

2021年04月24日进度汇报

屈原斌
首都师范大学
ybqu@cnu.edu.cn

1 今日进度

1. 更新离题检测实验
2. 梳理成语古诗文检错方案

2 工作详述

- 实验数据：
 - ICLE数据集，11个主题，共827篇作文，离题:切题=51:776 (1:15)
- 实验方案：
 - 聚类方案一：
 - * 计算聚类结果中可能离题的类与范文类的相似度
 - * distance_threshold从最小到最大距离进行调参，example_threshold设置为3
 - * 实验结果见表1和附件1
 - * distance_threshold、example_threshold与R@10的变化趋势见图1和附件2
 - * 结论：
 - distance_threshold值取较小时R@10可以达到最优值
 - example_threshold值与R@10成正比
 - example_threshold设置为3时使用tfidf表示指标最优
 - 聚类方案二 (one-class)：
 - * 计算全部作文与它们的质心的相似度
 - * 结论：
 - example_threshold设置为3时使用bert分类模型表示指标最优
 - SVR分数排序方案：
 - * 对SVR预测分数进行排序，五折交叉验证
 - * 实验结果见表3

		R@10	R@15	R@20	R@50
baseline		0.4793	0.4944	0.5247	0.5590
tfidf		0.3570	0.3624	0.3927	0.4323
doc2vec		0.5380	0.5380	0.5531	0.6783
分类模型	habilstm	0.4150	0.4150	0.4150	0.5455
	bert	0.4934	0.4934	0.5237	0.6200
生成模型	lstmabs	0.4083	0.4136	0.4591	0.5614
	bertabs	0.4869	0.4869	0.5172	0.6477

Table 1: 聚类方案一指标对比

Most University degrees are theoretical and do not prepare us for the real life. Do you agree or disagree?

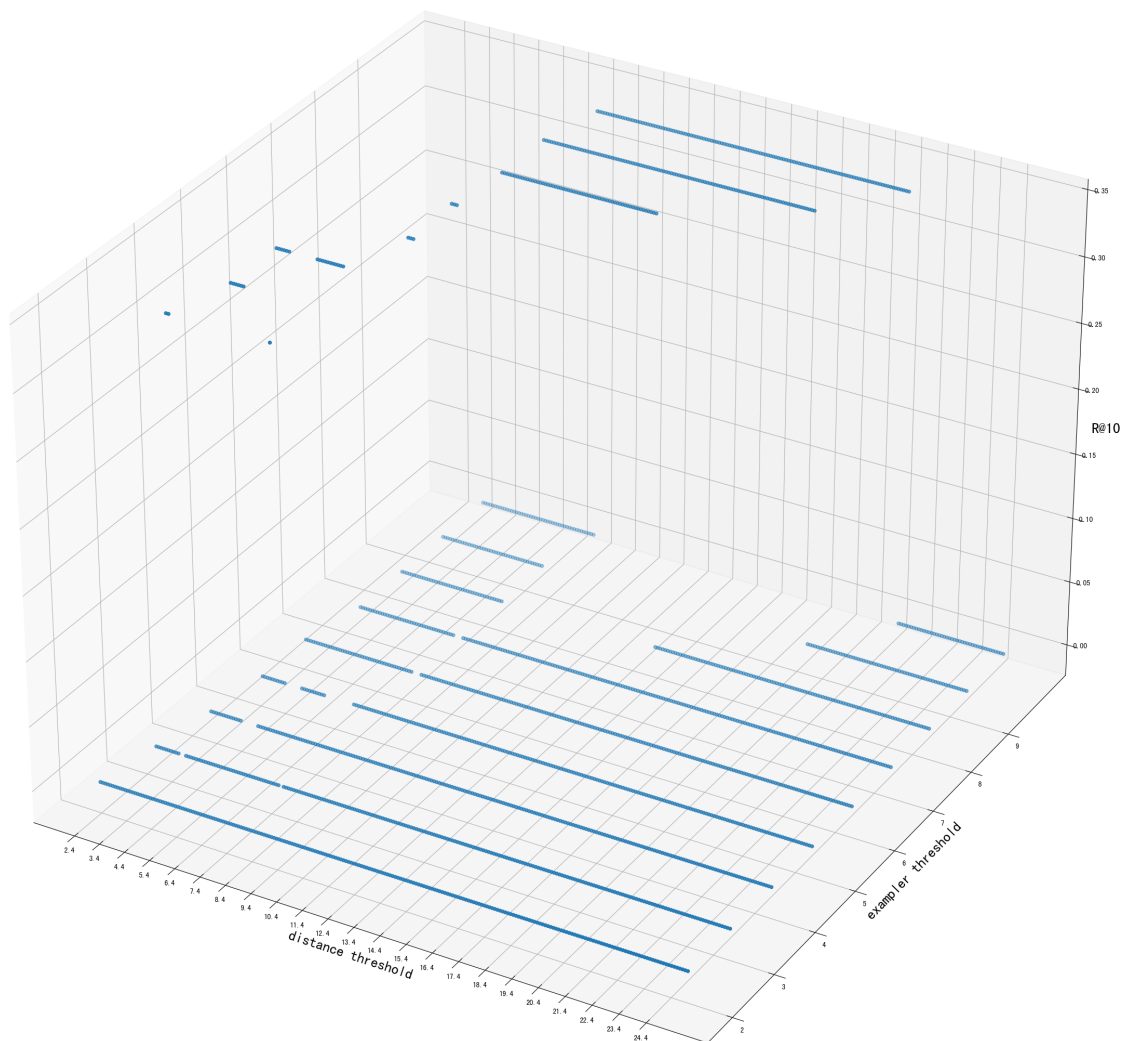


Figure 1: 变化趋势 (Topic : Most University degrees are theoretical and do not prepare us for the real life. Do you agree or disagree)

		R@10	R@15	R@20	R@50
baseline		0.4774	0.5109	0.5715	0.7148
tfidf		0.3272	0.4267	0.4472	0.6292
doc2vec		0.4471	0.4752	0.5540	0.7094
分类模型	habilstm	0.5933	0.6062	0.6062	0.7184
	bert	0.6062	0.6365	0.6971	0.8200
生成模型	lstmabs	0.3068	0.3706	0.4221	0.6328
	bertabs	0.5456	0.5608	0.5911	0.8057

Table 2: 聚类方案二 (one-class) 指标对比

	R@10	R@15	R@20	R@50
epsilon-SVR	0.3096	0.3652	0.4495	0.6571
nu-SVR	0.3109	0.3695	0.4539	0.6790

Table 3: SVR分数排序方案指标对比