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。

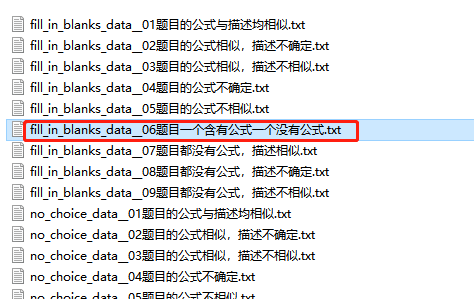
比较两个题目是否相似需要依次比较对应的各部分，其中比较的优先级是先比较公式信息，再比较题干描述的文字信息，对于选项也是先比较公式再比较选项描述。

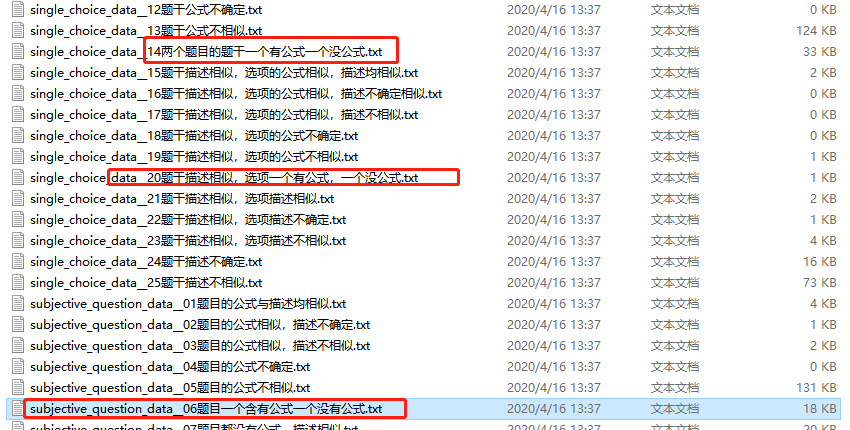
每个待比较的项目均需要给出两个阈值，其中一个为阈值上限，一个为阈值下限，我们规定相似度大于阈值上限的情况则判断为相似，小于阈值下限的则判断为不相似，在阈值上限与阈值下限之间的情况设置为不确定，需要工作人员人工校验题目的相似度。

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

**1.1.2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；