University of Glasgow | School of Computing Science

评估课程作业

课程名称	**文本即数据**			
课程数量	**1**			
截止日期	**时间**	**下午 4:30**	**日期**	**2024 年 3 月 12 日**
对最终航标	**20**			
独奏或小组演奏	**✓ 独奏**	**✓**	**组别**	
预计小时数	**30 小时**			
提交说明	**根据以下规格。**			
请注意：本课程作业不能重新评估				

提交课程作业的评估规则

课程文件中将公布提交正式评估课程作业的截止日期，逾期提交作业将受到如下处罚。

对于在公布的截止日期后提交的课程作业，将按以下方式计算初级成绩和二级分数段：

(i) 对于在截止日期后五个工作日内提交的作品

a. 作品将按常规方式进行评估；

b. 然后，将按迟交作业的每个工作日（或工作日的一部分）减少两个二级等 级。

(ii) 超过截止日期五个工作日后提交的作品将被评为 H 级。

如果有正当理由，则不会对迟交作业进行处罚。您应通过 MyCampus 提交证明理由充分的文件。

不遵守提交说明的处罚为 2 个等级

您必须通过 https://studentltc.dcs.gla.ac.uk/ 填写所有课程作业的 "自己的作

品"表格

文本即数据课件

导言

TaD 课程作业旨在评估您将文本处理技术应用于多选题答题系统的能力。

您的作业将通过 Moodle 提交，主要根据 **PDF 报告进行**评估。您的代码将作为一个或多个支持性 **Jupyter/Colab 笔记本**（作为单独的 .ipynb 文件）提交。这是一项**个人练习**，您应独立完成。如果您对本文档有任何疑问，请尽快联系课程讲师。

您的任务是建立并评估一个问题解答系统，该系统可从 4 个选项中选出最佳答案。例如

例如：1 汤匙水是多少 [正确答案：(c)

(a) 在英国、欧洲和大多数英联邦国家，汤匙是一种大勺，通常用于盛放食物。

(b) 在美国和加拿大部分地区，汤匙是用碗吃饭时使用的最大勺子。

(c) 这个汤匙的容量约为 15 毫升。

(d) 量匙

在本练习中，你可以假定所有选项都与事实相符；任务是确定哪项陈述提供了问题的答案，而不一定是哪项提供了与事实相符的信息。

数据集可通过以下链接下载： http://tinyurl.com/tad2024courseworkdata

Q1 - 数据集和预处理 [8 分]

从上面的链接开始下载 WikiQA 语料库的训练、验证和测试版本。将数据加载到笔记本中，并回答以下相关问题。请注意，数据附带的 README.txt 文件提供了有关数据格式、字段及其他构建信息的说明，这将有助于正确理解数据。

使用实验室 3 中的 `text_pipeline_spacy_special` 函数标记所有问题及其选项。然后回答下列问题：

(1.1) 每个单元有多少个问题和选项？**[1 分]**

(1.2) 训练集中每个问题的平均标记数是多少？**[1 分] (1.3)** 训练集中每个选项的平均

标记数是多少？**[1 分]**

(1.4) 训练集中每个正确选项的平均标记数是多少？**[1 分]**

(1.5) 对数据进行任何其他探索，只要你认为对这项选择题--回答任务有帮助。简要描述你的发现。

[4 分]

Q2 - 集合相似性测量 [10 分]

使用集合相似性度量，计算每个问题与四个相应答案的相似性得分。您应该使用 Q1 中的标记符。针对每个问题，选出相似度得分最高的答案。

(2.1) 通过测量准确度，报告每种相似性测量（重叠系数、索伦森-戴斯和贾卡德）在训练集和验证集上的表现。**[6 分]**

(2.2) 在每種相似度测量中，有多少次最相似答案的得分與另一個答案的得分相同？当最相似的答案得分相同时，您如何选择？为什么？

问题 3 - TF 向量的余弦相似性 [12 分]

生成每个问题的词频 (TF) 向量以及四个可能的答案。您应该使用默认设置的 CountVectorizer（但要使用与 Q1 和 Q2 中相同的 tokenizer）。对于每个问题，选出 TF 向量与问题 TF 向量余弦相似度最高的答案。

(3.1) 通过测量准确度来报告训练集和验证集的性能。讨论它们与 Q2 中的集合相似度测量方法的比较。**[6 分]**

(3.2) 提出、激励和评估对这一过程的**一项**修改，以改进这一方法。报告在训练集和开发集上的表现，并与未修改版本进行比较。[**6 分**）（提示：您可能需要检查该方法出错的问题，以激发修改的动机。）

Q4 - 来自无伯特基向量的余弦相似性 [12 分]

使用特征提取管道和基于贝尔特的非基化模型，从数据中创建上下文向量。
您应该使用代表 [CLS] 标记的上下文向量作为第一个向量。您应该使用代表 [CLS] 标记的上下文向量，这将是第一个向量。对于每个问题，选取其向量与问题向量之间余弦相似度最高的答案。

(4.1) 通过测量准确度来报告训练集和验证集的性能。**[8 分]**

(4.2) Q2、Q3 和 Q4 中使用的集合相似性和余弦相似性方法有什么局限性？**[4 分]**

问题 5 - 微调变压器模型 [18 分]

在该数据集上使用基于伯特的无基线模型训练自动序列分类模型。这将涉及数据转换，如下所述。您应该只对训练问题进行训练，并使用验证集进行评估。

将数据集转换成一个行表，每个行包含一个问题、一个选项和一个标签（1 或 0）（如果是正确答案）。该表（称为问题-选项对表示法）的行数应是原始问题数据集中问题行数的四倍。将每个问题和选项用"[SEP]"文本连接起来。例如，问题 "日本大阪在哪里 "和错误选项 "大阪城 "将变成标签为 0 的 "日本大阪在哪里[SEP]大阪城"。

在理想情况下，你可以通过超参数调整来确定最佳设置。由于计算成本的原因，使用这些设置应能提供合理的性能：

- 学习率 = 1e-5
- 批量大小 = 8

- 历时 = 4
- weight_decay = 0

(5.1) 报告对训练集和验证集的问题-选项对表示预测的准确度、精确度、召回率和 F1 分数**[10 分]**

(5.2) 报告这种方法在该模型的训练集和验证集上选择正确答案的准确率。请注意，这与第(a)部分中的数值不同。要做到这一点，请为每道题选择模型正类输出对数值最高的选项。**[6 分]**

(5.3) 为什么您认为这种方法优于 Q4 中描述的 [CLS] 向量的使用？[**2 分]**

问题 6 - 测试集性能 [4 分]

(6.1) 报告在测试集上使用最佳方法的准确率。根据验证集上的表现选择最佳方法。**[2 分]**

(6.2) 讨论达到的精度是否足以用于部署 **[2 分]**

未评分的额外内容 [0 分]

这一部分完全是可选项，不会被打分。

下一个可以研究的方法是 AutoModelForMultipleChoice，这是一种特殊的多选题 "拥抱脸 "架构。这需要将数据集精心准备成特定格式。研究相应的教程可能会有所帮助：

https://huggingface.co/docs/transformers/tasks/multiple_choice 合理的超

参数设置如下：

- 学习率 = 5e-5
- 批量大小 = 16
- 历时 = 3
- weight_decay = 0.1