周报，2021年1月18日

屈原斌
首都师范大学
ybqu@cnu.edu.cn

1 上周计划

1. [***] 数据集处理
2. [***] TF-IDF实验修改
3. [***] 评分模型复现

2 上周计划执行情况

1. [✓]
2. [✓]
2. [✓×]

3 本周部分重点工作详述

3.1 数据集处理

- 对全命题作文题目进行处理
- 统计结果见表1：

Table 1: Add caption

	全命题作文	半命题作文
题目数	18	17

3.2 TF-IDF实验修改

- TF-IDF获取表示方法：
  - 对训练集和测试集作文分词，去停用词以及低频词，使用剩余词构成词表
  - 分别在当前作文上计算词表中每个词的tf值，然后在所有作文上计算每个词的idf值，得到词表大小的篇章表示
  - 根据表示进行离题判定
- 实验结果见表2：
- 实验结论：
  - TF-IDF指标较差

3.3 分段训练分类模型

- 对作文句子按照3:4:3划分成三段进行训练
- 实验结果见表3：
- 实验结论：
  - 使用段落指标低于完整篇章

		Precision	Recall	F1-score
Baseline(Doc2vec)	离题	0.2188	0.5291	0.3095
	不离题	0.9892	0.9579	0.9733
TF-IDF	离题	0.0782	0.5058	0.1355
	不离题	0.9875	0.8673	0.9235

Table 2: 实验结果

	accuracy	precision	recall	f1
HBiLstm	0.577	0.686	0.5068	0.5645

Table 3: 分段实验结果

4 下周计划

1. [***] 完成评分模型
1. [***] 尝试聚类方法