# 题目去重项目报告——智慧学伴

## 一、背景介绍

当前学科部门导出的智慧学伴题目数据中存在相同或相似题目，题目去重项目的任务是筛选并去除相同或相似的题目。

由于当前题目数据为未标注数据，用于评估题目相似的算法尚未确定且对于相似题目的判定依据尚未确定，因此去重项目的任务可分为两个子任务：

* 确定最优的题目相似度评估方法。
* 通过相似度评估方法在题目数据上的结果，结合学科部分的经验，确定划分题目相似的阈值。

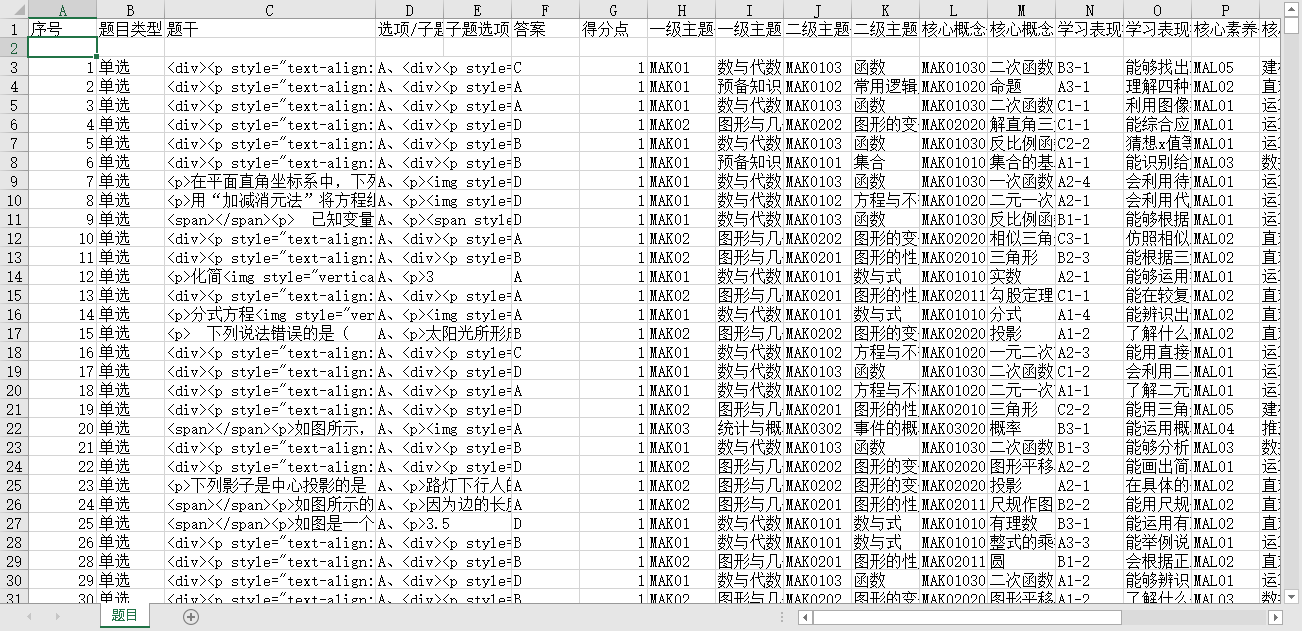
本报告内容主要分为三部分：

1. 介绍对智慧学伴题目数据的预处理。
2. 选出最优的题目相似度评估方法。采用的策略是将题目数据在常用的文本相似度方法上进行评估，对评估结果采用人工评价的方式进行比较，筛选出我们认为效果最好的相似度评估方法。
3. 题目去重。采用选基于编辑距离的评估方法对预处理后的题目进行题目去重。

## 二、数据预处理

**题目数据处理流程：**

1. 从网页导出智慧学伴题目数据，格式为.xls，存储在raw\_data作为原始题库数据文件。



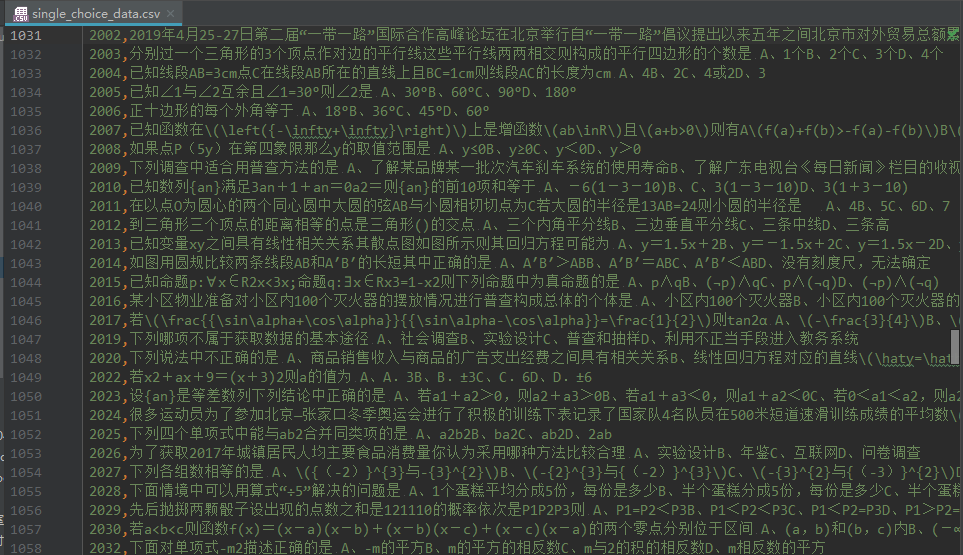
1. 根据科目的不同采用不同的解析文件进行解析，从html格式的数据中解析获取题目信息，生成的数据保存在data\_source中。如通过src/preprocess/phrase\_math\_data.py解析数学题目数据。



1. 将data\_source中的数据通过src/preprocess/split.py区分处理，结果保存在data\_split中，当前暂不考虑套题等情况，因此仅处理“正常”sheet的数据。



1. 将data\_split中的数据进行预处理，去除不合法字符及无意义字符，按照题目类型存放在不同的csv文件中，生成的结果存放在data/preprocess\_result\_xxx/中。如数学科目的数据存放在preprocess\_result\_math。



下面根据科目的不同，介绍不同科目数据不同的文本处理规则及数据介绍：预处理操作可以包括以下操作中的一种或多种：去掉网页标记、将各种类型的空格转换为中文空格、 将各种字符统一至同一编码(如UTF-8编码)、去除不合法字符及字符集外的乱码字符、当前我们对题目中的图片暂不做处理，因此去除题目中的图片信息。这几种操作并无时间先后顺序。

### 2.1数学数据

数据介绍：

数学题目共6677条。根据题目类型的标签，可以分为单选、填空、主观题三种。此外对于单选题及填空题处理过程中发现，有的单选题目没有选项信息，我们将其作为一种新题型，即无选项题。故数学题目类型共有四种不同题型。

数据特征：

* 数学题含有公式，需要区分处理，统一将题干分为公式部分与描述部分，将选项分为公式部分与描述部分。
* 单选题与填空题的判断同时考虑选项或答案信息，如单选题干内容为“下列正确的是”，因此对于单选题，将其选项的内容也作为题目部分内容进行比较；
* 主观题及无选项题目的比较仅考虑题干部分。题干部分需要考虑其中的公式部分与描述部分。

文本预处理规则：

1. 解析html源数据，去除html标签。
2. 对于空格、字符归一，例如题目中可能存在字符“〖^4〗”，“〖^9〗”等字符， 要将其归一化到“4”，“9”；将标点符号例如“【】”转化为“ [ ]”；将各种类型的空格转换为中文空格；统一编码格式为UTF-8。
3. 去除空格、换行、下划线及不合法字符如“□”等。
4. 去除图片信息如\_\_img321.png。
5. 去除题目中【教师题库】及（2018九下广东中考）这种不利于相似度比较的字符串。
6. 去除题目中的空括号及括号中有一个或多个空格的情况，如：（）（ ）()等。
7. 去除不规则的信息如：“2016九上海淀期末”。
8. 去除题目描述末尾的多余或者不匹配符号如：“=”、“？”、“（”、“[”等。
9. 对于选项及答案信息，对空格、字符归一，去除空格、换行符、下划线及不合法字符。
10. 对于选项及答案信息，去除选项处理后仅有序号的无意义情况如“ABCD”、“A、B、C、D、”、“A、B、C”等。
11. 对于单选题目，去除题干长度小于4的情况；对于其他类型题目，去除题干长度小于3的情况。

### 2.2语文数据

数据介绍：

语文题目共3060条。根据题目类型的不同分别有单选、填空、主观题、多选、判断、无选项题六类，其中无选项题目中为单选、多选没有选项的题目以及填空题没有答案的题目。

数据特征：

* 单选、多选、填空、判断题将选项信息或答案信息也作为题目信息一起进行比较。
* 主观题与无选项题只比较题干信息，主观题的答案会影响题目的判定，因此不考虑答案。

文本预处理规则：

1. 解析html源数据，去除html标签。
2. 对于空格、字符归一，例如题目中可能存在字符“〖^4〗”，“〖^9〗”等字符， 要将其归一化到“4”，“9”；将标点符号例如“【】”转化为“ [ ]”；将各种类型的空格转换为中文空格；统一编码格式为UTF-8。
3. 去除空格、换行、下划线及不合法字符如“□”等。
4. 去除图片信息如\_\_img321.png。
5. 去除题目中(教师题库)及（2018.西城二模）(2分)这种不利于区分的字符串。
6. 去除题目中的空括号如：（）、（ ）、()等。
7. 去除题目描述末尾的多余或者不匹配符号如：“=”、“？”、“（”、“[”等。
8. 对于选项及答案信息，对空格、字符归一，去除空格、换行符、下划线及不合法字符。
9. 对于选项及答案信息，去除选项处理后仅有序号的无意义情况如“ABCD”、“A、B、C、D、”、“A、B、C”等。
10. 对于单选题目，去除题干长度小于4的情况；对于其他类型题目，去除题干长度小于3的情况。

### 2.3历史数据

数据介绍：

历史题目共6000条。根据题目类型的不同分别有单选、填空、主观题、多选、判断、无选项题六类，其中无选项题目中为单选、多选没有选项的题目以及判断题与填空题没有答案的题目。

数据特征：

* 单选、多选、填空、判断题将选项信息或答案信息也作为题目信息一起进行比较。
* 主观题与无选项题只比较题干信息，主观题的答案会影响题目的判定，因此不考虑答案。

文本预处理规则：

1. 解析html源数据，去除html标签。
2. 对于空格、字符归一，例如题目中可能存在字符“〖^4〗”，“〖^9〗”等字符， 要将其归一化到“4”，“9”；将标点符号例如“【】”转化为“ [ ]”；将各种类型的空格转换为中文空格；统一编码格式为UTF-8。
3. 去除空格、换行、下划线及不合法字符如“□”等。
4. 去除图片信息如\_\_img321.png。
5. 去除题目中(教师题库)及（2018.西城二模）(2分)这种不利于区分的字符串。
6. 去除题目中的空括号如：（）、（ ）、()等。
7. 去除题目描述末尾的多余或者不匹配符号如：“=”、“？”、“（”、“[”等。
8. 对于选项及答案信息，对空格、字符归一，去除空格、换行符、下划线及不合法字符。
9. 对于选项及答案信息，去除选项处理后仅有序号的无意义情况如“ABCD”、“A、B、C、D、”、“A、B、C”等。
10. 对于单选题目，去除题干长度小于4的情况；对于其他类型题目，去除题干长度小于3的情况。

### 2.4物理数据

数据介绍：

物理题目共9999条。根据题目类型的不同分别有单选、填空、主观题、多选、判断、无选项题六类，其中无选项题目中为单选、多选没有选项的题目以及主观题与填空题没有答案的题目。

数据特征：

* 物理题目的处理与数学一样，考虑题目中可能出现公式的情况。
* 单选、多选、填空、判断题将选项信息与答案信息也作为题目信息一起进行比较。其中单选多选考虑选项信息，填空判断考虑答案信息。
* 主观题答案会影响判断，因此与无选项题一样只考虑题干信息

文本预处理规则：

1. 解析html源数据，去除html标签。
2. 对于空格、字符归一，例如题目中可能存在字符“〖^4〗”，“〖^9〗”等字符， 要将其归一化到“4”，“9”；将标点符号例如“【】”转化为“ [ ]”；将各种类型的空格转换为中文空格；统一编码格式为UTF-8。
3. 去除空格、换行、下划线及不合法字符如“□”等。
4. 去除图片信息如\_\_img321.png。
5. 去除题目中【教师题库】及（2018九下广东中考）这种不利于相似度比较的字符串。
6. 去除题目中的空括号及括号中有一个或多个空格的情况，如：（）（ ）()等。
7. 去除不规则的信息如：“2016九上海淀期末”。
8. 去除题目描述末尾的多余或者不匹配符号如：“=”、“？”、“（”、“[”等。
9. 对于选项及答案信息，对空格、字符归一，去除空格、换行符、下划线及不合法字符。
10. 对于选项及答案信息，去除选项处理后仅有序号的无意义情况如“ABCD”、“A、B、C、D、”、“A、B、C”等。
11. 对于单选题目，去除题干长度小于4的情况；对于其他类型题目，去除题干长度小于3的情况。

### 2.5化学数据

当前没有实际处理化学数据，根据学科部门的意见，参考数学科目来处理。

### 2.6道德与法律数据

当前没有实际处理道法数据，根据学科部门的意见，参考语文科目来处理。

### 2.7地理数据

当前没有实际处理地理数据，根据学科部门的意见，参考语文科目来处理。

### 2.8生物数据

当前没有实际处理生物数据，根据学科部门的意见，参考语文科目来处理。

### 2.9英语数据

数据介绍：

英语题目共9466条（不包含套题）。根据题目类型的不同分别有单选、填空、主观题、判断、无选项题五类，其中无选项题目中为单选、判断、填空题没有答案的题目。

数据特征：

* 英语中套题情况较多，对套题的处理在后面介绍。
* 听力题先不做查重，因为听力提问的问题基本一致，而且无法做语音判断，不太适合查重。听力题可以根据是否有video来判断，如：

1，“测试设备音频，请同学们在正式答题时务必点击此音频调试耳机设备 {video:97f7ae03-cf19-460f-b5c1-3edbe7a8e99d}  听录音，完成下面各题”

2，“听第7段材料，回答第8至9题。{video:c9840da5-bf9b-4dae-aadc-9ff1aed009c5}”

* 单选、填空、判断题将选项信息与答案信息也作为题目信息一起进行比较。其中单选多选考虑选项信息，填空判断题考虑答案信息。
* 主观题答案会影响题目的判断，因此与无选项题一样只考虑题干信息。
* 英语题目的处理中，不可以去除空格。

文本预处理规则：

1. 解析html源数据，去除html标签。
2. 对于空格、字符归一，例如题目中可能存在字符“〖^4〗”，“〖^9〗”等字符， 要将其归一化到“4”，“9”；将标点符号例如“【】”转化为“ [ ]”；将中文括号“（）”统一转换为英文括号“()”。
3. 将各种类型的空格转换为英文空格；统一编码格式为UTF-8。
4. 去除换行符号、下划线符号。
5. 去除不合法字符如“□”、“x0a\x0b\x0c\x0d\x0e\x0f\xc2\xa0” 等。
6. 去除题目中的听力题，听力题的特征为题目描述中包含“{video:97f7ae03-cf19-460f-b5c1-3edbe7a8e99d}”、“听”、“请听”。
7. 去重题目中的“（2014北京中考）”类信息
8. 去除题目中的空括号及括号中有一个或多个空格的情况，如：（）（ ）()等。
9. 去除题目描述末尾的多余或者不匹配符号如：“=”、“?”、“（”、“[”等。
10. 对于选项及答案信息，对空格、字符归一，去除空格、换行符、下划线及不合法字符。
11. 对于选项及答案信息，去除选项处理后仅有序号的无意义情况如“ABCD”、“A、B、C、D、”、“A、B、C”等。对于选项或者答案信息，去除其中的图片的信息。
12. 对于单选题目，去除题干长度小于4的情况；对于其他类型题目，去除题干长度小于3的情况。

## 三、相似度算法效果比较：

对于文本相似度比较的算法，主要有两种思路，一种基于字符进行比较，另一种基于向量比较。

下面两部分工作，分别是从七种相似度计算方法中比较，选取效果较好的两种；另外一部分是比较这两种方法的效果，进而确定最优的评估方法。

### 3.1比较七种不同的文本相似度算法：

#### 3.1.1基于字符比较的方法：

* Jaccard 通过两个字符串的交集并集的比例来计算相似度。取值区间[0,1]越接近1越相似。
* Levenshtein编辑距离 通过比较两个字符串的编辑距离进行比较。对于编辑距离的评估，取值区间x>=0越接近0表明两个字符串越相似，这里为了便于计算采用Levenshtein ratio来评估，取值区间[0,1]，越接近1表明字符串越相似。
* BM25 将句子s1分词后，对每个词与句子s2计算相关性，最后求和计算两个句子的相关性。取值区间x>=0越大越相似。
* Difflib 基于LCS的思路计算字符匹配， 取值区间[0,1]越接近1越相似。

#### 3.1.2基于向量比较：

取值区间[0,1]越接近1表明语句越相似。

* Word2vec + cosine
* Tf-idf + cosine
* Bert + cosine

#### 3.1.3实验结果分析：

1，通过向量计算相似度算法时间复杂度高，耗时长，效果普遍较差。

2，Word2vec + cosine评估效果较差，结果依赖于向量表示。由于数学题干中存在很多公式表示，因此分词效果较差， 转换为向量后进行比较的效果也较差，如下面的例子：

相似度：0.9999

题目1：497 下列函数中的偶函数是

题目2：905 下列函数中不是偶函数的是

3，tf-idf +cosine同样依赖于分词后的结果，但效果也不好，不符合我们对相似题目的定义，如下面的两个例子：

相似度：0.9995

题目1：975 下列说法正确的是

题目2：944 在同一平面内，有三条直线a、b、c，下列说法中正确的是

相似度：0.9995

题目1：628 下列说法正确的是

题目2：360 集合A＝{2，－1}，B＝{m²－m，－1}，且A＝B，则实数m＝()

4，通过字符来评估相似度的方法：时间复杂度较小，效率高。结果符合我们判断题干内容是否相似的定义。

5，Jaccard方法最为简单，但效果不好，由于将字符串拆开一一比较，忽略了字符间的次序信息，表现较差。

6，BM25算法效果最差，该方法基于统计的思想，结果不符合我们对相似题目判定的定义。

7，Difflib方法与编辑距离Levenshtein\_ratio方法效果较好，符合我们对相似题目判定的需求。下面重点比较这两种方法的效果。

### 3.2基于LCS的difflib方法与基于编辑距离的Levenshtein\_ratio的效果比较：

基于第一部分的结论，我们单独对基于LCS的difflib方法与基于编辑距离的Levenshtein\_ratio方法做评估，比较在三种不同题目类型数据上的效果。

对于题目描述完全一样的情况，两种方法计算的相似值均为1，没有差别，主要比较在题目描述字符串中有少量字符差异的情况，两种方法计算的文本相似度的值的比较。这里采用difflib与Levenshtein\_ratio值做评估，两种方法计算结果的取值都在[0,1]，越接近1认为两个字符串越相似。

效果比较过程：分别对三种不同题型数据采用两种方法计算相似度的值。

在单选题的比较中，我们发现编辑距离对字符变化更加敏感，便于区分两个字符串的差别。比如例子1中，题干内容一样，选项内容一样，但是选项顺序不一样，编辑距离方法认为这两个题目的相似度为0.89，difflib方法评估两个字符串的相似度为1，也就是认为两个题目相同。

同样在例2 例3中，同样可以发现，difflib对于字符的变化不敏感，对于题目相似度的评估均比编辑距离高。

在填空题与主观题数据中的表现也是相同情况。

例1：

Levenshtein\_ratio：0.896551724137931,difflib\_value: 1.0

2073 下列各选项中，线段的长度能组成直角三角形的是（）（）A、1，√2，√3B、2，3，4C、√2，√3，5D、3，5，6

2229 下列各选项中，线段的长度能组成直角三角形的是（）（）A、2，3，4B、1，√2，√3C、√2，√3，5D、3，5，6

例2：

Levenshtein\_ratio：0.8074534161490683,difflib\_value: 0.8571428571428571

2459 在长为18cm的线段AB上任取一点M，并以线段AM为边作正方形，则这个正方形的面积介于36cm2与81cm2之间的概率为A、5/6B、1/2C、1/3D、1/6

2888 在长为10cm的线段AB上任取一点P，并以线段AP为边作正方形，这个正方形的面积介于25cm2与49cm2之间的概率为A、3/10B、1/5　C、2/5D、4/5

例3：

Levenshtein\_ratio：0.7681159420289855,difflib\_value: 0.9565217391304348

1376 下列条件中，不能判定一个四边形是平行四边形的是（）A、两组对边分别平行B、两组对边分别相等C、两组对角分别相等D、一组对边平行且另一组对边相等

2211 下列条件中，不能判定四边形是平行四边形的是A、两组对边分别平行B、一组对边平行，另一组对边相等C、两组对边分别相等D、两组对角分别相等

**结论：经过比较发现基于字符的方法整体比基于向量的方法效果较好。其中基于编辑距离的Levenshtein\_ratio方法比基于基于LCS的difflib方法好。下面的工作中，我们将采用编辑距离的方法来评估题目的相似度。**

### 3.3基于编辑距离的Levenshtein介绍：

编辑距离是指由原字符串变化到另外一个字符串所需要的最小编辑次数或最小代价，编辑即是对字符串指定位置的单个字符进行插入、删除和替换的操作。

考虑长度分别为m和n的两个字符串str1和str2，构造矩阵，利用动态规划方法获得该矩阵元素即

在该矩阵中，矩阵右下角元素LD(n,m)即为字符串str1和str2之间的编辑距离。

例如字符串“abc”与字符串“ac”的编辑距离为1，表示字符串“abc”经过一次操作删除b即可变成字符串“ac”；同样的字符串“kitten”与字符串“sitting”的编辑距离为3，表示字符串“kitten”变成字符串“sitting”最少需要操作三次操作分别是将k替换为s，e替换i，添加g。

项目中使用python开发，直接调用python-Levenshtein库，来计算两个字符串的编辑距离，为了便于评估比较两个字符串的相似度，我们使用python-Levenshtein库的ratio()方法，该方法的算法原理是通过计算公式 ：

r = (sum - ldist) / sum

其中sum是指str1 和 str2 字串的长度总和，ldist是类编辑距离。

类编辑距离与编辑距离的差别在于，编辑距离中三种操作中每个操作均算作依次操作，距离值加一，而类编辑距离中，删除、插入操作距离值加一，但是替换操作距离值加二。（这样设计的目的在于：如计算ratio('a', 'c')时，sum=2,按照原始编辑距离通过公式计算为（2-1）/2 = 0.5，a','c'没有重合，显然不合算，但是采用类编辑距离替换操作距离值加二，（3-1）/2 = 1就可以解决这个问题。）

将两个待比较的字符串作为输入，返回这两个字符串的相似度r。相似度值r的取值范围为[0,1]，0表示两个字符串完全不相似，1表示两个字符串完全相同，越靠近1表示两个字符串越相似。

例如下面三个例子展示基于编辑距离计算得到的不同题目之间的相似度r：

1，相似度: 1.0

str1: 3046 已知向量a=（1m）b=（21）且a⊥b则m

str2: 3177 已知向量a=（1m）b=（21）且a⊥b则m

2，相似度: 0.7586

str1: 3026 抛物线\({y^2}=4x\)上到其焦点的距离为\(1\)的点的个数为

str2: 3299 抛物线y²=4x上到其焦点的距离为1的点的个数为

3，相似度: 0.5806

str1: 3115 \(-\frac{3}{2}\)的相反数是绝对值是倒数是

str2: 3941 3.2的相反数是倒数是绝对值是

## 四、题目去重的阈值设置：

经过上面的工作，我们对预处理后的数据进行评估，根据题目科目的不同，有不同的处理策略。

### 4.1数学题目

#### 4.1.1单选题、填空题

由于数学题目中含有大量的公式，需要单独处理，我们将数学的单选题分为四部分进行比较，分别考虑题目的题干与选项两部分信息，对于题干和选项分别考虑其中的描述部分与公式部分。

比较策略：

比较两个题目是否相似需要依次比较对应的各部分，其中比较的优先级是先比较题干再比较选项/答案信息，对于每部分先比较公式信息，再比较题干描述的文字信息，对于选项也是先比较公式再比较选项描述。

对于选择题的四部分数据需要依次进行确定，每次都会有三种情况，因此全部可能出现的情况为3\*3\*3\*3=81种情况。下面表格中行列分别表示两个待比较的题目，待比较的部分为对角线表格中的内容。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 题干公式 | 题干描述 | 选项公式 | 选项描述 |
| 题干公式 | 相似/不确定/不相似 |  |  |  |
| 题干描述 |  | 相似/不确定/不相似 |  |  |
| 选项公式 |  |  | 相似/不确定/不相似 |  |
| 选项描述 |  |  |  | 相似/不确定/不相似 |

每个待比较的项目根据设定的两个阈值，一个为阈值上限，一个为阈值下限，我们规定相似度大于阈值上限的情况则判断为相似，小于阈值下限的则判断为不相似，在阈值上限与阈值下限之间的情况设置为不确定，通常需要人工判定。

**经过与学科部门的讨论，阈值设定为0.95与0.8。具体规则如下：**

1两个题干均含有公式的情况：

题干公式相似0.95及以上 题干描述相似0.95及以上 选项公式相似0.95及以上 选项描述相似0.95及以上，判定相似

题干公式相似0.95及以上 题干描述相似0.95及以上 选项公式相似0.95及以上 选项描述不确定0.8-0.95，判定相似

题干公式相似0.95及以上 题干描述相似0.95及以上 选项公式相似0.95及以上 选项描述不相似0.8以下，判定不相似

题干公式相似0.95及以上 题干描述相似0.95及以上 选项公式不确定0.8-0.95， 选项描述相似0.95及以上，判定相似

题干公式相似0.95及以上 题干描述相似0.95及以上 选项公式不确定0.8-0.95，选项描述不确定0.8-0.95，人工判定

题干公式相似0.95及以上，题干描述相似0.95及以上，选项公式不相似0.8及以下，判定不相似

题干公式相似0.95及以上 题干描述不确定0.8-0.95 选项公式相似0.95及以上，判定相似

题干公式相似0.95及以上 题干描述不确定0.8-0.95 选项公式不确定0.8-0.95，人工判定

题干公式相似0.95及以上 题干描述不确定0.8-0.95，选项公式不相似0.8及以下 ，判定不相似

题干公式相似0.95及以上 题干描述相似0.95及以上 选项都没有公式且相似，判定相似

题干公式相似0.95及以上 题干描述相似0.95及以上 选项都没有公式且不确定，判定相似

题干公式相似0.95及以上 题干描述相似0.95及以上 选项都没有公式且不相似，判定不相似

题干公式相似0.95及以上 题干描述不确定0.8-0.95 选项公式相似0.95及以上，判定相似

题干公式相似0.95及以上 题干描述不确定0.8-0.95 选项公式不确定0.8-0.95，人工判定

题干公式相似0.95及以上 题干描述不确定0.8-0.95，选项公式不相似0.8及以下 ，判定不相似

题干公式相似0.95及以上 题干描述不相似0.8及以下 ，判定不相似

题干公式不确定0.8-0.95 题干描述相似 0.95及以上，参考如下判定方式：

在此基础上仍需继续判定选项描述及公式，两个当中如存在一个不确定或一个相似，则认为相似，如果存在2个及以上不确定，则人工判定；如存在任何一个不相似的情况，则判定不相似。

题干公式不确定0.8-0.95 题干描述不确定 0.8-0.95，参考如下判定方式：

如选项描述及公式存在一个不相似，则认为不相似，其他情况人工判定

题干公式不确定0.8-0.95 题干描述不相似0.8及以下，判定不相似

2对于题干或选项中出现的一个有公式一个没有公式的情况均判定为需要人工判定的情况

3对于两个题干都没有公式的情况：

题干描述相似，选项公式相似，选项描述相似，判定为相似

题干描述相似，选项公式相似，选项描述不确定，判定为相似

题干描述相似，选项公式相似，选项描述不相似，判定为不相似

题干描述相似，选项公式不确定，选项描述相似，判定为相似

题干描述相似，选项公式不确定，选项描述不确定，人工判定

题干描述相似，选项公式不确定，选项描述不相似，判定不为相似

题干描述不确定，选项公式相似，人工判定

题干描述不确定，选项公式不确定，人工判定

题干描述不确定，选项公式不相似，判定不相似

选项都不包含公式：

题干描述相似，选项描述相似，判定相似

题干描述相似，选项描述不确定，判定相似

题干描述相似，选项描述不相似，判定不相似

题干描述不确定，选项描述相似，判定相似

题干描述不确定，选项描述不确定，人工判定

题干描述不确定，选项描述不相似，判定不相似

#### 4.1.2对于主观题、无选项题的处理：

对数学的主观题、无选项题，在实际处理的过程中发现，主观题的答案会扰乱结果的准确性，只通过题干判定即可，题干信息的比较包括分别比较题干的公式部分及题干描述两部分。

比较两个题目是否相似需要依次比较对应的各部分，其中比较的优先级是先比较公式信息，再比较题干描述的文字信息，与单选题的阈值设定规则一样，设定两个阈值。

对于主观题、无选项题的两部分数据需要依次进行确定，每次都会有三种情况，因此全部可能出现的情况为3\*3=9种情况。下面表格中行列分别表示两个带比较的题目，待比较的部分为对角线表格中的内容。

|  |  |  |
| --- | --- | --- |
|  | 题干公式 | 题干描述 |
| 题干公式 | 相似/不确定/不相似 |  |
| 题干描述 |  | 相似/不确定/不相似 |

**经过与学科部门的讨论，阈值设定为0.95与0.8。具体规则如下：**

1两个题目都有公式

题干公式相似0.95及以上 题干描述相似0.95及以上 ，判定相似

题干公式相似0.95及以上 题干描述不确定 ，判定相似

题干公式相似0.95及以上 题干描述不相似0.8及以下 ，判定不相似

题干公式不确定0.8-0.95 题干描述相似0.95及以上，判定相似

题干公式不确定0.8-0.95 题干描述不确定0.8-0.95 ，判定人工核定

题干公式不确定0.8-0.95 题干描述不相似0.8及以下，判定不相似

题干公式不相似小于0.8 判定不相似

2一个题干有公式一个没有公式的情况需要人工判定

3两个题干都没有公式：

题干描述相似，判定相似

题干描述不确定，人工判定

题干描述不相似，判定不相似

### 4.2语文题目

#### 4.2.1单选题、多选、填空、判断的处理：

对于语文的单选题、多选题，实际比较中发现选项包含较多的信息，因此我们在比较的时候同时比较题干与选项信息，且语文题目中一般不考虑公式的情况，因此仅比较其文字描述部分即可。

对于语文的填空和判断题，基于同样的原因，将题目的答案部分也进行比较。因此对于语文题目的比较，会比较题目的两部分信息，分别是题干信息与选项/答案信息两部分。

比较策略与数学题目的思路一致，先比较题干的描述，再比较选项的描述。依次比较题目的题干与选项，因此全部可能出现的情况为3\*3=9种情况，如下表：

|  |  |  |
| --- | --- | --- |
|  | 题干描述 | 选项/答案描述 |
| 题干描述 | 相似/不确定/不相似 |  |
| 答案描述 |  | 相似/不确定/不相似 |

与数学题目的处理流程一样，在阈值设定的策略也一致。

**经过与学科部门的讨论，阈值设定为0.9及0.8。具体规则如下：**

题干描述相似0.9及以上 答案描述相似0.9及以上，认为相似

题干描述相似0.9及以上 答案描述不确定0.8-0.9，认为相似

题干描述相似0.9及以上 答案描述不相似0.8以下，不相似

题干描述不确定0.8-0.9 答案描述相似0.9及以上，认为相似

题干描述不确定0.8-0.9 答案描述不确定0.8-0.9，人工核定

题干描述不确定0.8-0.9 答案描述不相似0.8以下，不相似

题干描述不相似0.8及以下，认为不相似。

备注：语文题目的填空部分有一部分属于古诗默写方面的题目，当前还没找到合适的处理方法。通过核心概念维度信息的“优秀诗文”可以部分判定为古诗默写类题目，但是对于题目相似的判定情况需要进一步讨论。

#### 4.2.2对于主观题、无选项题的处理：

对于语文的主观题及无选项题，我们仅考虑题目的题干信息。

比较策略与数学题目的思路一致，只需要比较题干的相似度情况，根据给出的两个阈值将比较结果，全部可能出现的情况为3种情况，如下表：

|  |  |
| --- | --- |
|  | 题干描述 |
| 题干描述 | 相似/不确定/不相似 |

阈值设定策略与上面相同，由于当前这种情况的题目数据较少，与学科讨论后，**经过与学科部门的讨论，阈值设定为0.75-0.9。具体规则如下：**

题干描述相似0.9及以上，判定相似

题干描述不确定 0.75-0.9，人工校验

题干描述不相似 :0.75以下，判定不相似

### 4.3历史题目

#### 4.3.1单选题、多选、填空、判断题的处理：

对于历史题目的的单选题、多选题，填空和判断题在比较的时候同时比较题干与选项信息或者答案信息，且历史题目中一般不考虑公式的情况，因此仅比较其文字描述部分即可。

比较策略与语文题目的思路一致，先比较题干的描述，再比较选项的描述。依次比较题目的题干与选项，因此全部可能出现的情况为3\*3=9种情况，如下表：

|  |  |  |
| --- | --- | --- |
|  | 题干描述 | 选项/答案描述 |
| 题干描述 | 相似/不确定/不相似 |  |
| 答案描述 |  | 相似/不确定/不相似 |

与语文题目的处理流程一样，在阈值设定的策略也一致。

**根据与学科部门的沟通，阈值设定为0.8-0.9：具体规则如下：**

题干描述相似0.9及以上 答案描述相似0.9及以上，认为相似

题干描述相似0.9及以上 答案描述不确定0.8-0.9，认为相似

题干描述相似0.9及以上 答案描述不相似0.8以下，不相似

题干描述不确定0.8-0.9 答案描述相似0.9及以上，认为相似

题干描述不确定0.8-0.9 答案描述不确定0.8-0.9，人工核定

题干描述不确定0.8-0.9 答案描述不相似0.8以下，不相似

题干描述不相似0.8及以下，认为不相似。

#### 4.3.2对于主观题，无选项题的处理：

对于历史题目的主观题及无选项题，答案判定会扰乱结果的准确性，只通过题干判定即可，比较策略与语文题目的思路一致，只需要比较题干的相似度情况，根据给出的两个阈值将比较结果，全部可能出现的情况为3种情况，如下表：

|  |  |
| --- | --- |
|  | 题干描述 |
| 题干描述 | 相似/不确定/不相似 |

阈值设定策略与上面相同。

**根据与学科部门的沟通，阈值设定为0.75-0.9：具体规则如下：**

题干描述相似0.9及以上认为相似

题干描述相似0.75-0.9 人工校验

题干描述相似0.75以下，不相似

### 4.4物理题目

#### 4.4.1单选题、多选、填空、判断的处理：

物理题目的处理中与数学题一样需要考虑其中的公式，我们将物理的单选题，多选题、填空、判断分为四部分进行比较，分别考虑题目的题干与选项两部分信息，对于题干和选项分别考虑其中的描述部分与公式部分。

比较策略与数学题目一致，对于题目的四部分数据需要依次进行确定，每次都会有三种情况，因此全部可能出现的情况为3\*3\*3\*3=81种情况。下面表格中行列分别表示两个待比较的题目，待比较的部分为对角线表格中的内容。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 题干公式 | 题干描述 | 选项公式 | 选项描述 |
| 题干公式 | 相似/不确定/不相似 |  |  |  |
| 题干描述 |  | 相似/不确定/不相似 |  |  |
| 选项公式 |  |  | 相似/不确定/不相似 |  |
| 选项描述 |  |  |  | 相似/不确定/不相似 |

阈值设定清空与数学一致。

**经过与学科部门的讨论，阈值设定为0.95与0.8。具体规则如下：**

1两个题干均含有公式的情况：

题干公式相似0.95及以上 题干描述相似0.95及以上 选项公式相似0.95及以上 选项描述相似0.95及以上，判定相似

题干公式相似0.95及以上 题干描述相似0.95及以上 选项公式相似0.95及以上 选项描述不确定0.8-0.95，判定相似

题干公式相似0.95及以上 题干描述相似0.95及以上 选项公式相似0.95及以上 选项描述不相似0.8以下，判定不相似

题干公式相似0.95及以上 题干描述相似0.95及以上 选项公式不确定0.8-0.95， 选项描述相似0.95及以上，判定相似

题干公式相似0.95及以上 题干描述相似0.95及以上 选项公式不确定0.8-0.95，选项描述不确定0.8-0.95，人工判定

题干公式相似0.95及以上，题干描述相似0.95及以上，选项公式不相似0.8及以下，判定不相似

题干公式相似0.95及以上 题干描述不确定0.8-0.95 选项公式相似0.95及以上，判定相似

题干公式相似0.95及以上 题干描述不确定0.8-0.95 选项公式不确定0.8-0.95，人工判定

题干公式相似0.95及以上 题干描述不确定0.8-0.95，选项公式不相似0.8及以下 ，判定不相似

题干公式相似0.95及以上 题干描述相似0.95及以上 选项都没有公式且相似，判定相似

题干公式相似0.95及以上 题干描述相似0.95及以上 选项都没有公式且不确定，判定相似

题干公式相似0.95及以上 题干描述相似0.95及以上 选项都没有公式且不相似，判定不相似

题干公式相似0.95及以上 题干描述不确定0.8-0.95 选项公式相似0.95及以上，判定相似

题干公式相似0.95及以上 题干描述不确定0.8-0.95 选项公式不确定0.8-0.95，人工判定

题干公式相似0.95及以上 题干描述不确定0.8-0.95，选项公式不相似0.8及以下 ，判定不相似

题干公式相似0.95及以上 题干描述不相似0.8及以下 ，判定不相似

题干公式不确定0.8-0.95 题干描述相似 0.95及以上，参考如下判定方式：

在此基础上仍需继续判定选项描述及公式，两个当中如存在一个不确定或一个相似，则认为相似，如果存在2个及以上不确定，则人工判定；如存在任何一个不相似的情况，则判定不相似。

题干公式不确定0.8-0.95 题干描述不确定 0.8-0.95，参考如下判定方式：

如选项描述及公式存在一个不相似，则认为不相似，其他情况人工判定

题干公式不确定0.8-0.95 题干描述不相似0.8及以下，判定不相似

2对于题干或选项中出现的一个有公式一个没有公式的情况均判定为需要人工判定的情况

3对于两个题干都没有公式的情况：

题干描述相似，选项公式相似，选项描述相似，判定为相似

题干描述相似，选项公式相似，选项描述不确定，判定为相似

题干描述相似，选项公式相似，选项描述不相似，判定为不相似

题干描述相似，选项公式不确定，选项描述相似，判定为相似

题干描述相似，选项公式不确定，选项描述不确定，人工判定

题干描述相似，选项公式不确定，选项描述不相似，判定不为相似

题干描述不确定，选项公式相似，人工判定

题干描述不确定，选项公式不确定，人工判定

题干描述不确定，选项公式不相似，判定不相似

选项都不包含公式：

题干描述相似，选项描述相似，判定相似

题干描述相似，选项描述不确定，判定相似

题干描述相似，选项描述不相似，判定不相似

题干描述不确定，选项描述相似，判定相似

题干描述不确定，选项描述不确定，人工判定

题干描述不确定，选项描述不相似，判定不相似

#### 4.4.2对于主观题的处理：

对于物理题目的主观题及无选项题答案判定会扰乱结果的准确性，只通过题干判定即可，因此我们仅考虑题目的题干信息，题干信息的比较包括分别比较题干的公式部分及题干描述两部分。

比较策略与数学题目的思路一致，比较两个题目是否相似需要依次比较对应的各部分，其中比较的优先级是先比较公式信息，再比较题干描述的文字信息，

每个待比较的项目均需要给出两个阈值，其中一个为阈值上限，一个为阈值下限，我们规定相似度大于阈值上限的情况则判断为相似，小于阈值下限的则判断为不相似，在阈值上限与阈值下限之间的情况设置为不确定，需要工作人员人工校验题目的相似度。

对于主观题、无选项题的两部分数据需要依次进行确定，每次都会有三种情况，因此全部可能出现的情况为3\*3=9种情况。下面表格中行列分别表示两个带比较的题目，待比较的部分为对角线表格中的内容。

|  |  |  |
| --- | --- | --- |
|  | 题干公式 | 题干描述 |
| 题干公式 | 相似/不确定/不相似 |  |
| 题干描述 |  | 相似/不确定/不相似 |

**经过与学科部门的讨论，阈值设定为0.95与0.8。具体规则如下：**

1两个题目都有公式

题干公式相似0.95及以上 题干描述相似0.95及以上 ，判定相似

题干公式相似0.95及以上 题干描述不确定 ，判定相似

题干公式相似0.95及以上 题干描述不相似0.8及以下 ，判定不相似

题干公式不确定0.8-0.95 题干描述相似0.95及以上，判定相似

题干公式不确定0.8-0.95 题干描述不确定0.8-0.95 ，判定人工核定

题干公式不确定0.8-0.95 题干描述不相似0.8及以下，判定不相似

题干公式不相似小于0.8 判定不相似

2一个题干有公式一个没有公式的情况需要人工判定

3两个题干都没有公式：

题干描述相似，判定相似

题干描述不确定，人工判定

题干描述不相似，判定不相似

#### 4.4.3对于无选项题的处理：

针对no choice题，由于看不见图，不做“相似”判定，只做“人工判定”和“不相似”判定，其中人工判定阈值为0.8及以上，0.8以下为不相似

同样比较题干的两部分内容，包括公式与内容描述部分：

1两个题目都有公式

题干公式相似0.8及以上 题干描述相似0.8及以上 ，判定人工校验

题干公式相似0.8及以上 题干描述不相似0.8及以下 ，判定不相似

题干公式不相似小于0.8 判定不相似

2一个题干有公式一个没有公式的情况需要人工判定

3两个题干都没有公式：

题干描述相似0.8及以上，判定人工判定

题干描述不相似小于0.8，判定不相似

### 4.5化学题目

化学题目的处理与物理及数学一致，参考以上部分

### 4.6道德与法律题目

道德与法律题目的处理与语文及历史一致，参考以上部分

### 4.7地理题目

地理题目的处理与语文及历史一致，参考以上部分。

### 4.8生物题目

生物题目的处理与语文及历史一致，参考以上部分。

### 4.9英语题目

#### 4.9.1单选题、填空、判断题的处理：

对于英语题目的的单选题、填空和判断题在比较的时候同时比较题干与选项信息或者答案信息，英语仅比较其文字描述部分即可。

比较策略与语文题目的思路一致，先比较题干的描述，再比较选项的描述。依次比较题目的题干与选项，因此全部可能出现的情况为3\*3=9种情况，如下表：

|  |  |  |
| --- | --- | --- |
|  | 题干描述 | 选项/答案描述 |
| 题干描述 | 相似/不确定/不相似 |  |
| 答案描述 |  | 相似/不确定/不相似 |

与语文题目的处理流程一样，在阈值设定的策略也一致。

**经过与学科部门的讨论，阈值设定为0.9-0.95。具体规则如下：**

题干描述相似0.95及以上 答案描述相似0.95及以上，认为相似

题干描述相似0.95及以上 答案描述不确定0.9-0.95，认为相似

题干描述相似0.95及以上 答案描述不相似0.9以下，不相似

题干描述不确定0.9-0.95 答案描述相似0.95及以上，认为相似

题干描述不确定0.9-0.9 5答案描述不确定0.9-0.95，人工核定

题干描述不确定0.9-0.95 答案描述不相似0.9以下，不相似

题干描述不相似0.9及以下，认为不相似。

#### 4.9.2对于主观题，无选项题的处理：

对于英语题目的主观题及无选项题，答案判定会扰乱结果的准确性，只通过题干判定即可，比较策略与语文题目的思路一致，只需要比较题干的相似度情况，根据给出的两个阈值将比较结果，全部可能出现的情况为3种情况，如下表：

|  |  |
| --- | --- |
|  | 题干描述 |
| 题干描述 | 相似/不确定/不相似 |

阈值设定策略与上面相同。

**根据与学科部门的沟通，阈值设定为0.9-0.95：具体规则如下：**

题干描述相似0.95及以上认为相似

题干描述相似0.95-0.9 人工校验

题干描述相似0.9以下，不相似

#### 4.9.3对于套题的处理：

针对套题，如阅读题、完形填空题建议采用三层查重，当前的题目中其实是有套题的（由于智慧学伴设计问题，只在套题的第一题显示阅读文本材料，其他题被动变成了单选题）。

共生成合格的套题1796条。我们在处理过程中采取自动补全阅读文本材料并对id重新编码。如下面示例：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 9012 | 套题 | 阅读短文，根据短文内容，从方框中的五个选项中选出能填入空白处的最佳选项。每个选项只能用一次。（共6题） | He showed【1】courage at difficult times, which helped him to succeed. | [{"index":[1],"value":["extraordinary"]}] |
|  |  |  | The color blue in some culture【1】bad mood. | [{"index":[1],"value":["represents"]}] |
|  |  |  | She had always planned to train her daughter in playing some【1】, but her daughter showed no interest in music at all. | [{"index":[1],"value":["instrument/instruments"]}] |
|  |  |  | Some medicine has a side【1】. So before you take it, ask your doctor for advice. | [{"index":[1],"value":["effect"]}] |
|  |  |  | I am sorry if this is going to【1】you, but you failed the test. | [{"index":[1],"value":["disappoint"]}] |
|  |  |  | I was【1】sad when I heard the news of her departure. I am afraid I might never see her again. | [{"index":[1],"value":["extremely"]}] |

|  |  |  |  |
| --- | --- | --- | --- |
| 9012\_0 | 阅读短文根据短文内容从方框中的五个选项中选出能填入空白处的最佳选项每个选项只能用一次（共6题） | He showedcourage at difficult times, which helped him to succeed | extraordinary |
| 9012\_1 | 阅读短文根据短文内容从方框中的五个选项中选出能填入空白处的最佳选项每个选项只能用一次（共6题） | The color blue in some culturebad mood | represents |
| 9012\_2 | 阅读短文根据短文内容从方框中的五个选项中选出能填入空白处的最佳选项每个选项只能用一次（共6题） | She had always planned to train her daughter in playing some, but her daughter showed no interest in music at all | instrument/instruments |
| 9012\_3 | 阅读短文根据短文内容从方框中的五个选项中选出能填入空白处的最佳选项每个选项只能用一次（共6题） | Some medicine has a side. So before you take it, ask your doctor for advice | effect |
| 9012\_4 | 阅读短文根据短文内容从方框中的五个选项中选出能填入空白处的最佳选项每个选项只能用一次（共6题） | I am sorry if this is going toyou, but you failed the test | disappoint |
| 9012\_5 | 阅读短文根据短文内容从方框中的五个选项中选出能填入空白处的最佳选项每个选项只能用一次（共6题） | I wassad when I heard the news of her departure. I am afraid I might never see her again | extremely |

备注：如果是阅读材料里的选择，直接判断阅读文本是否相似即可，不用再判断题目，功能实现的前提上这样操作，“套题”先判断文本整体相似度，再借助单选的判定规则来确定，参考下表

|  |  |  |  |
| --- | --- | --- | --- |
|  | <=0.9 | 0.9<值<0.95 | >=0.95 |
| 阅读文本的整体相似度 |  |  |  |
| 题干本身相似度 |  |  |  |
| 选项/答案 |  |  |  |

对套题是否相似，依次比较阅读文本、题干、选项/答案三部分的相似度；具体规则如下：

**有一项小于0.9，则认为不相似**

**有两项及以上在0.9-0.95之间，则人工判定**

**有一项在0.9-0.95之间，两项在0.95以上，则认为相似**

**有两项或以上在0.95以上，则认为相似**

## 附录：

### 题目去重阈值选取策略-学科实验室研讨-20200419.docx

#### 题目去重阈值选取的策略

#### 题目去重认为中根据科目不同分别对语文和数学题目做不同处理

##### 1.1对于数学题目的处理：

根据题目类型的不同分别有单选 填空 主观题三种。此外对于单选题目，经过文本预处理后，若发现题目没有选项信息，则将其作为一种新题型，即无选项题目。故数学题目类型共有四种不同题型。

###### 1.1.1对于单选题的处理：

考虑题目的题干与选项两部分信息，每部分别包含题干描述与公式两部分。

比较两个题目是否相似需要依次比较对应的各部分，其中比较的优先级是先比较公式信息，再比较题干描述的文字信息，对于选项也是先比较公式再比较选项描述。

每个待比较的项目均需要给出两个阈值，其中一个为阈值上限，一个为阈值下限，我们规定相似度大于阈值上限的情况则判断为相似，小于阈值下限的则判断为不相似，在阈值上限与阈值下限之间的情况设置为不确定，需要工作人员人工校验题目的相似度。

对于选择题的四部分数据需要依次进行确定，每次都会有三种情况，因此全部可能出现的情况为3\*3\*3\*3=81种情况。下面表格中行列分别表示两个带比较的题目，待比较的部分为对角线表格中的内容。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 题干公式 | 题干描述 | 选项公式 | 选项描述 |
| 题干公式 | 相似/不确定/不相似 |  |  |  |
| 题干描述 |  | 相似/不确定/不相似 |  |  |
| 选项公式 |  |  | 相似/不确定/不相似 |  |
| 选项描述 |  |  |  | 相似/不确定/不相似 |

当前一项为不相似的时候，就不需要比较后面的信息，因此需要考察的情况如下，通过下面的比较情况来分别确定题干公式，题干描述，选项公式与选项描述四部分的八个阈值。

此类需人工判定

题干公式相似0.95及以上 题干描述相似0.95及以上 选项公式相似0.95及以上 选项描述相似0.95及以上，判定相似

题干公式相似0.95及以上 题干描述相似0.95及以上 选项公式相似0.95及以上 选项描述不确定0.8-0.95，判定相似

题干公式相似0.95及以上 题干描述相似0.95及以上 选项公式相似0.95及以上 选项描述不相似0.8以下，判定不相似

题干公式相似0.95及以上 题干描述相似0.95及以上 选项公式不确定0.8-0.95， 选项描述相似0.95及以上，判定相似

题干公式相似0.95及以上 题干描述相似0.95及以上 选项公式不确定0.8-0.95，选项描述不确定0.8-0.95，人工判定

题干公式相似0.95及以上，题干描述相似0.95及以上，选项公式不相似0.95及以上，判定不相似

题干公式相似0.95及以上 题干描述不确定0.8-0.95 选项公式相似0.95及以上，判定相似

题干公式相似0.95及以上 题干描述不确定0.8-0.95 选项公式不确定0.8-0.95，人工判定

题干公式相似0.95及以上 题干描述不确定0.8-0.95，选项公式不相似0.8及以下 ，判定不相似

题干公式不确定0.8-0.95 题干描述相似 0.95及以上，参考如下判定方式：

在此基础上仍需继续判定选项描述及公式，四个当中如存在一个不确定或一个相似，则认为相似，如果存在2个及以上不确定，则人工判定；如存在任何一个不相似的情况，则判定不相似

题干公式不确定0.8-0.95 题干描述不确定 0.8-0.95，参考如下判定方式：

如选项描述及公式存在一个不相似，则认为不相似

其他情况人工判定

题干公式不确定0.8-0.95 题干描述不相似0.8及以下，判定不相似

###### 1.1.2对于填空 主观题 无选项（来自单选题）的处理：

对数学的填空题及主观题，在实际处理的过程中发现，填空及主观题的答案包含的信息较少，且一般通过题干的描述就可以对其进行区分，因此我们仅考虑题目的题干信息，题干信息的比较包括分别比较题干的公式部分及题干描述两部分。

比较两个题目是否相似需要依次比较对应的各部分，其中比较的优先级是先比较公式信息，再比较题干描述的文字信息，

每个待比较的项目均需要给出两个阈值，其中一个为阈值上限，一个为阈值下限，我们规定相似度大于阈值上限的情况则判断为相似，小于阈值下限的则判断为不相似，在阈值上限与阈值下限之间的情况设置为不确定，需要工作人员人工校验题目的相似度。

对于选择题的两部分数据需要依次进行确定，每次都会有三种情况，因此全部可能出现的情况为3\*3=9种情况。下面表格中行列分别表示两个带比较的题目，待比较的部分为对角线表格中的内容。

|  |  |  |
| --- | --- | --- |
|  | 题干公式 | 题干描述 |
| 题干公式 | 相似/不确定/不相似 |  |
| 题干描述 |  | 相似/不确定/不相似 |

需要考察的情况如下，通过下面的比较情况来分别确定题干公式，题干描述的四个比较阈值。

题干公式相似0.95及以上 题干描述相似0.95及以上 ，判定相似

题干公式相似0.95及以上 题干描述不确定 ，判定相似

题干公式相似0.95及以上 题干描述不相似0.8及以下 ，判定不相似

题干公式不确定0.8-0.95 题干描述相似0.95及以上，判定相似

题干公式不确定0.8-0.95 题干描述不确定0.8-0.95 ，判定人工核定

题干公式不确定0.8-0.95 题干描述不相似0.8及以下，判定不相似

##### 1.2对于语文题目的处理：

根据题目类型的不同分别有单选 填空 主观题 多选 判断 无选项题六类，其中无选项题目中为单选 多选没有选项的题目以及主观题与填空题没有答案的题目。

###### 1.2.1对于单选题 多选 填空 主观题的处理：

对于语文的单选多选，实际比较中发现选项包含较多的信息，因此我们在比较的时候同时比较题干与选项信息，且语文题目中一般不考虑公式的情况，因此仅比较其文字描述部分即可。

对于语文的填空和主观题，基于同样的原因，将题目的答案部分也进行比较。因此对于语文题目的比较，会比较题目的两部分信息，分别是题干信息与选项/答案信息两部分。

比较策略与数学部分的相同，阈值设定的策略也相同。

因此全部可能出现的情况为3\*3=9种情况

|  |  |  |
| --- | --- | --- |
|  | 题干描述 | 选项/答案描述 |
| 题干描述 | 相似/不确定/不相似 |  |
| 答案描述 |  | 相似/不确定/不相似 |

需要考察的情况如下，通过下面的比较情况来分别确定题干描述，答案描述的四个比较阈值。

**备注：古诗默写方面存在较大争议，**

题干描述相似0.9及以上 答案描述相似0.9及以上，认为相似

题干描述相似0.9及以上 答案描述不确定0.8-0.9，认为相似

题干描述相似0.9及以上 答案描述不相似0.8以下，不相似

题干描述不确定0.8-0.9 答案描述相似0.9及以上，认为相似

题干描述不确定0.8-0.9 答案描述不确定0.8-0.9，人工核定

题干描述不确定0.8-0.9 答案描述不相似0.8以下，不相似

###### 1.2.2对于判断题 无选项题的处理：

对于判断题及无选项的单选多选及无答案的填空与主观题，我们仅考虑题目的题干信息。

比较两个题目的时候只需要比较题干的相似度情况，根据给出的两个阈值将比较结果，分为三种情况分别是相似，不确定相似情况，不相似。

阈值设定策略与上面相同，需要比较的情况有三种，

|  |  |
| --- | --- |
|  | 题干描述 |
| 题干描述 | 相似/不确定/不相似 |

备注：由于只有4道题，暂时基于经验确定

题干描述相似 0.95及以上

题干描述不确定 ：0.8-0.95之间

题干描述不相似 :0.8以下

**待研讨的难题（暂未解决）：**

**1.语文默写类题目按此判断不太准确，偏向人工核查，是否能识别默写类题目，默写类题目的标识，通过“优秀诗文”的核心概念抽炼所有题目？再验证。**

**2.数学题干和选项（答案），一个带公式，一个不带公式，偏向人工核查，核验latex和mathtype的一致性，联系技术请教？**

**总体调整方向：**

**1.在核准人工判定的基础上进一步确定转为机器判定的范围；**

**2.进一步验证其他相关学科的可行性：考虑参考数学，判定下物理（10000+题目）准确性；考虑参考语文，判定下历史（6000+题目）的准确性；**

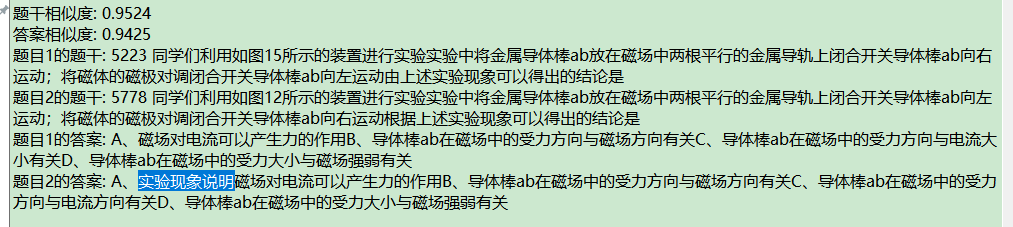
### 物理、历史查重反馈-学科实验室20200426.docx

#### 一、物理

1.填空题0.8-0.95判断标准即可，部分不一致，不太影响

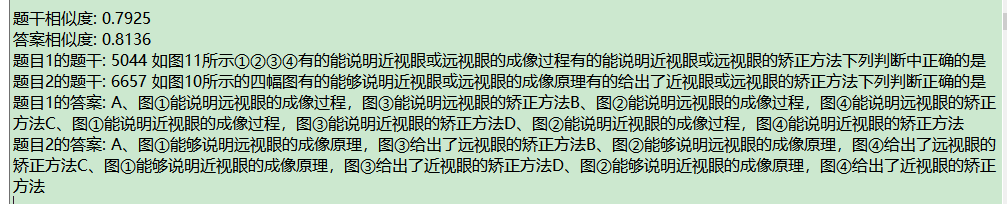
以下为填空题：

以下判定为相似，其实不相似：



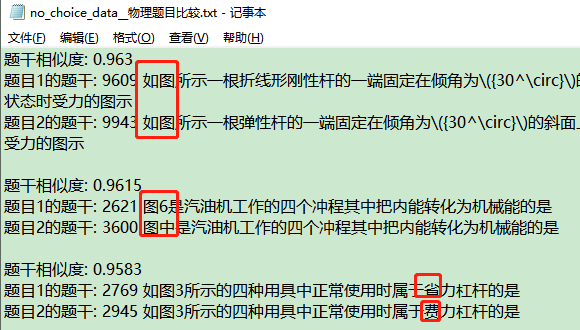
下方为多选题：

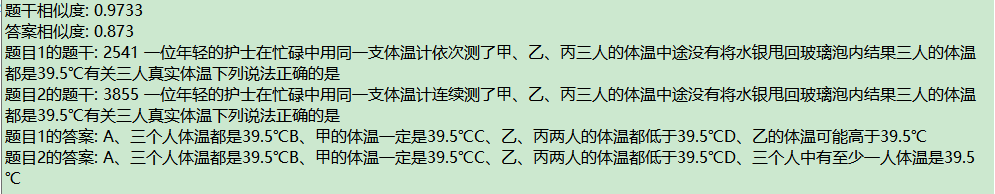
以下判定为不相似，但由于图片未知，实为不相似

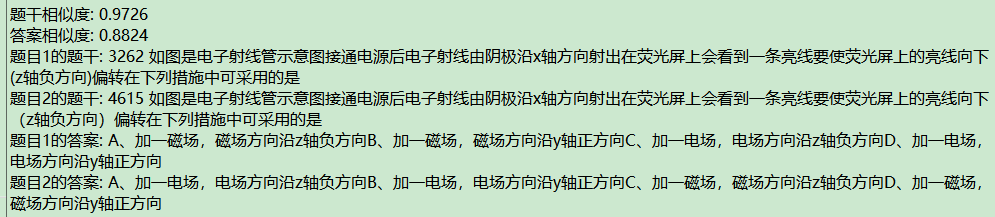


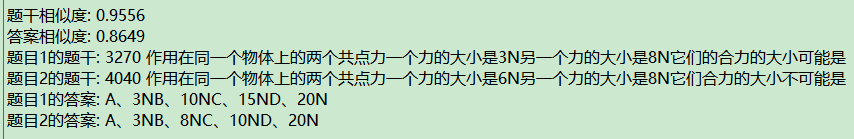
特别提醒：在系统设计时要考虑将相似度反馈给人工核查人员，方便试题管理

2.针对no choice题，由于看不见图，不做“相似”判定，只做“人工判定”和“不相似”判定，其中人工判定阈值为0.8及以上，0.8以下为不相似



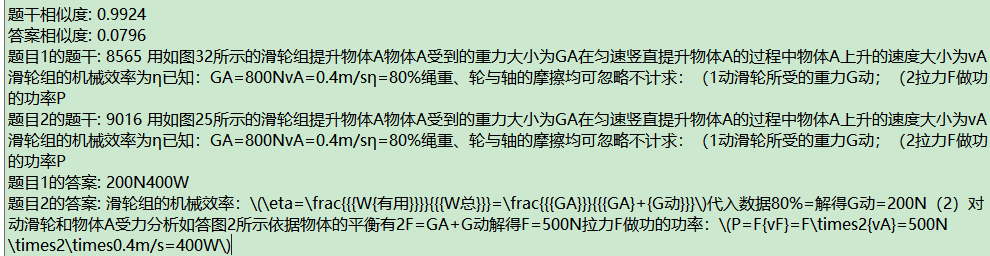


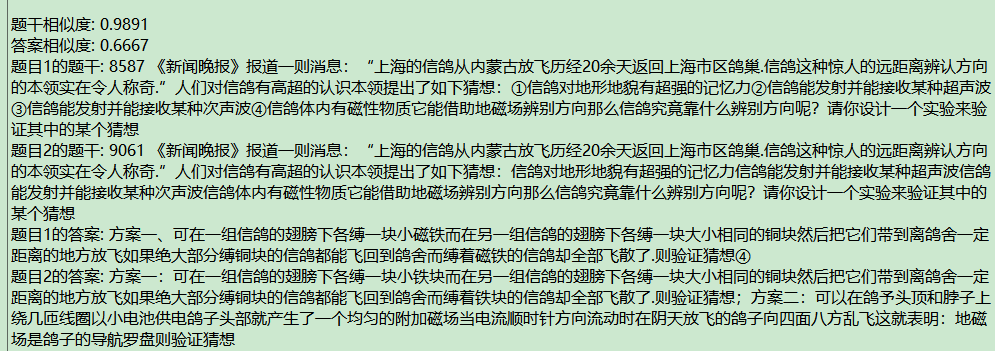


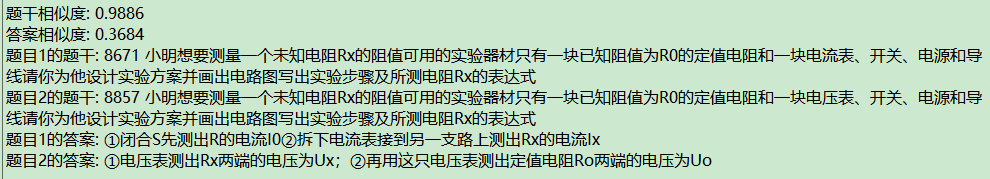


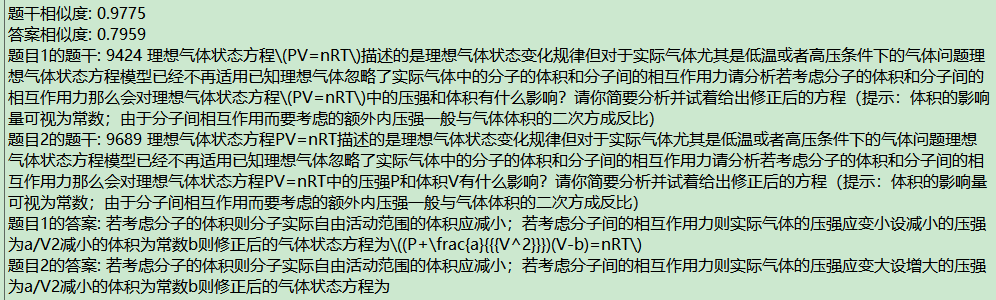
3.主观题-问答题subjective\_question，不做答案判定，答案判定会扰乱结果的准确性，只通过题干判定即可，采用题干0.8-0.95规则，主要考虑如下：

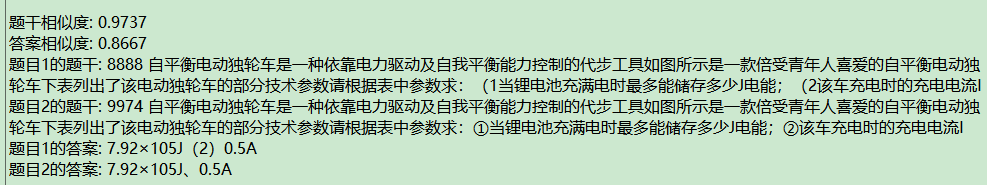
题干相似，0.95以上，但答案不相似（0.8以下）或者答案处于0.8-0.95之间，实际是相似的，参考如下：







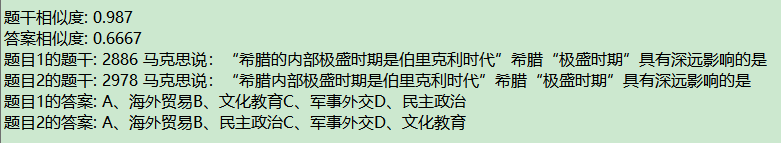


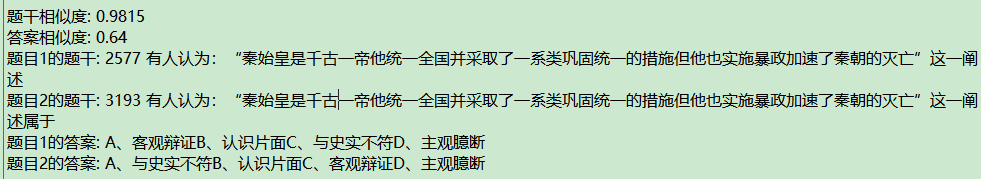


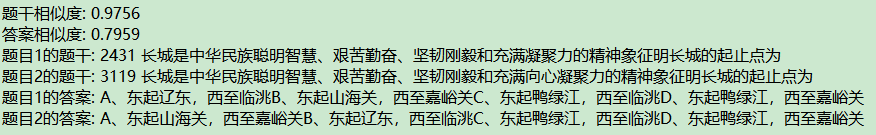
#### 二、历史：

1.建议增强选择题之间的对比，比如如下题目，其实只是调换了选项顺序，如果不要只定位A-A,B-B的操作，要增强两两对比，方能提高准确性

针对选项变换顺序的，建议特别处理







2.主观题-问答题，不做答案判定，答案判定会扰乱结果的准确性，只通过题干判定即可，采用题干0.75-0.9规则

三、总体规则方向：

1.数学、物理、化学：参考数学和物理的综合算法，0.8-0.95，考虑题干、公式、答案，参考之前阈值，整体调整主观题，主观题只判断题干，详参上述规则

2.语文、历史、道法、地理、生物：参考语文和历史的综合算法，填空、判断、单选、多选等均按照0.8-0.9，主观题判定题干0.75-0.9，详参上述规则

3.英语：考虑学科特殊性，阅读、语段、挖空等，不好确定规则，单独申请导出题库，再次验证，学科教育实验室协调

### 英语判断讨论结果-学科-20200508.docx

#### 英语查重研讨

#### 讨论结论：

整体临界值参考：0.9-0.95

1.听力题先不做查重，因为听力提问的问题基本一致，而且无法做语音判断，不太适合查重

2.针对套题，如阅读题、完形填空题建议采用三层查重，当前的题目中其实是有套题的（由于智慧学伴设计问题，只在套题的第一题显示阅读文本材料，其他题被动变成了单选题）

备注：如果是阅读材料里的选择，直接判断阅读文本是否相似即可，不用再判断题目，功能实现的前提上这样操作，“套题”先判断文本整体相似度，再借助单选的判定规则来确定，参考下表

|  |  |  |  |
| --- | --- | --- | --- |
|  | <=0.9 | 0.9<值<0.95 | >=0.95 |
| 阅读文本的整体相似度 |  |  |  |
| 题干本身相似度 |  |  |  |
| 选项/答案 |  |  |  |

有一项小于0.9，则认为不相似

有两项及以上在0.9-0.95之间，则人工判定

有一项在0.9-0.95之间，两项在0.95以上，则认为相似

有两项或以上在0.95以上，则认为相似

3.英语主观题不建议判断答案，主要判断题干即可

以下为特别案例，只作记录，暂无他用

* 如果按照规则，是相似的，实际不相似

