2021年04月07日进度汇报

屈原斌
首都师范大学
ybqu@cnu.edu.cn

1 今日进度

1. 英文作文数据爬取
2. 中文离题工作梳理
3. 成语古诗文检错

2 工作详述

2.1 英文作文数据爬取

- 爬取水滴英语作文，共爬取2267篇作文

2.2 中英文离题重点工作梳理

- 范文集的确定
  - 对单个题目聚类，选取聚类中心作为范文
  - 取不同范文数进行测试，观察范文数量的影响并根据结果确定范文数量
- 提高中文测试集质量
- 实验更新
  - 在英文数据上使用生成模型 ht特征做回归实验（对比baseline特征）
  - 开发集调阈值：尝试随机取最优阈值±2的值