周报，2021年07月12日

屈原斌
首都师范大学
ybqu@cnu.edu.cn

1 上周计划

1. 标题-正文匹配实验更新
2. 代码封装

2 上周计划执行情况

1. [×]
2. [✓]

3 本周部分重点工作详述

3.1 标题-正文匹配实验更新

- 数据集：
  - 构造三档数据
    - 构造方法，见图1
      1. 优 - 乐乐课堂/笔神作文
      2. 良 - 计算正文相似度，随机选择相似度较小作文，随机替换部分作文
      3. 中 - 计算正文相似度，随机选择相似度较小作文，替换正文
  - $Dataset_1$数据处理
    - 删除日记
    - 处理标题不完整和正文特殊字符
  - 数据分布，见表1

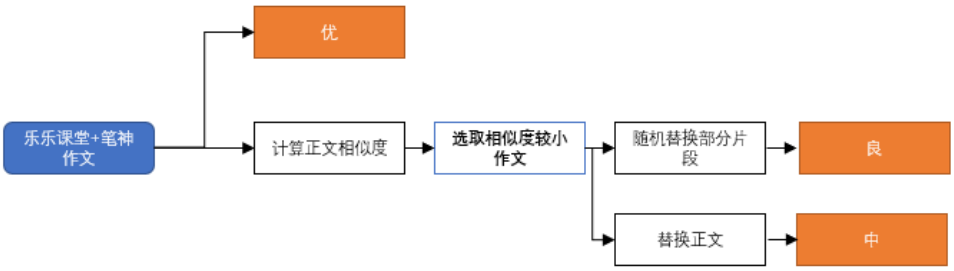

Figure 1: 数据集构造方法

- 指标更新见表2
- 结论：
  - 直接使用标题与正文计算相似度时TFIDF Marco-F1结果最优
  - AS-Reader(+SPP) 取平均结果优于取Last Time

		优	良	中	合计	数据来源
训练集		12000	3000	3000	18000	乐乐课堂，随机替换标题
验证集		500	50	50	600	
测试集	**Dataset_1**	199	127	78	407	同步作文（196原始数据+208篇构造数据，双标）
	Dataset_2	2000	200	200	2400	乐乐课堂，随机替换标题

Table 1: 数据集分布

		阈值	**Acc**	**Marco-F1**	跨二档率	相关系数	不离题**F1**	部分离题**F1**	完全离题**F1**
人**1-**人**2**		-	0.6832	0.6259	0.0173	0.7003	-	-	-
Baseline(线上系统**)**		0.80 / 0.95	0.3886	0.377	0.0743	0.2367	0.2879	0.4534	0.3896
标题与正文计算相似度	**TFIDF**	0.10 / 0.35	0.4505	**0.4382**	0.1287	0.3565	0.5552	0.32	0.4395
	Doc2vec	0.30 / 0.40	0.3936	0.3617	0.1015	0.1677	0.4928	0.3544	0.2378
	Skip-gram	0.45 / 0.60	0.4505	0.3003	0.1535	0.0942	0.6183	0.1398	0.1429
BERT-Gen HT(Topk相似度取最大值，**CLS**，**50W)**		100 / 0.60 / 0.80	0.5	**0.4374**	0.099	0.3223	0.665	0.352	0.2953
BERT-NSP			0.5347	0.4043	0.1386	0.2812	0.6831	0.2235	0.3063
AS-Reader(+SPP,取平均**, batch_size=16)**			0.5124	0.4091	0.1386	0.2277	0.6546	0.2769	0.2957
AS-Reader(+SPP,取**Last Time, batch_size=16)**			0.5223	0.347	0.1535	0.276	0.681	0.0296	0.3304

Table 2: $Dataset_1$指标更新

3.2 英文实验更新

- 数据集：10个主题，共784篇作文，离题:不离题=51:733
- 实验方案：
  - 正文与prompt计算相似度后进行排序
- 指标更新见表3
- 结论：
  - TFIDF searman结果优于Doc2vec和Skip-gram

	R@10	**R@20**	**R@50**	**P@1**	**P@5**	**P@10**	**spearman**	**ndcg**	**ndcg@10**
TFIDF	0.3227	0.5486	0.8137	0.3000	0.3000	0.2100	0.2624	0.9799	0.8378
Doc2vec	0.3341	0.4943	0.8294	0.1000	0.1400	0.1700	0.1676	0.9807	0.8759
Skip-gram	0.2424	0.4067	0.8451	0.2000	0.1200	0.1200	0.1634	0.9680	0.9038

Table 3: 英文数据集实验更新

4 下周计划

1. [***] 分析TFIDF结果，落实实验指标
2. [***] check测试集，确定离题规范
3. [***] 更新英文实验