《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	李浩天	专业	人工智能	班级	AI 一班	学号	09118111
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

高考志愿填报是每届高考生所要面对的最后一个难题。高考志愿填报要参考考生的高考成绩以及所在省份的排名等元素，并根据考生所处的不同的分数区间选择不同的填报策略。尽管高校每年的分数线都在变化，但其在每一个省份基于排名的定位基本不变，这保证了本实验的可行性。

本次暑期学校的实验项目为高考志愿填报助手，即基于 Web 的高考志愿推荐系统。该系统通过用户所在省份的高考成绩和排名给出推荐的志愿学校，同时设计了多个知识图谱与 AI 算法来提供更多的查询和可视化服务，从而起到辅助考生进行志愿填报的作用。

二、小组任务和个人任务

小组的名称为“知识图谱构建 A 组”，小组任务为以高校为中心，围绕专业以及其所对应的一级学科设计知识图谱，从而为后续知识推理提供可靠的底层支持，完成更多的查询功能。其中我的任务为：构建各个专业到一级学科的一对多映射关系。具体地说，我的任务包括：协调与 AI 算法 C 组负责消歧的同学之间的合作，分配组员任务，完成标准一级学科文件的构建，利用 Jaccard 距离消歧，初步构建专业到学科门类、一级学科、二级学科的映射词典。

三、个人任务需求分析

由于各个高校对于相同专业采取了不同的命名方式，我们很难找出针对同一专业的统计特征和针对不同专业之间的相似关系。为了解决这个问题，我们要做的就是统一命名方式，即完成专业名称到一级学科的映射，从而对专业名称进行消歧。

数据组提供的一级学科文件包含两位代码表示的学科门类以及四位代码表示的一级学科，利用该文件找出各个专业到一级学科的一对多映射关系就是我的任务。为了实现更准确的映射效果，二级学科文件也被用作参考来提高映射的细粒度。具体实验过程如下所述。

四、实验过程（需附上关键代码及相关说明）

1. 首先要观察该映射的定义域与值域的特点。定义域即文件中各个学校专业的名称，而值域就是所有一级学科。通关观察发现，定义域的所有专业命名大致可以分为五类：

   a) 正常命名的专业，例如“土木工程”、“工商管理”、“计算机科学”。该类专业可以直接映射到相应的一级学科。

   b) 实验班。以“实验班”命名的所有专业大体可分为两类。第一类的格式为“x 科实验班”，一共有三类，即“文科实验班”、“理科实验班”、“工科实验班”；第二类在原有 x 科实验班基础上增加了具体的专业名称，例如“工科实验班（电气类）”。对于前者，由于专业名称过于宽泛，无法界定相对应的一级学科，所以我们在原来一级学科的基础上增加了相应的名称，如下图所示。而对于后者，我们是直接根据括号里的专业名称来映射到相关的一级学科。

13	实验班
1301	文科实验班
1302	理科实验班
1303	工科实验班
1304	all
14	xx学院
1401	弘毅学堂
1402	茅以升学院
1403	徐特立英才班

   c) 未细分的专业，主要集中在清华大学与北京大学两所高校，其专业均已“all”命名。处理方法为在“实验班”学科门类下创建名称为“all”的一级学科。

   d) 以人名命名的专业，例如“徐特立英才班”、“茅以升学院”。处理该类专业的方式也和“x 科实验班”一样，人工定义新的一级学科完成映射。

   e) 其他特殊的或无意义的专业名称，如“少数民族预科班”、“合计”（这个专业名称显然是错误的），该类名称被直接删除。

2. 搞清专业的各种命名方式后就要考虑映射的方法。现有的文本相似度匹配方法有很多，我们分别尝试了基于 BERT 的词嵌入向量度量相似度的方法和基于 Jaccard 距离度量相似度的方法。我的任务是实现后者。Jaccard 距离是度量两个集合相似度的方法，其计算方法为两个集合的交集的元素个数除以并集的元素个数。Python 实现如下图所示。

```
11  def JaccardDistance(str1,str2):
12      # s1 = set(jieba.cut(str1))
13      # s2 = set(jieba.cut(str2))
14      s1 = set(str1)
15      s2 = set(str2)
16      if "(" in s1:
17          s1.remove("(")
18      if ")" in s1:
19          s1.remove(")")
20      if "(" in s2:
21          s2.remove("(")
22      if ")" in s2:
23          s2.remove(")")
24      intersection = len(list(s1 & s2))
25      union = len(list(s1 | s2))
26      return intersection / union
```

值得考虑的是，对于专业名称字符串的分割，是以字为单位，还是以词为单位。主观考虑的话应该是后者效果会更好一些，所以我首先尝试用 Jieba 库分别对专业和一级学科分词，计算两个词集合之间的 Jaccard 距离。之后也尝试了以字为单位的 Jaccard 距离的计算，并对比两种方法的结果，最后发现以字为单位的效果要略优于以词为单位的效果。此外，在分割字符串时还要注意去掉“（”“）”，否则很容易产生错误的映射结果。

为了完成专业到一级学科的映射，定义原专业名称到一级学科的直接映射似乎是理所当然的，但在实验过程中发现，相当一部分专业很难对应到一个合适的一级学科。拿“经济学”举例，“经济学”为第二个学科门类，其下有两个一级学科“理论经济学”和“应用经济学”。在某些学校，与经济相关的专业有“金融学”和“财政学”。显然基于 Jaccard 距离的相似度度量方法无法将这些专业映射到正确一级学科中。如果假设这些专业可以映射到“经济学”这个学科门类，但其对应的一级学科究竟是理论经济学还是应用经济学仍然无法决定。这时二级学科就派上了用场。查阅二级学科文件发现“金融学”和“财政学”都属于应用经济学，所以可以由二级学科的所属来间接映射到对应的一级学科。

另外一个问题是，如果某学校（实际也确实存在）关于经济学的专业名称就是“经济学”，那么该专业似乎可以同时映射到理论经济学和应用经济学，这就要求我的映射函数在一些情况下可以完成一对多的映射。但实际上考虑到代码的复杂性，实验中的代码在这种情况下会直接将该专业映射到学科门类“经济学”中。

总而言之，我忽略了学科门类（两位代码）、一级学科（四位代码）、二级学科（六位代码）之间的层次关系，而是将它们视为一类，构成了元素总数为 481 的值域集合，而消歧就是找到原专业与这 481 个元素中 Jaccard 距离最大的元素。完成以上映射后就初步构建好了专业映射的词典，词典中的元素格式（原专业名称，消歧后名称）。相应的代码示例如下。

```
40      # 原专业名称到学科门类、一级学科、二级学科的映射
41      for major in majors:
42          distances = []
43
44          if major =='all' or major =='All':
45              mixed.append("all")
46          else:
47              for name in names:
48                  distances.append(JaccardDistance(major, name))
49              sim_idx = distances.index(max(distances))
50              major_name = names[sim_idx]
51              mixed.append(major_name)
52
53
54              if IDs[sim_idx]<100:
55                  result["0"][counts] = major_name
56              elif 100<IDs[sim_idx]<10000:
57                  result["1"][counts] = major_name
58              else:
59                  result["2"][counts] = major_name
```

```
66      # 词典(Major1(origin), Major2(disambiguated))
67      test = pd.read_csv("test.csv")
68      diction = pd.read_csv("dict.csv")
69      number = len(test["Major"][:])
70      major1 = []
71      major2 = []
72      diction_pair = []
73      for i in range(number):
74          diction_pair.append((test["Major"][i], test["Mixed"][i]))
75
76      diction_pair = set(diction_pair)
77      print(len(diction_pair))
78
79      for pair in diction_pair:
80          major1.append(pair[0])
81          major2.append(pair[1])
82
83      diction["major1"] = major1
84      diction["major2"] = major2
85      pd.DataFrame.to_csv(diction, "dict.csv", encoding="UTF-8")
```

因为任务要求是完成专业名称到一级学科之间一对多的映射，因此还需对这个词典做进一步修正，判断消歧的结果是否是一级学科，具体地说，对于某专业消歧后的名称：

a) 如果该名称属于学科门类，则将该学科门类下的所有一级学科作为最终的映射的结果，实现一对多的映射。

b) 如果该名称属于一级学科，则无需进行任何操作。

c) 如果该名称属于二级学科，则要找到该二级学科对应的一级学科，将其所属的一级学科作为最终的映射的结果。

修正后的词典仍需人工干预。人工干预主要完成三个任务：（1）尽可能补充更多的一对多关系；（2）修改机器错误映射的结果；（3）删除某些特殊的和爬取错误的专业。

3. 最后只要根据构筑好的词典对原“高考录取分数线整合”文件中的专业列进行遍历与映射即完成了专业名称的消歧任务。

五、实验结果与分析

1. 初步构建的词典节选（消歧后的专业同时包含学科门类、一级学科、二级学科）：

工业工程（普通类）	矿业工程
工业工程类	矿业工程
工业工程类（中外合作办学）	工程力学
工业工程类（管理学Ⅰ）	管理学
工业设计	服装设计与工程
工业设计(双语类)	服装设计与工程
工业设计（单列）	服装设计与工程
工业设计（普通类）	服装设计与工程
工科试验班	材料科学与工程
工科试验班(土木与环境类)	环境科学与工程
工科试验班（大数据技术）	科学技术哲学
工科试验班（电子类）	机械电子工程
工科试验班（电气类）	电气工程

2. 最终词典节选（实现了专业名称到一级学科的一对多映射）：

工业工程（普通类）	轻工技术与工程
工业工程类	轻工技术与工程
工业工程类（中外合作办学）	轻工技术与工程
工业工程类（管理学Ⅰ）	轻工技术与工程
工业设计	轻工技术与工程
工业设计(双语类)	轻工技术与工程
工业设计（单列）	轻工技术与工程
工业设计（普通类）	轻工技术与工程
工科试验班	工科实验班
工科试验班(土木与环境类)	土木工程;环境科学与工程
工科试验班（大数据技术）	计算机科学与技术
工科试验班（电子类）	机械电子工程
工科试验班（电气类）	电气工程

3. 应用词典消歧后的结果：

36476	华东政法大学	2016	广东省	理科	金融学	586	应用经济学	202;
36477	华东政法大学	2016	广东省	理科	金融学（金融工程）	580	应用经济学	202;
36478	华东政法大学	2016	广东省	理科	社会学	579	社会学	303;
36479	华东政法大学	2016	广东省	理科	社会学（社会管理）	577	社会学	303;
36480	华东政法大学	2016	广东省	理科	社会工作	576	社会学	303;
36481	华东政法大学	2016	广西壮族自治区	文科	侦查学（经济侦查）	577	军队指挥学	1105;
36482	华东政法大学	2016	广西壮族自治区	文科	侦查学（刑事侦查）	561	军队指挥学	1105;
36483	华东政法大学	2016	广西壮族自治区	文科	治安学	544	工商管理;农林经济管理;公共管理;图书情报与档案管理	1202;1203;1204;1205;
36484	华东政法大学	2016	广西壮族自治区	文科	边防管理	552	工商管理;农林经济管理;公共管理;图书情报与档案管理	1202;1203;1204;1205;
36485	华东政法大学	2016	广西壮族自治区	理科	侦查学（经济侦查）	555	军队指挥学	1105;
36486	华东政法大学	2016	广西壮族自治区	理科	侦查学（刑事侦查）	525	军队指挥学	1105;
36487	华东政法大学	2016	广西壮族自治区	理科	治安学	543	工商管理;农林经济管理;公共管理;图书情报与档案管理	1202;1203;1204;1205;
36488	华东政法大学	2016	广西壮族自治区	理科	边防管理	505	工商管理;农林经济管理;公共管理;图书情报与档案管理	1202;1203;1204;1205;
36489	华东政法大学	2016	广西壮族自治区	文科	法学（民商法律）	606	法学	301;
36490	华东政法大学	2016	广西壮族自治区	文科	法学（刑事法律）	597	法学	301;
36491	华东政法大学	2016	广西壮族自治区	文科	法学（经济法）	602	法学	301;

总体而言，实验效果明显，基本实现了消除歧义的功能，且通过比较发现基于 Jaccard 距离的方法比基于 BERT 的方法（另一组同学实现）效果更好。

六、实验总结与心得体会

本次实验我主要完成并提交了三份文件：标准一级学科文件（包含实验班等门类），消歧代码（原专业名称到学科门类、一级学科、二级学科的映射），以及最终词典。通过这次实验，我完整体验了一次多人项目实现的流程，学会了多人合作下 git 仓库的使用，同时也切身体会到不同团队之间，同一团队的成员之间即时沟通和协商合作的重要性。这次课程的经验将为我以后的项目经历提供很大的帮助。

2020 年 9 月制