周报，2020年10月27日

屈原斌
首都师范大学
ybqu@cnu.edu.cn

1 上周计划

1. [**] 数据处理。
2. [***] 离题检测实验。
3. [***] 准备开题报告。

2 上周计划执行情况

1. [✓] 完成高中作文的处理，重新构建$DataSet_1$。
2. [✓×] 在新的数据集上进行测试，更新方案，完成部分实验。
3. [✓×] 未完成。

3 本周部分重点工作详述

3.1 数据处理

- 高中作文处理
  - 数据统计：见表1
    - 共18个题目下没有作文
- 重新构建$DataSet_1$
  - 将题目下所有一二三类卷作为不离题作文添加到测试集中，离题作文:不离题作文=
  - 数据分布：见表2

题目数	作文数	平均作文数	最大作文数	最小作文数
1274	1138251	880	24578	1

Table 1: 高中作文数据统计

	离题作文	不离题作文
作文数	341	85130

Table 2: $DataSet_1$数据分布

		方案一, strategy='topk', k=3				
		最优Threshold	A	P	R	F1
HABiLstm_C	离题	20	-	0.004	0.96	0.01
	不离题		-	0.996	0.03	0.07
	macro avg		0.04	0.5	0.5	0.04
HABiLstm_W	离题	20	-	0.02	0.55	0.04
	不离题		-	0.98	0.89	0.94
	macro avg		0.89	0.51	0.72	0.49

Figure 1: 更新数据集$DataSet_1$实验结果

		方案一, strategy='topk', k=3				
		最优Threshold	A	P	R	F1
HABiLstm_C	离题	10	-	0.48	0.52	0.5
	不离题		-	0.88	0.87	0.87
	macro avg		0.8	0.68	0.69	0.69
HABiLstm_W	离题	20	-	0.59	0.52	0.56
	不离题		-	0.89	0.91	0.9
	macro avg		0.84	0.74	0.72	0.73

Figure 2: 未更新数据集$DataSet_1$实验结果

3.2 离题检测实验

- 实验方案
  - 方案一：取topk，k=3
  - 方案二：取范文表示平均作为题目表示
  - 方案三：取全部一二类卷表示平均作为题目表示
- 实验结果：见图1、2
  - 增加离题/不离题作文比例整体指标都有下降
- 分析：
  - 默认添加一二三类卷作文为不离题作文，但实际仍然包含一部分乱写作文和离题的作文

4 存在问题

- 问题1：数据集还是有问题，一二三类卷中也存在离题作文

5 下周计划

1. [***] 补充分析实验结果。
2. [***] 尝试对向量进行归一化处理后计算相似度。
3. [***] 开题报告。
4. [**] 补充高中作文。