北京航空航天大学　2022 年秋季学期

《自然语言处理》期末考试试卷

注意事项：

1. 本课程为闭卷考试，答题时间为 120 分钟。

2. 请直接在答题纸上填写答案，不必抄题，字迹清晰。

一、单项选择题（共 10 题，每题 2 分，共 20 分）

1. 自然语言处理技术起源于_____

   A. 机器翻译　B. 语音识别　C. 信息检索　D. 自动摘要

2. 下列语言中_____为自然语言？

   A. 汇编语言　B. C 语言　C. 鸟语　D. 甲骨文

3. 判断两个连续汉字之间的结合强度，下列选项中_____更合适

   A. 条件熵　B. 互信息　C. 双字耦合度　D. 交叉熵

4. 计算英文(26 个字母和 1 个空格，共 27 个字符)信息源的熵：_____

   A. 4.55(bits/letter)　B. 4.75(bits/letter)　C. 4.82(bits/letter)　D. 4.69(bits/letter)

5. 假设 Σ{0,1}，0 和 1 都是正则表达式。如果令$x=0, y=1$，那么 $x|y^*$ 对应的正则集为_____

   A. $\{0, \varepsilon, 01, 011, 0111, \cdots\}$　B. $\{0, 1, 11, 111, \cdots\}$

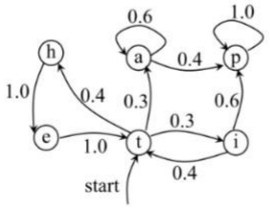
   C. $\{0, 01, 011, 0111, \cdots\}$　D. $\{0, \varepsilon, 1, 11, 111, \cdots\}$

6. 在乔姆斯基的语法理论中，文法被划分为 4 种类型，下列不属于乔姆斯基语法理论的是_____

   A. 正则文法　B. 上下文无关文法　C. 有约束文法　D. 上下文有关文法

7. 假设有一个马尔可夫链，其状态图如下图所示，计算 $p(t, a, a, p)$ 的概率值为_____

   A. 0.108　B. 0.0288　C. 0.0108　D. 0.072



8. 假设某个汉语分词系统在一测试集上输出 5260 个分词结果，而标准答案是 4510 个词语，根据这个答案，系统切分出来的结果中有 4110 个是正确的。那么分词系统的召回率为_____

   A.78.14%　B. 91.13%　C. 21.86%　D. 84.14%

9. 给定一个语料，“为人”出现 5 次，“为人民”出现 20 次，那么（为，人）的双字耦合度等于_____

   A. 0.4　B. 0.3　C. 0.25　D. 0.2

10. 字符串 eistaner 与单词 distance 之间的编辑距离是_____

    A. 1　B. 2　C. 3　D. 4

二、简答题（共 4 题，共 30 分）

1. 简述基于最大熵的消歧方法的基本思想。（本题 6 分）
2. 简述句法分析的任务，并列出句法分析的类型。（本题 6 分）
3. 设文法 G 由如下规则定义：（本题 8 分）

   $S \to AB \quad A \to Aa|bB \quad B \to a|Sb$

   给出下列句子形式的推导过程及产生的派生树：

   (1) $baabaab$　(2) $bBABb$
4. 给定正则文法 $G=(V_N,V_T,P,S)$，其中 $V_N=\{S,B\}$，$V_T=\{a,b\}$，$P=\{S\to aB,\ B\to bS|aB|a\}$ 构造与 G 等价的有限自动机，并画出有限自动机的状态图。（本题 10 分）

三、分析计算题（共 3 题，共 50 分）

1. 给定训练语料：（本题 15 分）

   “<BOS> John read Moby Dick <EOS>”

   “<BOS> Mary read a different book <EOS>”

   “<BOS> She read a book by Cher <EOS>”

   “<BOS> He buy a book from the store <EOS>”

   其中<BOS>是前缀，表示句子开头，<EOS>是句缀，表示句子结束。

   回答如下问题：

   (1) 请写出语料中所有的二元文法 2-gram 并统计其在训练语料中出现的次数。（如(<BOS>, John)，出现的次数是 1 次）。

   (2) 根据 2-gram 方法计算概率 $p(book|a)$ 和概率 $p(a|read)$。

   (3) 根据 2-gram 方法和加 1 数据平滑方法，计算文本“John buy a book”。

   (4) 对于 k-gram 方法，当 k 取值较大时，存在哪些问题？这些问题应该如何解决？
2. 给定隐马尔可夫模型 HMM $\mu=(A,B,\pi)$。（本题 15 分）

   $A=a_{ij}$ 是状态转移概率矩阵，$B=b_j(k)$ 是观察矩阵，π 是初始状态矩阵。

   $$\begin{cases} a_{ij}=p(q_{t+1}=S_j|q_t=S_i) \\ a_{ij}\geq 0 \\ \sum_{j=1}^{N} a_{ij}=1 \end{cases} \qquad \begin{cases} b_j(k)=p(O_t|q_t=S_j) \\ b_j(k)\geq 0 \\ \sum_{k=1}^{M} b_j(k)=1 \end{cases}$$

   其中，S_i 是状态i，O_t 是第 t 个观察到的数值。

   (1) 简述 HMM 模型的两个假设。

   (2) HMM 要解决的三个问题是什么？

   (3) 现给定一个 HMM 模型 μ，简述产生观察序列 $O=O_1O_2\cdots O_T$ 的生成过程。
3. 假设有如下文法 $G(S)$：（本题 20 分）

   ① $S\to P\ \ VP$　② $VP\to PP\ \ V$　③ $VP\to Prop\ \ VP$　④ $VP\to VP\ \ V$

   ⑤ $VP\to V\ \ Aux$　⑥ $VP\to VP\ \ N$　⑦ $PP\to Prop$

   有一部词典包含如下词条和每个词条对应的词性（如有词性兼类，两词之间用逗号隔开）：

# 从：Prop	# 了解：V	# 题：N	# 学会：V, N
# 从小：Prop	# 解：V, N	# 小：A	……
# 会：V, N	# 解题：V	# 小学：N	
# 了：Aux	# 他：P	# 学：V	

给定句子：他从小学会了解题。（假设不考虑句号）请完成下面的题目：

(1) 请给出正向最大匹配算法的分词过程（简述原理），并分别给出正向最大匹配算法和逆向最大匹配算法的分词结果。

(2) 请用逆向最大分词算法给出的分词结果和 CYK 分析算法分析给定句子，写出句法分析过程，并给出一种句法分析树。

(3) 如果有其他可能的句法分析树结构，请直接给出结果（不必写具体的生成过程）。