自然语言处理2025回忆版

全是计算，没有概念题，~~不喜欢背书的宝宝有福辣！~~

一（10）：给一句话，让你用正向和逆向的最大匹配法分别切分句子。
二（10）：给了一个矩阵。
1.让你用二元语法计算一个句子（i Want to eat chinese food）的概率（牢春还贴心地把<BOS>和<EOS>写出来了）
2.又给了几个概率，然后问你是更喜欢吃中国食物还是外国食物。

三（10）：考的文本分类（贝叶斯分类）

跟下面这个基本是一模一样

$$\hat{P}(c)=\frac{N_c}{N}$$

$$\hat{P}(w\mid c)=\frac{count(w,c)+1}{count(c)+|V|}$$

	Doc	Words	Class
Training	1	Chinese Beijing Chinese	u
	2	Chinese Chinese Shanghai	u
	3	Chinese Macao	u
	4	Tokyo Japan Chinese	v
Test	5	Chinese Chinese Chinese Tokyo Japan	?

先验概率：

$P(u)=\frac{3}{4}$

$P(v)=\frac{1}{4}$

条件概率：

P(Chinese|u) = (5+1) / (8+6) = 6/14 = 3/7
P(Tokyo|u) = (0+1) / (8+6) = 1/14
P(Japan|u) = (0+1) / (8+6) = 1/14
P(Chinese|v) = (1+1) / (3+6) = 2/9
P(Tokyo|v) = (1+1) / (3+6) = 2/9
P(Japan|v) = (1+1) / (3+6) = 2/9

各个类别：

P(u|d5) ∝ 3/4 * (3/7)³ * 1/14 * 1/14
≈ 0.0003

P(v|d5) ∝ 1/4 * (2/9)³ * 2/9 * 2/9
≈ 0.0001

四（10）：给了四个文本
1.让你根据这四个文本建立一个二元语法模型
2.根据建立的模型计算句子概率

大概跟下面这个一样：

- 举例： 文本1：中国万岁
  文本2：中国中国
  文本3：万岁中国
  文本4：万岁万岁

一元语法：

$p(\text{中国}) = 4/8$， $p(\text{万岁}) = 4/8$

二元语法：

$p(\text{中国} \mid \text{<BOS>}) = 2/4$， $p(\text{万岁} \mid \text{<BOS>}) = 2/4$， $p(\text{<EOS>} \mid \text{万岁}) = 2/4$
$p(\text{中国} \mid \text{中国}) = 1/4$， $p(\text{万岁} \mid \text{中国}) = 1/4$， $p(\text{<EOS>} \mid \text{中国}) = 2/4$
$p(\text{中国} \mid \text{万岁}) = 1/4$， $p(\text{万岁} \mid \text{万岁}) = 1/4$， $p(\text{<EOS>} \mid \text{万岁}) = 2/4$

有

$p(\text{中国万岁}) = p(\text{中国} \mid \text{<BOS>}) * p(\text{万岁} \mid \text{中国}) * p(\text{<EOS>} \mid \text{万岁}) = 2/4 * 1/4 * 2/4$

五（15）：牢春自定义了一手tf-idf向量，让你根据他给的定义计算几个词的tf-idf值，以及根据余弦相似度找近义词。
1.计算一个词的TF和IDF向量
2.计算两个词的if-idf向量
3.根据余弦相似度找近义词
六（30）：文本分类（logistic回归分类）+前向神经网络
1.给了你x,w,b，让你计算P（+|d）
2.假设文本d的真实类别为“-”，求交叉熵损失
3.给你tp、tn、fp、fn让你计算精确率召回率和F1测度
4.题目提前说明了隐藏层的节点数，让你画一个二分类的神经网络，并指明网络的连接情况、各层的激活函数，输入节点和输出节点的数量
5.跟四一样，只不过二分类变成三分类
6.计算多分类的精确率召回率和准确率
七（15）：考的HMM

1.给定隐状态，和观测值，让你求此隐状态下的观测值的概率

2&&3.第二三问分别考的正向传播和维特比算法

整理人：高宏达、肖俨哲、某不愿透露姓名的热心人士