**2023 年东华大学 试卷**

**38 自然语言处理 （ A 场）**

（本试卷考试时间 90 分钟）

**一、单选题 ( 本大题 10 道小题 ，每小题 2 分，共 20 分），从下面题目给出的A、B、C、D四个可供选择的答案中选择一个正确答案。**

1.以下是对文本的分析结果，实现了\_\_\_\_\_\_\_。

欢迎 来到 东华大学 2023 夏令营

v v ni nt n

**答案：A**

A.词性标注

B.命名实体识别

C.句法分析

D.篇章分析

2.词袋模型的主要缺点是\_\_\_\_\_\_\_。 **答案：D**

A.将句子表示为等长的数字向量，无法反映句子长度

B.只使用主题词表示文本，漏掉其他词语信息

C.只使用文本中关键句子表示整段文本信息

D.只考虑的文本中的词信息，忽略了词序、语法以及句法信息

**二、多选题 ( 本大题 5 道小题 ，每小题 2 分，共 10 分），从下面题目给出的A、B、C、D四个可供选择的答案中选择所有正确答案。**

1.以下关于中文分词描述，\_\_\_\_\_\_\_\_是错误的。 **答案：BCD**

A.将中文句子划分为以词为单位的词序列

B.去掉文本中于主题无关的句子

C.将文本转换为数字向量

D.从文本中抽取与主题有关的词语

**三、操作题**

三、 简答题（2道题，共 30分）

**注：答案需填写在C:\KS\answer.doc文档相应的题目后（按照文档中每道小题要求填写相应答案）。**

1. 知识图谱三元组抽取任务，通常也称为关系抽取任务（Relational Extraction），请论述该任务的主要功能，结合实例说明训练数据集的标注方法。说明常用的管道式（pipeline）抽取模型和联合（Joint）抽取模型的构建方法和特点。

1） RE任务描述。（5分）

2） 定义BIO模式标签，并进行样本标注。（5分）

目标：从文本“罗贯中，元末明初小说家，三国演义的作者”中抽取三元组：{ 作者，编著，图书}。

定义标签：B-sub，I-sub，O，B-obj，I-obj。

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 罗 | 贯 | 中 | ， | 元 | 末 | 明 | 初 | 小 | 说 | 家 | ， | 三 | 国 | 演 | 义 | 的 | 作 | 者 | 。 |

3） 管道式抽取模型和联合抽取模型的构建方法和特点。（5分）

四、 程序分析题（2道题，共40分）

**注：按照程序的给出的提示，补全程序中所缺的关键代码，将代码/运行结果截图/结果分析等填写在C:\KS\answer.doc文档相应的题目后（按照文档中每道题目要求填写相应答案）。**

**！！！编写完成的源程序.py文件、答案文件answer.doc需保留在C:\KS文件夹下，否则无法提交。**

1. 文本关键词提取，短文本（5条）相似度比较。源程序文件见C:\KS\prog-1.py（18分）

1）补全代码内容（12分），数据读入、TF-IDF特征表示、关键词提取、余弦相似度计算；

2）关键词输出截图（3分）；

3）文本相似度输出截图，并列出最相似的两条文本（3分）。