NLP学习过程

1. Bag of Words：生成词特征向量
2. TF-IDF
   1. TF：term frequency（词频）
   2. IDF：总文件数目/包含某词语的文件数目（用于为特征加权值，可以体现出特征在某一类别上的区分性）
3. Chi2算法：用于从众多特征中选出较好的Top-K个特征
4. POS tagging：词性标注
5. NER：命名实体识别
6. HMM：隐马尔科夫模型
7. CRF：条件随机场
   1. 随机场：一组随机变量，样本空间一样
   2. 马尔可夫性
   3. 满足马尔可夫独立性的随机场叫做马尔可夫随机场

MIT自然语言处理学习

1. NLP的应用 : 机器翻译，提取文本摘要（主题建模），对话系统【机器人客服】，语法检查，情感分类，作文评分，【语言】-【知识】，【语言】-知识-【语言】
2. 自然语言处理的难点：语义【Semantic】上产生的歧义，在语音识别中声音上产生的歧义，句法【Syntactic】层次的歧义，上下文关系【Context】中产生的歧义
3. 关于自然语言处理可能的解决方法：符号方法【将所有NLP需要的信息在计算机里进行编码】【我们需要的知识和信息很多，这些信息很难编码】，统计方法【收集大量文本-对于其中每个名词计算它和特定的限定词一起出现的概率-对于一个新名词，根据训练语料库中最高似然估计选择一个限定词，结果一般，但方法简单】，预测方法【学习分类函数，用于解决特定问题，如：一个词是否是复数，某个词是否第一次出现在文本中，是否是一个名词等等】，分类方法【学习一个X->Y的映射函数】，分类之外【很多时候分类框架并不适用于如语法分析，机器翻译以及自然语言生成等场景】
4. 句法【Syntactic】，语义【Semantic】，篇章【Discourse】
5. 平滑估计【Smoothed estimation】，对数线性模型【log-linear models】，概率上下文无关法【probabilistic context free grammar】，EM算法，联合训练【co-training】
6. 机器翻译【machine translation】，文本摘要【text summarization】，信息检索【information retrieval】
7. 语料库【Corpora，Corpus】【平行语料库，多语言组成，用于机器翻译等】【平衡语料库，主要指其语料的取样上是均衡的，有代表性的，这种语料可以用作得出有关某种语言特性的一般性的结论】
8. 语料库非常大会导致数据稀疏问题【Data Parsity】
9. 齐夫定律【Zipf’s Law】【一个单词出现的频率与它在频率表里的排名成反比，如Brown语料库中，频率排名第一的单词’The’出现的次数近似于排第二的单词’of’出现次数的2倍】
10. 分词【Tokenization】【将文本单词切分成单词序列】
11. 概率语言模型【向字符串赋予概率的方法称之为概率语言模型】【哪一个字符串更有可能或者更符合语法】【使用贝叶斯规则，通过最大似然估计法进行计算】
12. 如何构造一个概率语言模型【始于一个词汇集合V】-【根据V中的词汇得到一个训练样本】-【假设训练样本是由一些隐藏的分布P刻画的】-【学习一个概率分布P-p尽可能与P近似】-【根据训练样本计算出语言模型<分布P-p>的参数，如马尔科夫模型】-【对模型进行评估】
13. 如何计算一个概率语言模型【平滑算法-最大似然估计】【Good-Turing打折法】
14. 思考【在自然语言处理中，都有哪些常见的模型？这些模型分别适合用来解决什么养的应用问题？】
15. 自然语言处理之标注【在句子中为每个词标上合适的词性】【part of speech，POS】
16. 词性标注对许多领域都是非常重要的【语法分析，语言模型，问答系统和信息抽取，文本语言转换】
17. 各种粒度的多种标记集【Penn tag set 45tags】【Brown tag set 87tags】【CLAWS2 tag set 132tags】【CC-连接词，DT-限定词，JJ-形容词，NN-名词，RB-副词，VBD-动词】
18. 个人理解【词性标注是一种自然语言处理方法，为了完成词性标注的任务需要使用机器学习的方法，如HMM，CRF等等，同时完成词性标注的句子又可以作为训练集<标注的词性可作为特征或者生成特征的元素>来帮助完成如句法分析，机器翻译或者语音识别等应用】
19. POS的方法【基于转换的学习，TBL】【隐马尔科夫模型，HMM】【对数线性模型】
20. 词性标注的总结

a) 标注是一个相对比较简单的任务，至少在一个有监督框架下对于英语来说（Tagging is relatively easy task (at least, in a supervised framework, and for English)）  
b) 影响标注器性能的因素包括（Factors that impact tagger performance include）:  
　i. 训练集数量（The amount of training data available）训练集的平衡性  
　ii. 标记集（The tag set）  
　iii. 训练集和测试集的词汇差异（The difference in vocabulary between the training and the testing）  
　iv. 未登录词（Unknown words）  
c) TBL和HMM框架可用于其他自然语言处理任务（TBL and HMM framework can be used for other tasks）

1. 最大熵和对数线性模型【什么是最大熵模型？】【最大熵模型的原理及应用】