周报，2021年08月09日

屈原斌
首都师范大学
ybqu@cnu.edu.cn

1 上周计划

1. 更新特征实验
2. 更新格式正确

2 上周计划执行情况

1. [×]
2. [×]

3 本周部分重点工作详述

3.1 特征实验更新

- 测试集：
  - $Dataset_1$: 395篇，不离题:部分离题:离题=197:125:73
  - $Dataset_1$: 1000篇，不离题:部分离题:离题=500:300:200
- 方案：
  - 抽取特征进行三分类
  - 特征：
    1. 标题-正文相似度：TFIDF、Doc2vec、Word2vec、IDF-Embedding、BERTGen(生成标题与原标题)
    2. 标题关键词
       - 关键词在正文中出现的比例
       - 关键词排名位置（是否出现在top5）
       - 关键词是否出现在开头/结尾
       - 包含所有关键词句子比例
    3. 正文关键词（TFIDF值排序Top5）
       - 关键词与标题相似度
    4. BertNSP特征向量
- 指标更新，见表1、表2
- 结论：
  - 使用特征Marco-F1指标最优
  - 特征方案SVC分类器结果最优
  - 特征+BertNSP特征对$Dataset_2$指标影响较大

		阈值	模型预测						
		-	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		-	0.6835	0.6213	0.0172	0.6967	-	-	-
Baseline(线上系统)		0.80 / 0.95	0.3873	0.3745	0.0785	0.2250	0.2902	0.4548	0.3784
标题-正文计算相似度	TFIDF	0.05 / 0.25	0.4734	0.4390	0.0962	0.3401	0.6196	0.3359	0.3614
	Doc2vec	0.30 / 0.40	0.3848	0.3475	0.0911	0.1729	0.5042	0.3077	0.2308
	Skip-gram	0.55 / 0.75	0.4380	0.3861	0.0785	0.2081	0.5503	0.3696	0.2385
	IDF-Embedding	0.40 / 0.55	0.4253	0.3940	0.1089	0.2289	0.5488	0.2977	0.3356
BERT-Gen HT (Topk相似度取最大值，CLS，50W)		100 / 0.60-0.80	0.5030	0.4390	0.0911	0.3374	0.6716	0.3496	0.2958
BERT-NSP			0.5392	0.4055	0.1241	0.3241	0.6979	0.1739	0.3448
AS-Reader(+SPP,取平均, batch_size=16, lr=0.002)			0.4911	0.3584	0.1392	0.1847	0.6585	0.2316	0.1852
AS-Reader(+SPP,取Last Time, batch_size=16)			0.5190	0.3423	0.1595	0.2331	0.6777	0.0305	0.3286
特征工程	SVC(31维)	C=50	0.4962	**0.4513**	0.1241	0.2669	0.6192	0.3590	0.3758
	SVC(+BertNSP特征，768维，CLS)	C=1	0.5266	0.3725	0.1367	0.2732	0.6842	0.1419	0.2913
	SVC(+BertNSP特征，128维，CLS)	C=1	0.5342	0.3847	0.1291	0.3044	0.6906	0.1529	0.3107
	SVC(+BertNSP特征，768维，Last1avg)	C=1	0.5266	0.3602	0.1418	0.2569	0.6814	0.1438	0.2553
	SVC(+BertNSP特征，128维，Last1avg)	C=1	0.5241	0.3605	0.1418	0.2487	0.6828	0.1538	0.2449
	DecisionTree	max_depth=10	0.4278	0.3948	0.1266	0.1913	0.5435	0.3231	0.3179
	RandomForest	max_depth=3	0.4557	0.4232	0.1089	0.2904	0.5918	0.3385	0.3394
	GaussianNB		0.4709	0.4172	0.1392	0.3300	0.6477	0.2376	0.3663

* +128维度表示取BertNSP特征前128维

Table 1: $Dataset_1$指标更新

		阈值	模型预测						
		-	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
Baseline(线上系统)		0.70 / 0.95	0.4440	0.4478	0.0200	0.4347	0.3215	0.4885	0.5333
标题-正文计算相似度	TFIDF	0.05 / 0.20	0.5240	**0.5275**	0.1040	0.5255	0.5065	0.4114	0.6645
	Doc2vec	0.30 / 0.40	0.4660	0.4743	0.0660	0.3946	0.4962	0.4121	0.5118
	Skip-gram	0.65 / 0.75	0.4210	0.4017	0.1760	0.2102	0.5178	0.2942	0.3932
BERT-Gen HT (Topk相似度取最大值，CLS，50W)		100 / 0.60 / 0.80	0.6310	0.5674	0.0290	0.6484	0.7804	0.3004	0.6214
BERT-NSP			0.8210	0.8120	0.0380	0.7683	0.8552	0.7403	0.8406
AS-Reader(+SPP,取平均)			0.6770	0.6339	0.0630	0.5990	0.7743	0.4694	0.6581
AS-Reader(+SPP,取Last Time)			0.6520	0.6224	0.0700	0.5650	0.7398	0.4654	0.6619
特征工程	SVC(31维)	C=50	0.6120	0.6028	0.0620	0.5853	0.6605	0.3918	0.7561
	SVC(+BertNSP特征，768维，CLS)	C=1	0.8390	**0.8349**	0.0290	0.8061	0.8390	0.7500	0.8921
	SVC(+BertNSP特征，128维，CLS)	C=1	0.8380	0.8329	0.0270	0.8104	0.8639	0.7429	0.8921
	SVC(+BertNSP特征，768维，Last1avg)	C=1	0.8280	0.8224	0.0230	0.8113	0.8588	0.7205	0.8878
	SVC(+BertNSP特征，128维，Last1avg)	C=1	0.8310	0.8257	0.0240	0.8111	0.8604	0.7290	0.8878
	DecisionTree	max_depth=10	0.5900	0.5960	0.0700	0.5497	0.6114	0.4463	0.7302
	RandomForest	max_depth=3	0.6080	0.6066	0.0650	0.5928	0.6437	0.4410	0.7350
	GaussianNB		0.6060	0.5697	0.0910	0.5564	0.6868	0.3399	0.6825

* +128维度表示取BertNSP特征前128维

Table 2: $Dataset_2$指标更新

4 下周计划

1. [***] 验证特征方案实验
2. [***] 更新格式正确实验