A**
 - 交集型歧义
 - 求链长
 - 组合型歧义
 - 未登录词识别（集外词OOV） ppt31, p132
 - 分类（ppt：2类；书：4类）
 - 汉语自动分词基本原则
 - 分词性能评价
 - 正确率、召回率、F测度值
 - 自动分词基本算法
 - 基本分类
 - 有词典切分/无词典切分
 - 基于规则的方法/基于统计的方法
 - 最大匹配法
 - 有词典切分
 - 正向\逆向\双向
 - 最少分词法（最短路径法）
 - 基本原理+缺点
 - 基于语言模型的分词方法
 - 基本原理、优缺点分析
 - 由字构词（基于字标注）的分词方法
 - 基本思想、HMM模型建立、优缺点分析
 - 未登录词的识别

- 命名实体的概念
 - 中文姓名 **A**
 - 中文地名
 - 中文机构名
 - 使用CRF模型（书p152）
- 词性标注
 - 面临的问题
 - 词性兼类歧义（四种ppt37）
 - 词性标注方法
 - 基于字符串匹配的字典查找算法
 - 基于统计的算法
 - HMM模型的应用
- 第9章 语义分析**
 - 定义 ppt39
 - 语义歧义
 - 语义消歧
 - 有监督词义消歧
 - 贝叶斯分类：将上下文看作一个无结构词集，整合上下文中众多的词汇信息。（具体原理见ppt40,书p247）
 - 基于互信息的方法：仅仅考虑上下文中的一个信息特征，其可以灵敏地反映上下文结构，但需要谨慎地选取此特征（ppt上没有，书P245）
 - 基于词典的消歧
 - 基于语义定义的消歧
 - 实现算法：书/ppr41
 - 基于义类词典的消歧
 - 实现算法：书/ppt41.7
- 词向量与深度神经网络**
 - 词向量表示 ppt42
 - 定义（字面意思）
 - 方法
 - 独热编码
 - 缺点
 - 词嵌入
 - 基本原理 ppt42.6
 - 基本方法

- CBOW：通过上下文来预测当前词
 - Skip-gram：通过当前词来预测上下文
 - 其他：Glove, ELMo, Bert
- 人工神经网络
 - 基本原理 ppt43.4
- 深度神经网络
 - 卷积神经网络
 - 基本层次
 - 数据输入层
 - 卷积层 **A**
 - 激活层 **A**
 - 池化层 **A**
 - 全连接层
 - 输出层
 - 全连接层+softmax
 - Dropout ppt47
 - 损失函数 ppt47
 - 模型训练过程 ppt47.5
 - 前向传播
 - 后向传播
 - 循环神经网络
 - 网络结构 ppt48.5
 - LSTM
 - 单元结构
 - 遗忘门
 - 输入门
 - 输出门
 - 应用
 - 文本分类
 - 表示方式：①使用最后的隐变量作为句向量；②使用每个词的隐变量拼成句向量
 - 命名实体识别
 - LSTM
 - LSTM+CRF
- 第13章 文本分类与聚类

- 文本表示p418, ppt51

- 向量空间模型: $D(t_1, w_1; t_2, w_2; \dots; t_n, w_n)$

- 文本特征选择方法 t_i

- 基于文档频率

- 信息增益法

- 本质: 不考虑任何特征时文档的熵和考虑特征t后文档的熵的差值

- CHI统计量

- 计算 ppt52 **A**

- 互信息法

- 特征权重计算方法 w_i

- TF-IDF

- 余弦相似度

- 分类器设计

- 朴素贝叶斯分类器NB

- 贝叶斯决策理论

- 计算 ppt53-6 **A**

- 注意平滑公式

- SVM

- 线性判别函数

- 最大间隔准则

- KNN

- 算法描述

- 性能标准

- 正确率、召回率、F测度值

- 微平均和宏平均

- 文本聚类

- 2个假设

- 与文本分类的区别

- 方法

- K-means

- K-medoids

- ...

- 基本概念整理 (记不住的还是记不住🔥)

- NLP的研究困难: 结构歧义、语义歧义、未知语言歧义

- 懒了 (见上)

- 填空简答盲猜（我要我的考点光环😄）

- 均衡分布的熵最大
- 懒了

- 应用分析盲猜（猜对了多少👁👁）

- 噪声信道模型 ppt10
- 语言模型概率计算 ppt17-1
- 前向算法计算（算法步骤ppt22, P113）
- 后向算法计算 ppt23
- viterbi算法 ppt24

- 1.简要叙述文本分类的主要任务和模型,请设计一个中文文本分类的系统实现方案

- 文本分类的主要任务：文本分类是在预定义的分类体系下，根据文本的特征（内容或属性），将给定文本与一个或多个类别相关联的过程。
- 模型：p147（数学模型描述文本分类任务）
- 系统实现方案
 - 1、预处理
 - 将文本按照不同类别归纳到不同目录中
 - 对各个文本进行中文分词
 - 训练集和测试集划分
 - 2、文本表示
 - 根据向量空间模型，对文本进行结构化表示
 - 步骤1 文本特征选择：先基于文档频率对文本特征进行选择
 - 步骤2 特征权重计算：将文本特征的TF-IDF值作为其权重
 - 3、分类器选择
 - 可基于朴素贝叶斯分类器进行文本分类
 - 4、性能评测
 - 可通过正确率、召回率、F-测度值对分类结果进行评估

- 2.简要叙述语义消歧的主要任务,请分别设计一个基于有监督的以及基于词典的实现方案

- 主要任务：确定一个多义词在给定的上下文语境中的具体含义
- 有监督实现方案：基于贝叶斯分类器的消歧方法（p247算法描述）
- 基于词典的实现方案：基于词典语义定义的消歧方法（p249算法描述：假设...）