学院____专业____班级____本专 学号____姓名____座号

密封线　学生须将答案写在此线以下

鲁东大学 2022 — 2023 学年第 1 学期

2020 级　人工智能　专业 本科卷 A

课程名称 自然语言处理

课程号（2220188202）　考试形式（闭卷）　时间（120 分钟）

题目	一	二	三	总分	统分人	复核人
得分						

得分	评卷人

一、完成下列各题（共 5 小题，每小题 4 分，满分 20 分）

1、列举自然语言处理应用（至少四项）。

2、语料库的定义。

3、文本分类的基本流程。

4、共现矩阵生成词向量的过程。

5、LSTM 对 RNN 的主要改进及其改进作用。

得分	评卷人

二、简述题（共 5 小题，每小题 10 分，满分 50 分）

1、简述主题模型：

(1) LDA 中“文档-词项”的生成模型；

(2) PLSA 和 LDA 两者的区别。

2、简述注意力机制在机器翻译领域中的作用。

3、相比于 CNN，RNN 为什么在自然语言处理领域占据主流位置。

4、简述词嵌入方法 Word2vec 中 CBOW 的基本原理。

5、什么是平滑技术？列举出两种平滑技术。

得分	评卷人

三、计算题（共 2 小题，每小题 15 分，满分 30 分）

1、给定句子：海南一直是我向往的地方，椰子树、大海在我的心里是那样的神秘。(假设不考虑句号)。

(1) 简述基于统计分词算法的实现过程；

(2) 简述最大匹配分词算法的实现过程；

(3) 请给出利用正向最大匹配分词算法对给定句子的分词结果，并简单给出分析过程。

分词字典：

海 | 南 | 海南 | 一直 | 是 | 我 | 向往的地方 | 向往 | 的 | 地方 | 椰子 | 树 | 椰子树 | 大海 | 大 | 在 | 我的 | 在我的 | 心 | 心里 | 里 | 那样 | 神秘

2、下表是由 15 个样本组成的贷款申请训练数据，数据包括贷款申请的 4 个特征（年龄、是否有工作、是否有自己的房子、信贷情况），最后一列表示是否同意贷款，利用该训练数据，通过信息增益准则选出最优的分类特征变量。

(1) 简述以 ID3 算法为例决策树生成流程；

(2) 构建决策树。

参考公式：　信息熵：$H(X)=-\sum_i P(x_i)\log(P(x_i))$

信息增益：$IG(T,a)=H(T)-H(T\mid a)$

$\log\frac{1}{15}=-3.91$　$\log\frac{2}{15}=-2.91$　$\log\frac{3}{15}=-2.32$　$\log\frac{4}{15}=-1.91$　$\log\frac{5}{15}=-1.58$　$\log\frac{6}{15}=-1.32$

$\log\frac{7}{15}=-1.10$　$\log\frac{8}{15}=-0.91$　$\log\frac{9}{15}=-0.74$　$\log\frac{10}{15}=-0.58$　$\log\frac{11}{15}=-0.45$　$\log\frac{12}{15}=-0.32$

$\log\frac{1}{4}=-2$　$\log\frac{3}{4}=-0.42$

ID	年龄	工作	自己的房子	信贷情况	类别
1	青年	否	否	一般	否
2	青年	否	否	好	否
3	青年	是	否	好	是
4	青年	是	是	一般	是
5	青年	否	否	一般	否
6	中年	否	否	一般	否
7	中年	否	否	好	否
8	中年	是	是	好	是
9	中年	否	是	非常好	是
10	中年	否	是	非常好	是
11	老年	否	是	非常好	是
12	老年	否	是	好	是
13	老年	是	否	好	是
14	老年	是	否	非常好	是
15	老年	否	否	一般	否