
哈爾濱工業大學
HARBIN INSTITUTE OF TECHNOLOGY

# 信息检索

## 实验二：问答系统设计与实现

School of Computer Science and Technology

Harbin Institute of Technology

# 1 实验目标

本次实验目的是对问答系统的设计与实现过程有一个全面的了解。实验主要内容包括：对给定的文本集合进行处理、建立索引；找出问题的候选答案句并排序；答案抽取，逐步调优。

# 2 实验环境

编程语言为：推荐使用 Python

**本实验中子模块 3.2、3.3、3.4 要求使用机器学习的方法实现**

# 3 实验内容及要求

本次实验中，同学们首先要自己建立一个检索系统，从文本库中检索到与问题最相关的文档（可以是一个或者多个）。然后对文档中的候选答案句进行排序，抽取出最相关的候选答案句。最后，在最相关的候选答案句中抽取最精简的答案，这个答案可能是一个词或者几个词。为了方便同学们使用机器学习/深度学习的模型，我们提供了一部分有标注的数据作为训练集和开发集，同学们需要提交的那部分是去掉了标注的数据，最终我们会通过同学们提交的答案和标准答案的相似度（BLEU-1 值）来评价本次实验的效果。

## 3.1 文本集合进行处理、建立索引

**任务描述：**本实验提供的文本集合保存在 passage_multi_sentences.json 中，每行是一个标准格式的 json 文件，包含文档内容和 id。同学们需要对所有文档分词、分句（已使用上个实验中介绍的 LTP 进行分词、分句操作），并建立索引，作为问答系统的检索语料。检索系统的实现没有具体要求，可以使用开源库，也可以自己实现（**自己实现的检索系统有加分**）。同学们可以根据后面的任务，自己设计合适的建立索引的方法。另外，同学们可以使用有标注的 train.json 数据，来检验自己检索系统的准确性。

**提交要求：**只需要提交相应的程序（prerpocessed.py 或 prerpocessed 文件夹）即可。

## 3.2 问题分类

**任务描述：**问题分类是问题理解中的重要部分，问题类别信息对答案抽取有很大帮助。该

小节实验期望同学们能够训练一个问题分类模型，得到问题类别信息，然后将其融入到候选答案句排序和答案抽取的任务中，以取得更好的效果。同学们可以通过对照实验来分析问题类别信息带来的收益。**该部分实验必须使用机器学习的方法实现。**

问题分类使用的数据是 train_questions.txt 和 test_questions.txt，分类介绍见“哈工大 IR 研究室问题分类体系 ver2.doc”，由于使用的问答数据集上没有问题类别的标注，因此无法用于训练问题分类任务。本实验提供了额外的问题分类数据集（train_questions.txt 和 test_questions.txt，介绍见“哈工大 IR 研究室问题分类体系 ver2.doc”），用于训练问题分类任务，训练得到的模型可用来获取问答数据集中的问题类别信息。关于如何将问题类别信息使用到后续任务中，需要同学们自己设计。

**提交要求 :**提交训练和测试代码(question_classification.py 或者 question_classification 文件夹)

### 3.3 候选答案句排序

**任务描述：**对于每个问题，其对应的相关文档中所有的句子称为候选答案句。同学们需要对所有候选答案句按照**其包含正确的可能性**进行排序，可能性越大的越靠前。为了不将检索系统的误差累计到这一步，同学们直接使用真实的相关文档训练模型，而不需要通过检索系统检索。另外同学们需要从训练集中分出开发集，来同学们可以根据有标注的 train.json 数据来检验自己候选答案句排序系统的有效性。

候选答案句排序（也叫候选答案句抽取）任务在问答系统中是一个经典任务，主流方法是采用机器学习、深度学习的方法构建分类模型或者排序模型。**本实验要求同学们必须使用机器学习的方法完成**，通过结合老师上课讲过的文本特征，以及查阅相关文献，设计模型并完成实验。

**提交要求：**只需要提交相应的程序（answer_sentence_selection.py 或 answer_sentence_selection 文件夹）即可。

### 3.4 答案抽取

**任务描述：**本次实验的目的是从候选答案句中抽取精简的答案片段，这个片段可以是一个短语或者一个词。例如对于问题“姚明出生于哪一年？”，其最相关的候选答案句为“姚明，前中国男篮国家队队员，生于 1980 年 9 月 12 日”，那么我们期望的答案片段为“1980 年”。

答案抽取任务有一定难度，**本实验要求同学们使用机器学习或者基于规则的方法完成**，同学们可以通过查阅文献来找到相应的解决方案。

**提交要求：**提交相应的程序（answer_span_selection.py 或 answer_span_selection 文件夹）

## 4. 实验提交

完成上述所有实验后，同学们需要将检索系统、候选答案句排序模型以及答案抽取模型串起来。对于 test.json 中的每个问题，先通过检索系统得到相关文档，候选答案句排序模型对相关文档中的答案句进行排序，并取出若干最相关的句子（一般是一个），最后通过答案抽取模型抽取片段级别的答案，并保存在 test_answer.json 中并提交。

test_answer.json 格式如下，其中 “qid” 表示 question ID，:“answer_pid” 表示在 passage.json 中检索到相关文档的 ID（数组，长度不超过 3），“answer” 表示最终抽取的精简答案片段（无需分词）。

{ “qid” : 1, “question” : “… …” , “answer_pid” : [10504], “answer” : “2010 年 10 月” }

**注：每条 json 数据占一行！由于该结果文件是脚本自动评测的，因此每个字段名及字段数据类型系要严格遵守上述要求，文件编码为 utf-8。**

我们最终将通过综合考量 test_answer.json 中 answer_pid 的 F1 值和 answer 的 BLEU1，以及使用的算法来对本次实验打分。

本次实验的实验报告请严格按照 “信息检索实验报告模板.docx” 的格式完成，并**导出为 PDF 格式**，按 “学号+姓名+实验 2 问答系统设计与实验报告.pdf” 命名，例如：1140310421_张三+实验 2 问答系统设计与实验报告.pdf。

请同学们按照各个实验模块的提交要求**正确命名提交文件，**然后将所有文件打包命名为 “学号+姓名+实验 2 问答系统设计与实现.zip”，例如：1140310421_张三_实验 2 问答系统设计与实验报告.zip。发送到邮箱 :。

最后提醒同学们，报告或代码发现抄袭现象，该实验部分将按 0 分处理。