数学学院案例分析报告

案例名称：初中历史主观题自动阅卷案例分析

报告负责人：魏丹怡（202032164）

完成时间：2021.05.19

目　录

一、数据来源和基本情况

1、数据来源

该数据集是学生自己收集，关于初中历史主观题的题面以及学生作答以及标准答案和评分。

2、基本情况

此数据集为初中历史主观题答题集，共三个表格分别为题目、作答以及学生。其中题目包括题目 ID、题面出处、题目文本题目分值以及标准答案；作答包括学生 ID、题目 ID 学生作答文本、得分以及得分点；学生包括学生 ID、姓名、年龄、年级、性别以及所属学校。其中作答中的学生 ID 可以连接学生表格，题目 ID 可以连接题目表格。其中题目以及作答表格中都包含 153 条数据，而学生表格中共包含 56 条数据。

	A	B	C	D	E
1	学生id	题目id	该生作答文本	该生作答得分	该生作答得分点
2	202032195	17-1	族，民权，民生	4	分）；民权（1分）；民生
3	202032195	17-2	飞机，火车	2	火车（1分）；飞机（1分）
4	202032195	17-3	《南京条约》中	5	《南京条约》，英国（1
5	202032181	17-1	生	3	分）；民生（1分）
6	202032181	17-2	船	3	分）；汽车（1分）
7	202032181	17-3	中，英国割占了	6	分）；香港岛（1分）；《马

	A	B	C	D	E
1	题目id（唯一）	题面出处	题目文本	题目分值	该题标准答案（含得分点）
2	17-1	https://wenku.baidu.com/view/200e6f1b	列举孙中山领导辛亥革命的指导思想	4	分）；民权（1分）；民生
3	17-2	https://wenku.baidu.com/view/200e6f1b	具	4	分）；汽车（1分）；飞机
4	17-3	https://wenku.baidu.com/view/200	列举《南京条约》《马关条约》中被列强割占	6	《南京条约》，英国（1

	A	B	C	D	E	F
1	学生id（唯一）	姓名	年龄	年级	性别	所属学校
2	202032195	刘家良	23	2020级应用统计专硕	女	西北大学数学学院
3	202032181	张少羿	24	2020级应用统计专硕	女	西北大学数学学院
4	202032191	李帛洋	22	2020级应用统计专硕	女	西北大学数学学院
5	202032175	梁奕宁	23	2020级应用统计专硕	女	西北大学数学学院
6	202032142	彭驰	23	2020级应用统计专硕	男	西北大学数学学院
7	202032162	徐馨宇	22	2020级应用统计专硕	女	西北大学数学学院
8	202032179	文新雨	22	2021级应用统计专硕	女	西北大学数学学院
9	202032143	薛重阳	24	2020级应用统计专硕	男	西北大学数学学院
10	202032194	胡天澍	22	2020级应用统计专硕	男	西北大学数学学院

二、文本匹配技术路线

1、技术路线示意图

为了实现主观题自动阅卷，我们设计了如下技术路线：

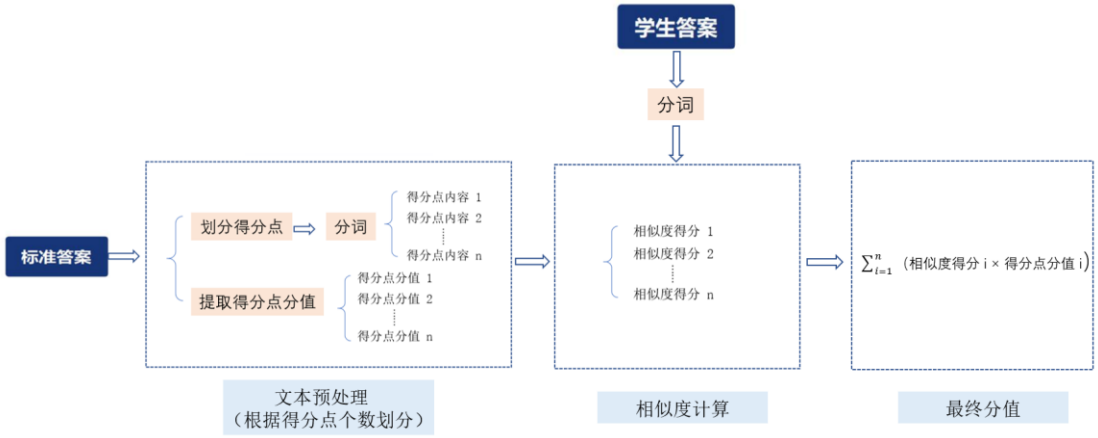

图 1　技术路线

在实现主观题自动阅卷的过程中，最重要的步骤为比对学生作答与标准答案之间的相似度。假设学生给出的答案和标准答案为两段中文段落，我们的目的是计算出这两个中文段落的相似度，而中文文本一般由段落组成，段落根据标点符号可以划分成句子，句子根据分词可以划分成词汇。因此，我的基本思路为：根据两个句子中词汇的相似度计算出句子的相似度，再根据句子的相似度计算出段落的相似度，最后根据段落的相似度计算出文本的相似度。

由于学生作答文本难以统一格式，无法找到合适的字符对其进行划分，因此，我们将学生作答文本视为句子，直接进行分词处理，将其与每个得分点内容进行对比，则每道题可以得到 n 个相似度得分（n 为该题目得分点个数）。

2、技术路线详细描述

（1）分句

在本案例中，我希望按照得分点个数来划分标准答案，因此，我对标准答案的格式有所要求，即每个得分点之后用中文括号标注得分点分值，同时在文本中出现的其他括号为英文括号（这一格式也有利于提取得分点分值）。之后，即可按照中文右括号将便准答案划分为多个句子。

（2）分词

中文分词指的是将一个汉字序列切分成一个一个单独的词。分词就是将连续的字序列按照一定的规范重新组合成词序列的过程。中文分词根据实现特点大致可分为两个类别：基于词典的分词方法、基于统计的分词方法。

我尝试了两种方法对文本进行分词处理，均为基于词典的分词方法，其常见的扫描策略有：正向最大匹配、逆向最大匹配、双向最大匹配和最少词数分词。我们分别调用了 HanLP（Han Language Processing）和 jieba 两个工具包，将句子通过分词划分成多个词语的集合，并通过停用词表去掉没有意义的词语。HanLP 是面向生产环境的多语种自然语言处理工具包，基于 PyTorch 和 TensorFlow 双引擎，目标是普及落地最前沿的 NLP 技术。HanLP 具备功能完善、性能高效、架构清晰、语料时新、可自定义的特点。而 jieba 是目

前最好的 Python 中文分词组件，它支持 3 种分词模式，支持繁体分词并且支持自定义词典。其中，精确模式试图将句子最精确地切开，适合文本分析。使用 jieba.cut 方法进行分词，它所返回的结构是一个可迭代的 generator，可使用 for 循环来获得分词后得到的每一个词语，或者直接使用 jieba.lcut 直接返回 list。在 jieba.cut 接受的 3 个参数中，cut_all 参数是指是否使用全模式；HMM 参数用来控制是否使用 HMM 模型。

HMM 模型也就是隐马尔科夫模型，隐马尔科夫模型是结构最简单的动态贝叶斯网络。描述由一个隐藏的马尔科夫链随机生成不可观测的状态随机序列，再由各个状态生成一个观测而产生随机序列的过程。隐藏的马尔科夫链随机生成的状态的序列称为状态序列，每个状态生成一个观测，称为观测序列。HMM 模型是词性兼类的消歧常采用的概率的方法。所谓词性兼类的消歧，是指在分词中存在词性歧义现象，主要是由词性兼类所引起的。词性兼类是指自然语言中一个词语的词性多余一个的语言现象。对人来说，词性歧义现象比较容易排除，但是对于没有先验知识的机器来说是比较困难的。

隐马尔科夫做了两个基本的假设：

①齐次马尔科夫假设，即假设隐藏的马尔科夫链在任意时刻 t 的状态只依赖于前一时刻的状态，去其他观测状态无关；

②观测独立性假设，即假设任意时刻的观测只依赖于该时刻的马尔科夫链的状态，与其他观测以及状态无关；

隐马尔科夫模型由初始状态概率向量π，状态转移概率矩阵A，以及观测概率矩阵B决定。在词性标注问题中，初始状态概率为每个语句序列开头出现的词性的概率，状态转移概率矩阵由相邻两个单词的词性得到，观测序列为分词后的单词序列，状态序列为每个单词的词性，观测概率矩阵 B 也就是一个词性到单词的概率矩阵。

隐马尔科夫模型有三个基本问题：

概率计算问题，给出模型和观测序列，计算在模型λ下观测序列 O 出现的概率；

②学习问题，估计模型$\lambda =$（A,B,π）参数，使得该模型下观测序列$P(O|\lambda)$最大，也就是用极大似然的方法估计参数；

③观测问题，已知模型 λ 和观测序列 O，求对给的观测序列条件概率$P(I|O)$最大的状态序列 I，即给的观测序列，求最可能的状态序列。

在词性标注问题中，需要解决的是学习问题和观测问题。学习问题即转移矩阵的构建，观测问题即根据单词序列得到对应的词性标注序列。

（3）计算余弦相似度

余弦相似性通过测量两个向量的夹角的余弦值来度量它们之间的相似性。0 度角的余弦值是 1，而其他任何角度的余弦值都不大于 1；并且其最小值是-1。从而两个向量之间的角度的余弦值确定两个向量是否大致指向相同的方向。两个向量有相同的指向时，余弦相似度的值为 1；两个向量夹角为 90° 时，余弦相似度的值为 0；两个向量指向完全相反的方向时，余弦相似度的值为-1。这结果是与向量的长度无关的，仅仅与向量的指向方向相关。余弦相似度通常用于正空间，因此给出的值为-1 到 1 之间。这上下界对任何维度的向量空间中都适用，而且余弦相似性最常用于高维正空间。

（4）计算句子相似度

设两个句子S_1和S_2分词后的词语集合分别为：

$$\{W_{11}, W_{12}, \ldots\ldots, W_{1m}\}$$

$$\{W_{21}, W_{22}, \ldots\ldots, W_{2n}\}$$

由此得出两个句子的相似度矩阵MS，其中W_{1i}、W_{2j}是句子S_1中的词语W_{1i}和句子S_2中的词语W_{2j}的相似度。在计算句子相似度时，首先取矩阵中的最大值MS_1放入序列$maxMS$中，然后将此最大值所在的行和列删除，形成新的矩阵，重复此过程直到矩阵为空，得到最大值序列$maxMS$。则两个句子的相似度可以通过对$maxMS$求合并除以S_1，S_2词语长度的较大值得到。

（5）计算总分

调用 re 库中的函数提取得分点分值，在得到相似度得分并提取每个得分点分值后，我将其对应相乘再求和，即可得到该题学生得分。这样做的好处在于根据学生答案对每一个得分点内容的描述完整程度，给出了学生在该得分点中所得的分数，之后再求学生该题得分总分，充分考虑到了每个得分点分值不同，占该题总分比例不同这一问题。

三、数据预处理

为了可以让算法更好更精确的运行，需要将数据转化成我们需要的格式。由于给出的算法只需要用到学生作答文本、学生作答得分、题目分值、题目标准答案以及题目 ID，为了要对应于学生本人我保留学生 ID 一项，所以我通过 excel 中自带的 powerquery 功能进行合并，共保留七项分别为：题目 ID、题目文本、题目分值、该题标准答案、学生 ID、该生作答文本以及该生作答得分。

1、数据整理

由于后续工作需要用 python 中的函数进行自动的得分点切分，所以需要在数据整理中统一答案所用的符号。利用 Excel 中自动查找替换功能将得分点的括号全部换为中文括号，并删去括号后的符号如句号逗号分号等，最后将其余的括号都改为英文符号。同时删去一些没有在标准答案中标注得分点内容和得分点分值的数据，最后得到 122 条可用数据。处理结果如下图所示：

	A	B	C	D	E	F	G
1	题目id	题目文本	题目分值	该题标准	学生id	该生作答	该生作答
2	17-1	列举孙中	4	三民主义	202032195	三民主义	4
3	17-1	列举孙中	4	三民主义	202032181	民权、民	3
4	17-1	列举孙中	4	三民主义	202032191	辛亥革命	4
5	17-2	列举第一	4	轮船（1分	202032195	飞机，火	2
6	17-2	列举第一	4	轮船（1分	202032181	汽车、火	3
7	17-2	列举第一	4	轮船（1分	202032191	1.第一次	4
8	17-3	列举《南	6	《南京条	202032195	《南京条	5
9	17-3	列举《南	6	《南京条	202032181	在《南京	6
10	17-3	列举《南	6	《南京条	202032191	《南京条	6

2、数据处理

接下来对于已经整理好的原始数据中该题标准答案一项进行得分点切分以及提取各个得分点分值。主要利用 python 中的 re 包库中的 split 函数以及 findall 函数。

首先利用 split 函数针对中文右括号进行分割，将原始的标准答案划分为多个得分点内容。如下图所示：

```
In [125]: jieguo

Out[125]: [['三民主义（1分', '民族（1分', '民权（1分', '民生（1分'],
           ['三民主义（1分', '民族（1分', '民权（1分', '民生（1分'],
           ['三民主义（1分', '民族（1分', '民权（1分', '民生（1分'],
           ['轮船（1分', '火车（1分', '汽车（1分', '飞机（1分'],
           ['轮船（1分', '火车（1分', '汽车（1分', '飞机（1分'],
           ['轮船（1分', '火车（1分', '汽车（1分', '飞机（1分'],
           ['《南京条约》，英国（1分', '香港岛（1分', '《马关条约》，日本（1分', '辽东半岛
           （1分', '台湾（1分', '澎湖列岛（1分'],
           ['《南京条约》，英国（1分', '香港岛（1分', '《马关条约》，日本（1分', '辽东半岛
           （1分', '台湾（1分', '澎湖列岛（1分'],
```

接下来利用 findall 函数，findall 函数可以用来进行模式的发现并提取所需要的相应内容。这里需要提取得分点内容，以及得分点分值两部分，所以利用 findall 函数。首先发现模式一堆文字加一个中文括号加数字再加文字，提取出其中的数字即为我想要的得分点分值，接下来发现一堆文字加一个中文括号并提取其中的文字即为我们需要的得分点内容。同时两者是匹配的且都有 122 条数据，这样的数据处理为我之后实现自动阅卷提供便利。提取结果如下图所示：

```
In [131]: A

Out[131]: [[1, 1, 1, 1],
           [1, 1, 1, 1],
           [1, 1, 1, 1],
           [1, 1, 1, 1],
           [1, 1, 1, 1],
           [1, 1, 1, 1],
           [1, 1, 1, 1, 1, 1],
           [1, 1, 1, 1, 1, 1],
           [1, 1, 1, 1, 1, 1],
```

```
In [132]: B

Out[132]: [['体现的民族精神：视死如归、宁死不屈的民族气节；',
           '不畏强暴、血战到底的英雄气概；',
           '天下兴亡、匹夫有责的爱国情怀；',
           '百折不挠、坚韧不拔的必胜信念。',
           '原因：中国人民巨大的民族觉醒，空前的民族团结和英勇的民族抗争。'],
           ['战争给人类造成深重的灾难和破坏。',
           '第一次世界大战引发了一系列革命运动，俄国十月革命建立第一个社会主义国家。',
           '第一次世界大战改变了资本主义列强的战略战局。欧洲在世界上的中心地位开始动摇，美国和日本迅速崛起。',
           '第一次世界大战在客观上推动了科学技术的发展。'],
           ['迅速发展。',
           '苏联制定并实施了两个五年计划；',
           '苏联人民文化、技术水平的提高；',
           '利用西方经济危机的时刻，大量购买西方先进机器设备，招聘技术家。'],
```

在数据集的标准答案中，有存在任答一点得一分情况，我无法对这种情况进行分句，所以删掉了这种情况的数据。也由此说明，我的分句不够智能，无法识别这种情况，这点是我需要改进的地方。

四、数据探索

1、文本挖掘简介

文本挖掘是抽取有效、新颖、有用、可理解的、散布在文本文件中的有价值知识，并且利用这些知识更好地组织信息的过程。1998 年底，国家重点研究发展规划首批实施项目中明确指出，文本挖掘是“图像、语言、自然语言理解与知识挖掘”中的重要内容。文本挖掘是信息挖掘的一个研究分支，用于基于文本信息的知识发现。文本挖掘利用智能算法，如神经网络、基于案例的推理、可能性推理等，并结合文字处理技术，分析大量的非结构化文本源（如文档、电子表格、客户电子邮件、问题查询、网页等），抽取或标记关键字概念、文字间的关系，并按照内容对文档进行分类，获取有用的知识和信息。

文本挖掘是一个多学科混杂的领域，涵盖了多种技术，包括数据挖掘技术、信息抽取、信息检索，机器学习、自然语言处理、计算语言学、统计数据分析、线性几何、概率理论甚至还有图论。

2、文本挖掘技术的发展

数据挖掘技术本身就是当前数据技术发展的新领域，文本挖掘则发展历史更短。传统的信息检索技术对于海量数据的处理并不尽如人意，文本挖掘便日益重要起来，可见文本挖掘技术是从信息抽取以及相关技术领域中慢慢演化而成的。

随着网络时代的到来，用户可获得的信息包含了从技术资料、商业信息到新闻报道、娱乐资讯等多种类别和形式的文档，构成了一个异常庞大的具有异构性、开放性特点的分布式数据库，而这个数据库中存放的是非结构化的文本数据。结合人工智能研究领域中的自然语言理解和计算机语言学，从数据挖掘中派生了两类新兴的数据挖掘研究领域：网络挖掘和文本挖掘。

网络挖掘侧重于分析和挖掘网页相关的数据，包括文本、链接结构和访问统计（最终形成用户网络导航）。一个网页中包含了多种不同的数据类型，因此网络挖掘就包含了文本挖掘、数据库中数据挖掘、图像挖掘等。文本挖掘作为一个新的数据挖掘领域，其目的在于把文本信息转化为人可利用的知识。

3、文本挖掘步骤

（1）数据集。我们将题目作为我们的文本数据集。

（2）文本预处理。去除标点符号，防止在统计词频时会统计进去。

（3）文本分词。把文本数据集分词。

（4）过滤词表。运用过滤词表优化掉常用词，比如：“的”“了”等。

（5）读取分词。经过循环读出每个分词，再判断每个分词是否在常用词表中或结果是否为空，如果不是，那么是我们优化后的分词。

（6）词频统计。获取前 100 个最高频的词。

（7）绘制词云图。

4、文本挖掘结果

词云图如下所示：

图 2 词云图

5、结论

由上图可以看出，“辛亥革命”、“工业革命”、“孙中山”、“指导思想”、“交通工具”、“新式”、“领导”属于较高频词汇。可以看出同学们在寻找历史题目时倾向于选择近代世界史以及中国史，并且同学们更加关注对于中国历史有很大推动作用的人物以及事件比如云词图中出现的孙中山先生、唐太宗、汉文帝、辛亥革命、新文化运动等。

五、结果分析

1、评价标准

（1）平均绝对误差 MAE：

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|(r_i - \hat{r}_i)|$$

（2）均方根误差 RMSE：

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(r_i - \hat{r}_i)^2}$$

其中 N 表示题目个数，$\hat{r}_i$表示人工给分，r_i模型计算得分。

2、结果展式

（1）模型计算得分

在数据预处理中，已将标准答案分割为各个得分点并提取出了得分点分值，之后我按照计算路线图计算相似度即可，每道题目模型计算得分如下所示：

```
In [136]: r
          2.0,
          1.380952380952381,
          0.0,
          1.368421052631579,
          2.0,
          2.0,
          6.0,
          5.333333333333333,
          1.8666666666666667,
          0.9181818181818182,
          0,
          6.0,
          1.4666666666666668,
          1.75,
          1.7430555555555556,
          0.7976190476190477,
          0.7692307692307693,
          0.25,
          0.5358851674641147,
          1.6666666666666667]
```

（2）模型评价

上图将得到的模型计算得分与提取出的人工给分带入评价标准公式， 得到平均绝对误差 MAE 和均方根误差 RMSE 的值分别为 1.73 和 2.29。

（3）结果可视化

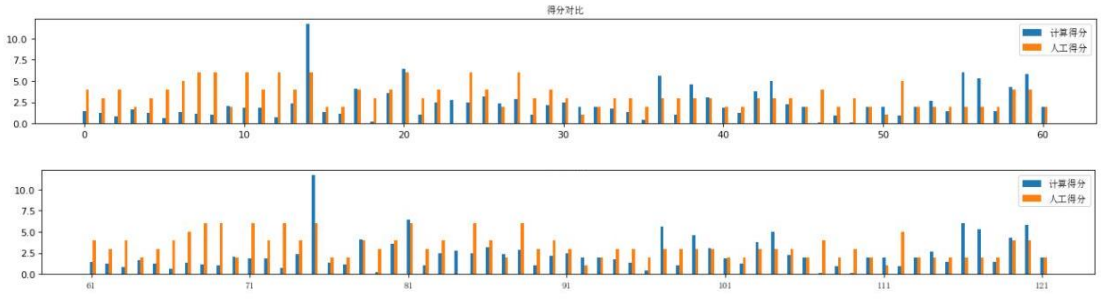

图 3 得分对比

如上图所示，两幅图分别为前 61 道题目和后 61 道题目的得分对比，通过人工给分以及模型计算得分的对比条形图可以看出，模型计算得分有的高于人工给分，有的低于人工给分，在个别题目中差异比较明显，大部分原因在于没有考虑到“任答一点给一分”这类情况，还需进一步改进。

六、实践总结

本次实验实现了主观题自动阅卷，通过学习了自然语言处理相关知识，比如如何分词、如何进行文本间相似度匹配。并通过小组讨论以及查阅相关资料找到切实可行的主观题自动阅卷流程方案，针对该方案所需要的数据样式，对原始初中历史数据集进行数据预处理。将标准答案按照得分点分句，再与学生答案进行匹配，利用余弦相似度得到模型计算得分，最后将平均绝对误差和均方根误差作为评价标准对此次自动阅卷模型进行评价。此外，利用词云图针对于同学们找到的历史问题题目进行数据挖掘，得到一些有用信息。

本次实验中不足在于：在对数据集进行数据预处理时，符号删去还是有一些过于人工，尤其是针对中英文括号的转化问题。需要将分数的括号转化为中文括号，而其他文本中的括号仍然保留为英文括号，这一点无法用 excel 直接实现。在接下来的实践中我可以尝试通过更加灵活地编程语言比如 python 进行实现。并且在数据集的标准答案中，有存在任答一点得一分情况，我无法对这种情况进行分句，所以删掉了这种情况的数据。也由此说明，我的分句不够智能，无法识别这种情况，这点是我需要改进的地方。