第六届“泰迪杯”数据挖掘挑战赛——

C 题：智能阅读模型的构建

一、赛题背景

近年来，自然语言处理（NLP）作为人工智能的一个重要领域得到了飞速发展，并且相关技术及其应用的需求日益广泛。在国家政策的推动下，目前市面上有众多的创业公司，对 NLP 的人才需求相当大。

目前，作为人工智能中自然语言处理的代表产品之一，“智能交互技术”已经逐渐渗透到我们周围的很多产品中。但是很多所谓的智能产品，仅仅可以识别一些特定命令，例如，当输入为“打开 QQ”，就能够启动 QQ，但输入改为“看一下 QQ”，就会毫无反应，更不用说一般的语言交流了。而对于普通大众来说，他们希望机器更加“智能”，能够通过自然语言就可以跟机器交流，让机器为我们服务，最大程度上减少额外学习负担，所以未来自然语言处理的一个发展方向就是如何让用户“更自然”、“更低成本”地实现人与机器的交流。

本赛题聚焦于智能交互在电子书阅读的应用。

日常生活中人们要阅读大量的 txt 文本，其内容可能是小说、教程、文集、词典等。很多情况下我们只是需要从文本中查找某一些片段来解决我们的问题。比如，通过查找法律文献中的一些段落来解决我们的法律疑惑，这时并不需要精读整个法律文献；对于小说，有时候我们也只是想知道其中一些特殊细节，并不想花时间去通读整个小说；因此我们希望智能阅读技术能够在这方面提供一些帮助。下面是两个典型的智能阅读的使用场景：

场景一：

TXT：汽车的说明书

问题：1、怎样打开远光灯？ 2、后排要不要系安全带？

需求：定位到 txt 中能帮我们回答问题的所在行，或者给出明确的答案

场景二：

TXT：《射雕英雄传》小说全文

问题：1、“江南七怪”分别是谁？ 2、九阴真经的作者是谁？

需求：定位到 txt 中能帮我们回答问题的所在行，或者给出明确的答案

本题希望能够构建一个智能的文本挖掘模型，针对自然语言输入的问题，能够根据 txt 内容给出需要的回答。

二、任务设计

本赛题的目标是建立一个理解自然语言问题并检索相关材料的模型，具体来说，就是对于用户提出的问题，能够在给定的文本数据库（可能是一本小说、一本法律、一本说明书等）中找出与之相关的章节、段落甚至是句子。

我们希望构造的模型具有一定的通用性，即对于任意给定的文本文件，都能够根据文本文件来回答相关问题。考虑到问题本身的困难性，我们设计了如下的参考步骤（注：参赛选手可以参考下述步骤来完成赛题，但并不要求完全遵循步骤进行）。

1. 对用户所提的问题进行聚类（也就是总结问题的类型，具体的类别、定义可以参考已有的文献，也可以自行整理），以确定答案的形式。例如：“美国的首都是什么”是一个事实类问题，所期望的回答应该是一个命名实体；“天空为什么是蓝的”也是一个事实类问题，但期望的回答应该是一段长文本；“你觉得广州怎么样？”则是一个非事实类问题，期望的回答也应该是一段文本。

2. 根据给定问题和输入材料（材料一般有足够多的段落，并且大多数都是跟问题不相关的），并结合第一步所得到的问题类别，筛选出跟问题有关的若干段落，并进行排序，参赛选手需要在论文中清晰地论述排序模型的合理性。

3. 在前两点的基础上，逐步定位答案所在句子。考虑到问题种类的多样性（事实类/非事实类、实体类/论述类等），一般情况下，定位到所在的句子即可。选手也可以根据问题的不同种类定制不同颗粒度的方案。

总的来说，基本的要求是模型能够针对问题具有辅助阅读的功能，并在这个基础上尽量精准地回答问题。

方法说明：参赛选手可以将这个问题理解为搜索引擎的改进，也可以理解为一个阅读理解任务，这两方面都已经有很多现成的工作可以参考。总之，对于用

户来说，如果能够识别用户用自然语言提出的问题、并加快用户寻找该问题的答案，那么我们的模型已经产生了价值，参赛选手在设计自己的论文和模型时，需要坚持这个原则。

三、数据与评估

1. 数据集。数据集分为训练集和测试集。

训练集格式如下：

问题	篇章	标签
广州白云山有多高	白云山号称“羊城第一秀”，是广州著名的风景区。方圆 28 平方公里，主峰摩星岭，海拔 382 米。白云山风景秀丽，古迹众多，经历年开发，现有七个游览区：明珠楼、摩星岭、三台岭、鸣春谷、飞鹅岭、云台花园、麓湖。	1
	浙江丽水市白云山 丽水白云山森林公园位于浙南山区、瓯江中游的丽水市北郊 2.5 公里处，以白云山为主的北部山区，面积为 2848 公顷（约 4 万亩）。森林总蓄积量为 17 万立方米。因山阿时有白云涌出，可占晴雨，故名。	0
天空为什么是蓝的	阳光进入大气时，波长较长的色光，如红光，透射力大，能透过大气射向地面；而波长短的紫、蓝、青色光，碰到大气分子、冰晶、水滴等时，就很容易发生散射现象。被散射了的紫、蓝、青色光布满天空，就使天空呈现出一片蔚蓝	1
	小鱼为什么不睡觉?石头为什么不怕疼?小鸟在哪里洗澡?小草为什么不能说话?…童年时的我对世界充满了无限好奇!但是我最好奇的是：天空为什么是蓝色的?	0

测试集为上述格式去掉了标签列。

2. 允许参赛选手自行寻找外部语料作为补充，但需要在论文中注明语料来源。

3. 评价方式：所有参赛选手需要通过自己的模型对测试集进行评分（完全相关为 1，完全没关系为 0），并提交结果文件。对于进入视频答辩的队伍会使用

新的数据集来评估模型的泛化能力，以决出特等奖及以上队伍。