中国科学院大学

试题专用纸（A卷）

考试日期：2019年7月9日

课程编号：091M4044H

课程名称：自然语言处理

任课教师：宗成庆、赵军、张家俊

姓名[redacted] 学号[redacted] 成绩_____

说明：本次考试为闭卷考试，卷面成绩满分为60分。本课程的最终成绩由两部分组成：闭卷考试成绩+项目作业成绩（满分为40分）。

一、完成下列各题（满分30分）：

(1) 请给出短语“中国科学院大学”与“国科大”之间的编辑距离。

(2) 有如下复杂特征集：

$$\begin{pmatrix} \text{Cat} = \text{S} \\ \text{Subject} = \begin{pmatrix} \text{Cat} = \text{Pron} \\ \text{Number} = \text{Plur} \\ \text{Person} = \text{First} \\ \text{Lex} = \text{他} \end{pmatrix} \\ \text{Objective} = \begin{pmatrix} \text{Cat} = \text{Pron} \\ \text{Number} = \text{Sing} \\ \text{Person} = \text{Third} \\ \text{Lex} = \text{她} \end{pmatrix} \\ \text{Predicator} = \begin{pmatrix} \text{Cat} = \text{Verb} \\ \text{Lex} = \text{帮助} \end{pmatrix} \\ \text{Tense} = \text{Past} \\ \text{Voice} = \text{Active} \end{pmatrix}$$

他

请写出该复杂特征集所描述的句子。

(3) 举例说明词性标注（消歧）中的“并列鉴别规则”。

(4) 假设 $X \sim p(x)$，$q(x)$ 为用于近似 $p(x)$ 的一个概率分布，则 $p(x)$ 与 $q(x)$ 的交叉熵定义为 $H(p,q) = H(p) + D(p \| q)$。请证明：$H(p,q) = -\sum_x p(x)\log q(x)$。

(5) 请标出下面这段文字中全部的命名实体，并说明命名实体的类型：

钓鱼岛，亦称钓鱼台、钓鱼屿、钓鱼山，是中国东海钓鱼岛及其附属岛屿的主岛，是中国自古以来的固有领土。位于北纬25°44.6′，东经123°28.4′，距浙江温州市约358千米、福建福州市约385千米、台湾基隆市约190千米，周围海域面积约为17.4万平方公里。

(6) 请给出句子“我们选他当书记”的依存关系图，并说明该句子是否满足依存句法理论的基本约束。

二、简述题（满分12分）

1. 请简述基于中间语言的机器翻译方法（Interlingua-Based Machine Translation）的基本原理及其优点和弱点，并写出机器翻译译文质量自动评价指标BLEU的计算公式。

2. 知识图谱是人工智能基础研究和互联网应用融合的产物，请简要阐述语义网络(Semantic Network)、语义网(Semantic Web)和知识图谱的区别与联系。

三、分析计算题（满分 8 分）：

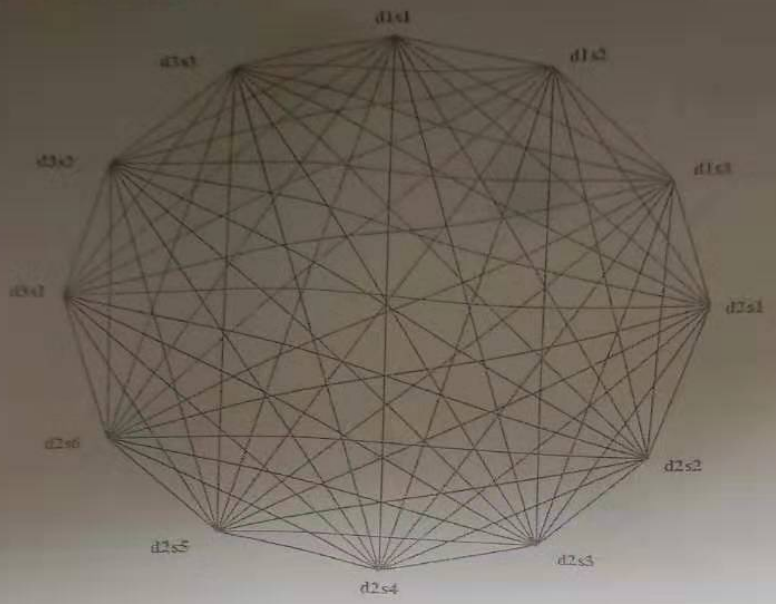

图 1：面向多文档摘要的图结构表示

表 1：句子间相似度得分统计，例如第一行的第五个元素 0.18 表示第一个句子 d1s1 与第五个句子 d2s2 之间的相似度得分。

1	0.13	0.14	0.12	0.18	0.10	0.14	0.17	0.07	0.47	0.27	0.15
0.13	1	0.24	0.04	0.11	0.04	0.03	0.09	0.09	0.14	0.11	0.12
0.14	0.24	1	0.04	0.10	0.09	0.03	0.17	0.05	0.14	0.14	0.20
0.12	0.04	0.04	1	0.15	0.26	0.37	0.20	0.07	0.07	0.14	0.10
0.18	0.11	0.10	0.15	1	0.16	0.20	0.11	0.09	0.09	0.03	0.18
0.10	0.04	0.09	0.26	0.16	1	0.50	0.11	0.08	0.12	0.21	0.05
0.14	0.03	0.03	0.37	0.20	0.50	1	0.14	0.08	0.14	0.29	0.02
0.17	0.09	0.17	0.20	0.11	0.11	0.14	1	0.15	0.21	0.09	0.07
0.07	0.09	0.05	0.07	0.09	0.08	0.08	0.15	1	0.04	0.02	0.04
0.47	0.14	0.14	0.07	0.09	0.12	0.14	0.21	0.04	1	0.37	0.13
0.26	0.11	0.14	0.14	0.03	0.21	0.28	0.09	0.02	0.37	1	0.16
0.15	0.12	0.20	0.10	0.18	0.05	0.02	0.07	0.04	0.13	0.16	1

对于抽取式多文档自动摘要任务，其核心模块是计算每个句子的重要性得分。假设输入三篇文档，第一篇文档包含 3 个句子，第二篇文档包含 6 个句子，第三篇文档含有 3 个句子。如果每个句子作为一个节点（例如 d1s1 表示第一篇文档的第一个句子），句子之间的相似度作为边的权重（表 1 给出了任意两个句子之间的相似度得分），从而构成一个图（如上面的图 1 所示）。假设每个句子的重要性得分都初始化为 0.1，即 S(v)=0.1，其中 v 表示图中任意的节点（句子）。那么，请根据上述信息写出基于图的自动摘要方法 TextRank 中句子重要性得分的迭代计算公式，并计算 d1s1 和 d3s1 经过第一次迭代后的重要性得分。

四、计算题（满分 10 分）：

实体消歧的核心问题是计算待消歧实体之间的相似度，该相似度主要由待消歧实体上下文的语义关联决定，其中，上下文为待消歧实体所在句子中的主要词语。例如，给定如下句子及词语语义图：

S1：苹果是一家高科技的公司。

S2：Iphone 是苹果公司的主要产品。

S3：苹果营养丰富，味道甜美。

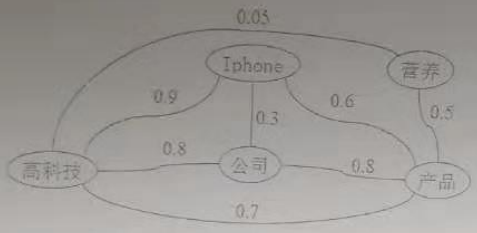

图 2：S1-S3 词语的语义图

在句子 S1、S2、S3 中，“苹果”是待消歧实体，下划线词语是句子中的主要词语（例如：S1 中的“高科技”）。如果要计算 S1 与 S2 中的“苹果”之间的相似度，需要计算 S1 与 S2 中的主要词语“高科技”和“Iphone”之间的语义相似度。

句子间主要词语w_i和w_j的语义相似度（Sim_{ij}）计算公式如下：

$$Sim_{ij} = 0.5 \times (Sr(i \to j) + Sr(j \to i))$$

$$Sr(i \to j) = \alpha A_{ij} + \beta \sum_{l \in N_{ij}} \frac{A_{il}}{d_i}$$

其中，$Sr(i \to j)$是语义图上节点i到节点j的语义关联，N_{ij}是与节点i、节点j都直接相连的词语节点集合（如图 2 所示，与两节点“公司”、“高科技”都直接相连的词语节点集合为{“Iphone”，“产品”}），A_{il}是语义图中两节点i和l之间边的权重（如节点“营养”与“产品”之间边的权重为 0.5），d_i是节点i的度（在无向图中，节点的度指图中与其相连的节点个数，如图 2 中节点“高科技”的度为 4），α和β是加权项，$\alpha = 0.6$，$\beta = 1$。

(1) 请计算 S1 与 S2 中“苹果”的相似度Sim_{12}（即计算“高科技”与“Iphone”的语义相似度）（最终计算结果保留两位小数）；

(2) 请计算 S1 与 S3 中“苹果”的相似度Sim_{13}（最终计算结果保留两位小数）。