《软件实践》课程实验报告

暑期学校实验项目：高考志愿填报助手

小组名称	知识图谱构建 A 组						
姓　　名	庄祎	专业	人工智能	班级	091181	学号	09118118
实验时间	2020.8.31-2020.9.23		指导教师	孔祥龙		成绩	

一、实验背景和目的

实验背景：高考志愿填报，以其信息多样性、繁杂性，选择个性化，困扰着许多考生。面对这种情况，亟需高考志愿填报助手。

实验目的：对考生来说，帮助考生更轻松、更个性化地填报大学，做出对他来说的相对最优选择。对小组来说，构建基于 web 开发，基于知识图谱的高考志愿填报系统。对个人来说，锻炼团队协作能力和用代码进行数据处理的能力。

二、小组任务和个人任务

小组任务：以高校为中心，围绕一级学科构建知识图谱，为后续知识推理、图形界面设计打基础。具体流程为：专业消歧；创建可导入 neo4j 的 csv 文件；生成知识图谱；构建显示网页。

个人任务：和组内李春澍、李浩天共同完成专业消歧。专业消歧，即：把不同学校不同叫法的专业统一成一个或多个一级学科中的专业。个人的具体任务有：

1.机器消歧。将李春澍同学经过 Bert 语言模型 encoding 后的词向量采取合适的距离度量，构建专业消歧前后的映射对。并与李浩天、张硕所得的用 Jaccard 距离消歧结果进行比较优劣，采取更优的一方作为初步消歧结果。

2.人工审查。和组内外同学分工，将机器消歧后的约 2800 条结果进行人工审查，修改其中不正确的映射对。

3.映射成表。将得到的映射应用于约 16w 条高校数据中。

三、个人任务需求分析

专业消歧总需求分析：由于各个高校对于相同专业采取了不同的命名方式，我们很难在后续的知识图谱中构建专业有关的联想。所以，通过专业消歧，统一不同专业的名称是很有必要的。输入各个大学专业分数表，输出增加 2 列消歧后专业名、对应专业编号的大学专业分数表。

个人任务具体到每个步骤需求分析：

1.机器消歧：输入李春澍同学用 bert 模型处理后的专业词向量，输出消歧前、后专业组成的映射对。

2.人工审查：输入 1 中消歧得到的专业映射对，输出手动修改后的专业映射对。

3.映射成表。输入 2 中人工审查后的专业映射对（csv 格式），输出输出增加 2 列消歧后专业名、对应专业编号的大学专业分数表。

四、实验过程（需附上关键代码及相关说明）

1.机器消歧：

首先，和李浩天、李春澍讨论某些特殊专业情况的消歧。例如：类似于“理科、文科实验班”映射为原名；“All”映射为“all”；类似于“茅以升班”人名命名的映射为原名；其他情况（例如北航的数据，印象很深，好几次都在这里出错）直接删除处理。由此构建新的一二级学科表（李浩天负责）。

在李春澍同学用 bert 得到每个专业的 encoding 的词向量后，我编写如下函数计算离某专业最近的一二级学科名，作为该专业消歧后的结果。

In [41]:
```
# Contributor: 庄祎

import scipy

def get_min_major(encoded_major):
    min_major = "all"
    temp_dis = 4

    for item in subject1_dict.keys():
        a = scipy.spatial.distance.cosine(encoded_major, subject1_dict[item])
        if temp_dis > a:
            min_major = item
            temp_dis = a

    for item in subject2_dict.keys():
        a = scipy.spatial.distance.cosine(encoded_major, subject2_dict[item])
        if temp_dis > a:
            min_major = item
            temp_dis = a

    return min_major
```

再经过简单的代码处理，即得到初步的映射函数。在此不作截取。

2.人工审查：

机器初步消歧后，发现机器消歧错误很多，例如一对多的消歧关系问题。我们组与其他组成员共同修改映射结果文件。修改前后结果部分见实验结果分析。

同时，因需求变更，学科不再被映射为二级学科，统一映射为一级学科。

3.映射成表：

读取 csv 并将其转化为字典，并进行切片后（代码略），用如下所示代码进行映射，并自动增加相应的一级学科编号。

In [29]:
```
import re
after_disambiguate=[]
after_ID=[]
for major in result.iloc[:,4]:

    if major in reflect_dict.keys():

        temp=reflect_dict[major]
        after_disambiguate.append(temp)

        temp_split = re.split('[;, ]',temp)
        ID_str=""
        for item in temp_split:
            ID_str =ID_str + str(major_to_ID[item]) + ';'
        after_ID.append(ID_str)
    else:
        after_disambiguate.append(major)
        after_ID.append('None')
result['after disambiguate']=after_disambiguate
result['ID']=after_ID
result.to_csv('result_after.csv')
```

五、实验结果与分析

1.机器消歧：

用 bert 结果节选如下，左、右分别为消歧前、后专业结果：

In [45]:
```
disamb_dict
```

Out[45]:
```
{'all': '数量经济学',
 '外国语言文学类': '欧洲语言文学',
 '财政学类': '高等教育学',
 '法学': '法学',
 '工商管理类': '工商管理',
 '国际政治': '国际政治',
 '金融学类': '理论经济学',
 '经济学类': '政治经济学',
 '理科试验班': '教育经济与管理',
 '人力资源管理': '企业管理（含：财务管理、市场营销、人力资源管理）',
 '人文科学试验班': '思想政治教育',
 '社会科学试验班（管理学科类）': '教育经济与管理',
 '社会学类': '社会学',
 '新闻传播学类': '广播电视艺术学',
 '法语（中外合作办学）': '民商法学（含：劳动法学、社会保障法学）',
 '国民经济管理（中外合作办学）': '民商法学（含：劳动法学、社会保障法学）',
 '金融学（中外合作办学）': '民商法学（含：劳动法学、社会保障法学）',
 '绘画（加权成绩）': '力学（可授工学、理学学位）',
 '美术学（艺术管理与策划方向，加权成绩）': '美术学',
```

与李浩天所得的用 Jaccard 距离的结果比较后，发现他们的效果更好（目测得出的公认结论）。故在下一阶段选择基于他们的初步消歧结果进行人工审查。

2.人工审查：

如下图所示，每个人分配完修改条数后，直接在 csv 文件中修改。此为修改后的部分最终映射表。

major (origin)	major2 (disambiguated)
all	all
All	all
人力资源管理	工商管理
人力资源管理（人力资源管理、公共事业管理）	工商管理
人力资源管理（含公共事业管理专业、人力资源管理专业）	工商管理
人力资源管理（国家专项）	工商管理
人力资源管理科	工商管理
人工智能	计算机科学与技术
人文地理与城乡规划	地理学;建筑学
人文地理与城乡规划（汉）	地理学;建筑学
人文地理与城乡规划(单列)	地理学;建筑学
人文地理与城乡规划(南疆单列)	地理学;建筑学
人文地理与城乡规划（藏）	地理学;建筑学
人文社科类	文科实验班
人文科地理科与城乡规划	地理学;建筑学
人文科学试验班	文科实验班
人文科学试验班(人文艺术传播类)	艺术学
人文科学试验班（可选专业：汉语言文学、历史、哲学）	中国语言文学;历史学;哲学
人文科学试验班（可选专业：汉语言文学、历史学、哲学）	中国语言文学;历史学;哲学
人文科学试验班（汉语言文学、历史学、哲学、法学）	中国语言文学;历史学;哲学;法学
人文科学试验班（民考汉）	文科实验班
人文科学试验班(弘毅学堂)	弘毅学堂
人文科学试验班（国学方向）	文科实验班
人文科学试验班（国家专项）	文科实验班
儿科学	临床医学
儿科学(五年)	临床医学
儿科学（五年制）	临床医学

3. 映射成表：

如下图所示，新的 csv 文件中，红色方框多出来的两列就是消歧后的专业名字和专业 ID。

26	清华大学	2019	安徽	文科	all	644	09118101高捷	all	1304;
27	清华大学	2019	黑龙江	理科	all	669	09118101高捷	all	1304;
28	清华大学	2019	黑龙江	文科	all	620	09118101高捷	all	1304;
29	清华大学	2019	河北	理科	all	675	09118101高捷	all	1304;
...	...	...	...	...	...	...	...	...	...
163491	兰州大学	2017	重庆	理科	药学	570	61518431郁航远	药学	1007;
163492	兰州大学	2017	重庆	理科	医学检验技术	573	61518431郁航远	基础医学	1001;
163493	兰州大学	2017	重庆	理科	化学类	590	61518431郁航远	化学	703;
163494	兰州大学	2017	重庆	理科	环境科学与工程类	599	61518431郁航远	环境科学与工程	830;
163495	兰州大学	2017	重庆	理科	计算机类	585	61518431郁航远	计算机科学与技术	812;
163496	兰州大学	2017	重庆	理科	经济学类	590	61518431郁航远	理论经济学;应用经济学	201;202;
163497	兰州大学	2017	重庆	理科	理论与应用力学基地班	577	61518431郁航远	物理学	702;
163498	兰州大学	2017	重庆	理科	临床医学	580	61518431郁航远	临床医学	1002;
163499	兰州大学	2017	重庆	理科	生物科学类	580	61518431郁航远	化学工程与技术	817;

六、实验总结与心得体会

本次实验主要做了数据处理方面的任务。在该任务中，我有如下几点不同方面体会：

1.使用技术方面，虽然我们将人工智能相关方法融入到语义消歧中——用了近几年表现较好的语言模型 bert。然而，在该任务中，效果却没字面相似性（Jaccard 距离）好。这让我了解了该技术的局限性。

2.数据方面，在运行程序过程中，出现了许多来自数据方面的报错；而且海量的数据也带来了运算量的问题（bert 编码与编码后的距离比较）；再就是大量数据也难以考虑周全，需要提前或是编程遇到问题后进行讨论，统一标准（例如特殊专业的消歧问题）。让我体会到了数据处理并不是一个轻松的工作，充满了不确定与系统性的繁琐。

3.团体协作方面，我们几个还是很好的分好了工作，各司其职，并没遇到工作前后冲突等问题。之后的协作可以从这次实践中汲取成功经验。

2020 年 9 月制