周报，2021年02月01日

屈原斌
首都师范大学
ybqu@cnu.edu.cn

1 上周计划

1. [***] 训练集数据处理
2. [***] 实验更新

2 上周计划执行情况

1. [✓]
2. [✓]

3 本周部分重点工作详述

3.1 训练集数据处理

- 删除作文题目/去除转写错误作文/去除句子数小于10的作文
- 对作文数过多主题下作文进行抽样
- 数据统计，见表1：
- 分类模型指标，见表2：

	训练集	验证集	测试集	合计
作文数	127507	15949	16055	159511

Table 1: 训练集数据

	Accuracy	Precision	Recall	F1
HABiLstm	0.694	0.7004	0.6826	0.6856

Table 2: 分类指标

3.2 聚类实验更新

- 聚类方法：K-means
- 数据：全部不离题作文，35个主题共59560篇作文
- 实验指标（purity）,见表3：
  - BERT的表示较差（分类指标低于lstm）

		purity_score
HABiLstm	未去除题目	0.5194
	去除题目	0.5666
BERT	未去除题目	0.4797
	去除题目	0.4993
HABiLstm	未去除题目	0.6946
	去除题目	0.7419
BERT	未去除题目	0.4663
	去除题目	0.4859

Table 3: k-means指标

4 下周计划

1. [***] 完成分类实验/seq2seq生成模型
2. [***] 完善聚类实验结果
3. [***] 更新离题实验