1、**NLP 定义：**NLP 研究的是如何通过机器学习等技术让计算机学会处理人类语言，乃至实现终极目标--理解人类语言或人工智能。

2、**NLP 层次：**语音、图像和文本 ‖中文分词、词性标注和命名实体识别 ‖信息抽取 ‖文本分类与文本聚类 ‖句法分析 ‖语义分析和篇章分析 ‖其他高级任务。

3、NLP 流派：（1）基于规则的专家系统；（2）基于统计的学习方法；（3）基于深度学习的方法。

1、**词的定义：**在基于词典的中文分词中，词典中的字符串就是词。

2、**词的性质：**齐夫定律——一个单词的词频与它的词频排名成反比。

3、**分词定义：**分词就是将句子、段落、文章这种长文本，分解为以字词为单位的数据结构，方便后续的处理分析工作。（分词算法：基于词典规则与基于机器学习两种）

4、**切分算法：**（1）完全切分；（2）正向最长匹配；（3）反向最长匹配；（4）双向最长匹配　　【词典分词局限性：准确率不高、无法区分歧义、无法召回新词】

5、**字典树：**字典树中每条边都对应一个字，从根节点往下的路径构成一个个字符串并在终点节点上做个标记"该节点对应词语的结尾"。（用于匹配算法：速度快并且省内存）。

6、**字典树优点：**当词典大小为 n 时，虽然最坏情况下字典树的复杂度依然是 O(logn) (假设子节点用对数复杂度的数据结构存储，所有词语都是单字)，但它的实际速度比二分查找快。这是因为随着路径的深入，前缀匹配是递进的过程，算法不必比较字符串的前缀。

7、**基于字典树的其他算法：**（1）首字散列其余二分的字典树；（2）双数组的字典树；（3）AC 自动机；（4）基于双数组字典树的 AC 自动机。

8、**中文分词中的 P,R,F1 值计算：**A={标准答案的所有区间} B={预测结果的所有区间}，准确率,召回率，F1 值。

9、中文分词优缺点：（1）优点：速度快，成本低（2）缺点：适应性不强，不同领域差异大。

1、**语言模型定义：**对语言现象的数学抽象，给定一个句子，语言模型就是计算句子出现概率，而统计的对象就是人工标注的语料库。

2、**语言模型问题：**（1）数据稀疏；（2）计算代价大。

3、**马尔可夫链：**使用马尔可夫假设简化语言模型，假设每个事件的发生概率只取决于前一个事件，那么这串事件构成的因果链被称作马尔可夫链。这种只与前一个事件相关的性质被成为马尔可夫性质。前一个事件与后一个事件的关联关系称为状态转移。

4、**二元语法和 n 元语法：**由于语料库中二元连续的重复程度要高于整个句子的重要程度，所以缓解了数据稀疏的问题，另外二元连续的总数量远远小于句子的数量，存储和查询也得到了解决。【OOV 是 n 元语法的硬伤】

5、**数据稀疏与平滑策略：**n 元语法 n 越大，数据稀疏问题越严重，利用低阶 n 元语法平滑高阶 n 元语法。

6、**二元语法训练与预测：**（1）加载语料库并进行词频统计（2）对词频文件生成词网（3）词网生成词图并使用维特比算法。

7、**维特比算法：**动态规划算法，（1）初始化（2）递推（3）终止（4）回溯

8、基于统计优缺点：（1）优点：适应性较强（2）缺点：成本较高，速度较慢。

1、**序列标注原因：**二元语法词语级别的模型无法应对 OOV，因此需要更细粒度的模型，即字符。

2、**序列标注模型：**序列标注指的是给定一个序列 ，找出序列中每个元素对应标签求解序列标注问题的模型一般称为序列标注器。

3、**HMM：**初始状态概率向量、状态转移矩阵、发射概率矩阵——观测序列与状态序列

4、**HMM 三个基本用法：**（1）样本生成：给定模型λ=(π,A,B)，生成满足模型约束的样本，即系列观测序列及其对应的状态序列{(x^(i),y^(i))}（2）序列预测：给定训练集{(x^(i),y^(i))}，估计模型参数λ=(π,A,B)（3）模型训练：已知模型参数λ=(π,A,B)，给定观测序列 x，求最可能的状态序列 y。

1、**HMM 缺陷：**特征仅限于两种: 其一，前一个标签是什么；其二，当前字符是什么

2、**线性分类模型：**用一条线性的直线（决策边界）或高维平面（超平面）将数据一分为二。

3、**感知机算法：**损失函数、梯度下降、学习率、随机梯度下降。（投票感知机和平均感知机）——读入训练样本执行预测，将预测结果与标准答案比较，更新参数

4、**感知机算法改进:**(1)创造更多特征,将样本映射到更高维空间,使其线性可分;(2)切换算法,例如 SVM;(3)对感知机算法打补丁,例如投票和平均感知机.

5、**结构化感知机：**结构化预测的过程就是给定一个模型 λ 及打分函数 score，利用打分函数给一些备选结构打分，选择分数最高的结构作为预测输出。

6、**结构化学习算法：**（1）读入样本进行结构化预测（2）与正确答案相比若不相等，则更新参数: 奖励正确答案触发的特征函数的权重，否则进行惩罚:（3）还可改学习率。

7、**与感知机算法比较：**（1）结构化感知机修改了特征向量（2）结构化感知机的参数更新赏罚分明。

8、**生成式模型【HMM,朴素贝叶斯模型】与判别式模型【CRF、神经网络、结构化感知机、SVM】 ‖ 有向图与无向图**。

9、无向图模型将概率分解为所有最大团上的某种函数之积(最大团:满足所有节点相互连结的最大子图　因子节点:只连接部分节点,组成更小的最大团)

1、**CRF 模型：**是一种给定输入随机变量 x，求解条件概率 p(y| x) 的概率无向图模型。用于序列标注时，特例化为线性链(linear chain)条件随机场。此时，输入输出随机变量为等长的两个序列。

2、**对比结构化感知机：**（1）**相同点**：特征函数相同、权重向量相同、打分函数相同、预测算法相同、同属结构化学习。（2）**不同点**：感知机算法属于在线学习。感知机更新参数时，只使用一个训练实例，没有考虑整个数据集，难免顾此失彼；而条件随机场对数似然函数及其梯度则使用了整个数据集 ‖ 条件随机场更新参数更加合理,感知机奖励正确答案,但仅惩罚错误最厉害的那个,而 CRF 同时惩罚所有大难,分摊惩罚总量。

3、**CRF 三类问题：**（1）概率计算问题（2）预测问题（3）学习问题。

1、**词性定义：**在语言学上，词性指的是单词的语法分类，也称为词类。

2、**词性用处：**词性的作用是提供词语的抽象表示，词的数量是无穷的，但词性的数量是有限的。词性也可以直接用于抽取一些信息。

3、**词性标注难点:**(1) 一个词语多个词性,但在具体语境下词性唯一　(2)OOV 是任何 NLP 任务的难题

4、**词性标注性能衡量：**P = 预测正确的标签总数/预测的标签总数||R= 预测正确的标签总数/测试集中的标签总数||F1 = 2PR/(P+R)||Accuracy =预测正确的标签总数/标签总数

1、**命名实体：**文本中有一些描述实体的词汇。比如人名、地名、组织机构名、股票基金、医学术语等。**命名实体共性：**（1）数量无穷（2）构词灵活（3）类别模糊。

2、**命名实体识别：**统计为主，规则为辅。命名实体的边界可以通过{B,M,E,S}确定，其类别可以通过 B-nt 等附加类别的标签来确定。

3、**自定义领域命名实体识别:**(1)标记领域命名实体识别语料库【将生语料转换为熟语】料 (2)训练领域模型

4、**命名实体识别性能：**P = 正确识别该类命名实体数/识别出的该类命名实体总数 || R= 正确识别该类命名实体数/该类命名实体总数 || F1 = 2PR/(P+R)

1、**信息抽取定义：**从非结构化文本中提取结构化信息的一类技术。这类技术依然分为基于规则的正则匹配、有监督学习和无监督学习等各种实现方法。

2、**新词提取：**（1）提取出大量文本(生语料)中的词语，无论新旧。用词典过滤掉已有的词语，于是得到新词（2）片段外部左右搭配的丰富程度，可以用信息熵来衡量，而片段内部搭配的固定程度可以用子序列的互信息来衡量。

3、**关键词提取：**（1）词频统计（2）TF-IDF 算法（3）TextRank (PageRank 的变种)

4、**短语提取：**与新词提取相同，字符串替换成单词列表即可。

5、**关键句提取：**BM25+TextRank。

1、**聚类：**指的是将给定对象的集合划分为不同子集的过程，目标是使得每个子集内部的元素尽量相似，不同子集间的元素尽量不相似。这些子集又被称为簇，一般没有交集。

2、**软聚类与硬聚类：**硬聚类每个元素被确定地归入一个簇。软聚类每个元素与每个簇都存在一定的从属程度（隶属度),只不过该程度有大有小。

3、**文本聚类流程：**特征提取（词袋模型）+向量聚类（k 均值算法或者重复二分聚类） ———————— 划分式与层次化

4、**K-mean 算法:**(1)选取 k 个点作为 k 个簇的初始质心(2)将所有点分别分配给最近的质心所在的簇(3)重新计算每个簇的质心(4)重复步骤 2 和步骤 3 直到质心不再发生变化。(更快的准则函数+初始质心的选择)

5、**重复二分算法:**挑选一个簇进行划分.利用 k 均值算法将该簇划分为 2 个子集。重复步骤 1 和步骤 2，直到产生足够舒朗的簇。

6、**词袋模型：**布尔词频: 词频非零的话截取为 1，否则为 0，适合长度较短的数据集 // TF-IDF: 适合主题较少的数据集 // 词向量: 如果词语本身也是某种向量的话，则可以将所有词语的词向量求和作为文档向量。适合处理 OOV 问题严重的数据集。// 词频向量: 适合主题较多的数据集。

1、**文本分类概念：**将一个文档归类到一个或多个类别中的 NLP 任务。文本的类别有时又称作标签。

2、**文本特征提取：**使用词袋向量作为特征向量，步骤为：（1）分词；（2）卡方特征选择（卡方非参数检验过滤与类别相关程度不高的词语）；（3）词袋向量

3、**分类器：**（1）朴素贝叶斯分类器；（2）支持向量机（SVM）：支持向量机的学习策略就是尽量找出离正负样本的间隔最大的分离超平面，以降低测试集上的风险。

4、**朴素贝叶斯训练：**首先计算先验概率分布 P(Y=Ck)，通过统计每个类别下的样本数: 然后计算 P(X=x|Y=Ck)，这个难以估计，为此朴素贝叶斯法“朴素”的假设了所有特征是条件独立的: 于是，又可以利用极大似然来进行估计:预测时，朴素贝叶斯法依然利用贝叶斯公式找出后验概率 P(Y=Ck|X=x) 最大的类别 Ck 作为输出 y: 将贝叶斯公式带入上式得: 最终，由于分母与 Ck 无关，可以省略掉，然后将独立性假设带入，得到最终的分类预测函数:

5、**文本分类评测**：对每一个类别的分类结果，正确分入该类的样本数量记作 TP，错误分入该类的样本数量记作 FP，本该分入该类却错误地分入其他类的样本数量记为 FN

6、**情感分类挑战：**（1）背景与极性（2）确定主观性与语气（3）识别挖苦与讽刺（4）中性消息，无法归类于任何类别。

1、**语法分析：**分析句子的语法结构并将其表示为容易理解的结构（通常是树形结构）。

2、**短语结构树（上下文无关文法）：**（1）终结符集合（2）非终结符集合（3）推导规则。

3、**依存句法树：**依存语法理论认为词与词之间存在主从关系，是一种二元不等价关系，其中修饰词成为从属词，被修饰词称为支配词，两者间的语法关系称为依存关系。

4、**现代依存语法：**（1）有且只有一个词语不依存于其他词语【根节点唯一性】（2）除此之外所有单词必须依存于其他词【连通性】（3）每个单词不能依存于多个单词【无环性】（4）如果单词 A 依存于 B，那么位置处于 A 和 B 之间的单词 C 只能依存于 A、B 或 AB 之间的单词。【投射性】　　【支配词指向从属词】

5、**依存句法分析：**（1）基于图的依存句法分析（2）基于转移的依存句法分析【Arc-Eager 转移系统】

6、**依存句法评测：（1）无标记依存正确率 UAS：**对应忽略标签的 F1 值，测试集中找到其正确支配词的词（包括没有标注支配词的根结点）所占总词数的百分比。**（2）带标记依存正确率 LAS：**对应包括标签的 F1 值，测试集中找到其正确支配词的词，并且依存关系类型也标注正确的词（包括没有标注支配词的根结点）占总词数的百分比。**（3）依存正确率 DA：**测试集中找到正确支配词非根结点词占所有非根结点词总数的百分比。**（4）根正确率 RA：**有二种定义，一种是测试集中正确根结点的个数与句子个数的百分比。另一种是指测试集中找到正确根结点的句子数所占句子总数的百分比。**（5）完全匹配率 CM：**测试集中无标记依存结构完全正确的句子占句子总数的百分比。【给定两棵树，一棵树为标准答案（来自测试集），一棵树为预测结果，评测的目标是衡量这两棵树的差异。如果将树的节点编号，拆解为依存弧并分别存入两个集合 A(标准答案)和 A*(预测结果)，则可以利用分类任务的评价指标。】

1、**传统方法局限：**（1）数据稀疏【可用稠密向量解决】（2）特征模板【多层网络自动提取特征】（3）误差传播【端到端设计】

2、**深度学习优缺点：**（1）优点：准确率高、适应性强（2）缺点：成本高，速度慢。

3、**Word2Vec 和 CBOW 模型：**CBOW 是一种基于窗口的语言模型，利用上下文来预测中心词。一个窗口指的是句子中的一个固定长度的片段，窗口中间的词语称为中心词，窗口中其他词语称为中心词的上下文。CBOW 模型通过三层神经网络接受上下文的特征向量，预测中心词是什么。

4、**NLP 进阶：**（1）在词嵌入的预训练方面,word2vec——fasttext——ELMO（2）Transformer 和注意力机制提取特征。