练习题

编辑距离的计算

编辑距离（Edit Distance）可以用来度量两个字符串之间的距离，其定义是把一个字符串变成另一个字符串所需要的插入和删除操作的数目。若有两个字符串x和y，其中x = abbcde，y = bcduvhe，则x和y之间的编辑距离是（　　）。

A: 3　B: 4　C: 5　D: 6

词法分析例题

基于统计的分词模型不包括下列哪个选项（　　）

A：N元文法模型　　B：隐马尔可夫模型

C：最大匹配模型　　D：最大熵模型

深度学习模型例题

CNN模型可以用来进行关系分类。一个CNN模型可以包含多个卷积层和池化层，假设一个CNN模型的某卷积层卷积核的大小为3*100，该层共有10个卷积核，则在不考虑偏置的情况下，该卷积层的参数个数为（　　）。

A: 300　B: 10　C: 100　D: 3000

Perplexity(困惑度)的计算

- Let's suppose a sentence consisting of random digits.
- What is the perplexity of this sentence according to a model that assign P=1/10 to each digit?

$$
\begin{aligned}
\mathrm{PP}(W) &= P(w_1 w_2 \ldots w_N)^{-\frac{1}{N}} \\
&= \left(\frac{1}{10}^{N}\right)^{-\frac{1}{N}} \\
&= \frac{1}{10}^{-1} \\
&= 10
\end{aligned}
$$

文本表示及信息抽取

- 假设词典的大小为|V|，在使用one-hot表示词典中的词语时，每个词语对应的向量维度为_____；若使用最大熵模型进行命名实体识别（标记的个数为t），采用当前词、前一个词和后一个词作为特征，则模型理论上的特征数为_____。

简答与计算举例

- 各种任务的评价方法
  - 如：命名实体识别的评价方法
- 各种信息抽取任务的定义
- Skip-gram的实现过程
- HMM定义和参数估计
- …