第二届“中电慧治”杯政府治理大数据应用

算法大赛试题

中电科大数据研究院有限公司

二零一九年六月

一、竞赛题目及要求：

竞赛题目：政策公文中的语义搜索

语言要求：主程序 python3

测试运行环境：Linux

二、竞赛题目描述：

1、问题背景

当前，“互联网+”等信息技术与政府治理融合日趋深入，智慧治理作为一种新型的政府治理模式正在推动新一轮的政府转型。在社会由信息化、数字化向智能化、智慧化迈进的环境下，积极推进“大数据+政务服务”，是贯彻落实党中央、国务院决策部署，把简政放权、放管结合、优化服务改革推向纵深的关键环节，对加快转变政府职能，提高政府服务效率和透明度，便利群众办事创业，进一步激发市场活力和社会创造力具有重要意义。

目前，政府各委办局的官方网站上都提供了政策公文检索功能，方便公众查询需要的政策信息。但由于公众普遍无法提出准确的政策关键字，而是直接抛出具体问题，如“请问生育津贴领取的条件是如何？依据的法律规定是什么？”，导致公文检索效果不佳。因此，基于语义搜索算法帮助公众准确检索政策公文与具体条例是十分有价值的。

2、目标问题

需要根据输入的问题，返回对应的政策内容，以及相应的公文名称和具体条例（格式在下面说明）。

3、数据下载及说明

1、根据百度网盘链接及密码，登录网盘进行下载。

2、QQ 群内群文件内进行下载

网盘链接：https://pan.baidu.com/s/1vcXbyjKFs6uSz4Qxe1XRWg

提取码：ysc3

（1）训练集

主办方将分别提供训练数据集和公文条例集。其中，训练数据集包含 4000 条左右原始问题及相应的答案、政策性公文、公文条例。公文条例集包含 290 条左右公文、条例和内容，可用于回答所有训练集和测试集中的问题。两个数据集用于模型训练或知识图谱构建，数据格式采用 excel 格式（UTF-8 编码），参赛者自行划分验证集。数据格式样例如下图所示：

原始问题	参考答案（根据问题，从条例中抽取）	出自公文1	出自条例1
目前医保已经中断3个月，想个人补缴医保应该怎么处理呢？	个体参保人员中断生育保险缴费2个月以上(不含2个月)，视为重新参保，欠费期间的保费不能补缴；城镇职工基本医疗保险欠费4个月以上的视为中断，欠费期间的保费不能补缴。	个体参保人员补缴医疗、生育保险的时限规定	正文

公文名	条例名	内容
办事指南-计划生育手术医疗费结算	申请材料-1	1单位参保职工：填报《成都市生育、计划生育手术医疗费审批表》一式两份并加盖行政公章。手术费原始票据、病情证明、本人身份证、婚姻证明（原件及复印件）。

政策公文主要是成都市社保类政策：养老、医疗、工伤、生育、失业及综合类（可能包含五个险种的一个或多个）。

（2）测试集

由于问题可以根据主办方提供的公文集，人工手动进行回答，所以不予公布。回答测试集问题所需的公文政策及条例、内容均包含在公布的文件之中，不再另行提供。训练集和测试集中的原始问题都具有如下的一些共同难点：

(a)口语的多义问题

(b)推理问题

(c)多条问句的真实意图理解

4、结果提交说明

（1）提交内容

源代码及相关文档（包括思路简述，架构简述，算法包简述等）。

程序入口要求：主文件统一命名为 searcher.py,并能接受一个文件路径的参数，searcher.py 按如下方式组织，主办方将调用此文件进行评分。

```
# -*- coding: utf-8 -*-
import sys
def answer(path_test_file):
    'do something(path_test_file)'
def main(arg1):
    answer(arg1)
if __name__ == '__main__':
    main(sys.argv[1])
```

说明：path_test_file 参数是 test_file 的绝对路径地址，test_file 为 txt 文件格式，utf-8 编码格式，只有问题一列，如下截图：

"目前医保已经中断3个月，想个人补缴医保应该怎么处理呢？"
①农村户口的自由职业该可以购买医疗保险吗②个人办理社保需要满足哪些条件以及办理流程

结果保存要求：参赛队伍将运行的结果存于当前工程下面，结果文件统一命名为：result。result 文件要求为 txt 格式，utf-8 编码格式，输出格式要求：问题答案|出自公文 1|出自条例 1|出自公文 1|出自条例 2……，问题存在多个对应公文和条例的用|分割即可。务必按照这个格式排列，否则

后续因格式问题而影响分数的情况，由队伍自行负责。

（2）提交方式

要求：所有提交结果需打包（tar）上传，命名方式：队名+队长名。

A、打包自行上传到百度网盘，将参赛队伍名、网盘路径、密码发送至作品提交邮箱或私聊发给 QQ 群内赛委会陆老师（1546180093）；（推荐）

B、分解打包，通过作品提交邮箱进行提交 cetcconsultation@163.com;

5、评估指标

（1）意图识别的 F1 值；

（2）答案生成准确率；

（3）算法逻辑结构与稳定性。

（4）加分项：如果参赛队伍的算法中采用了知识图谱技术，将给予一定加分；若图谱的构建采用了非纯人工的方式则将根据构建方法给予额外加分。

三、注意事项：

1、作品提交时间：8 月 12 号 23:59 截止；

2、本次竞赛对象为国内外全日制在校大学生（包括全日制本科、硕士、博士研究生等非 2019 届毕业生），其他类人员如果对比赛感兴趣亦可以参加，可计入排名，但不参与赛事奖励。最终排名产生后需要参赛学生开具所在学校在读证

明，主办方会与学校进行核实，如核实有误，取消其奖励资格。

3、作品初步评选结束后，将抽取排名前 10 名的队伍现场进行答辩，最终的排名将结合初评结果及现场答辩情况得出，如前 10 中有参赛人员不符合赛事要求，抽取次位将延后一位。具体比赛结果将在大数据院公众号上进行公布，同时邮件通知各参赛队队长。

4、技术咨询：比赛开始之后技术人员会加入到赛事 QQ 群，如有技术上相关问题可咨询群中“技术咨询”人员。

技术咨询联系人：技术咨询陈老师（561029113）

技术咨询刘老师（2547789944）

技术咨询电话：15908106057

5、友情提示：本次大赛期间，主办方及合作单位不会以任何形式向参赛者收取费用或押金，敬请各位参赛者知晓并转告，以免受骗损失财物。本赛事最终解释权归中电科大数据院所有。

敬请关注公众号：中电科大数据院。