**基于知识图谱的优质文章识别**

**背景**

在移动互联网、大数据的时代背景下，各类自媒体文章呈爆发式增长，在搜索和推荐等信息分发场景下，甄别出优质文章并分发给用户具有重要的研究意义和实际应用价值。除了文章本身的写作质量以外，优质文章应该具备内容的深度和新颖性等，因此单纯依赖文章本身内容无法完整识别优质文章。本任务将引入文章相关的外部知识图谱，结合文章内在的知识逻辑，在对文章更深入的语义理解基础上实现优质文章识别。

# 任务描述

本任务要求参赛者利用文章的知识图谱进行建模，实现优质文章分类。

* 输入：文章标题、文章正文、文章的知识图谱
* 输出：是否是优质

# 数据集介绍

本任务数据集来自国内主流自媒体平台百度百家号，数据集中共有6000篇文章，每篇文章均有标题、正文、是否为优质文章的标签和相应知识图谱信息，其中5000篇用于训练集，500篇用于验证集，500篇用于测试集。优质文章的评判维度主要有写作、可信度、深度、新热性、正能量，具体细则参见附录。

数据格式如下：

{

"id": 1,

"title": "标题",

"date": "发布时间",

"content": "正文，段落之间有</br>作为分隔标记",

"label": 1, // 1表示为优质文章0表示为非优质文章

"entities": {

"实体1": "实体1的百度百科链接",

"实体2": "实体2的百度百科链接",

}

}

输出格式如下：

{

"id": 1,

"label": 1, // 1表示为优质文章0表示为非优质文章

}

# 评价指标

本任务采用精准率（Precision）、召回率（Recall）以及F1-Score作为评价指标。

# 任务提交方式

本次任务将采取排行榜的方式，验证集发布后，允许参赛队伍每天至多向平台提交三次结果，文件命名为“result.txt”，格式与数据集描述中的输出格式相同，排名每十分钟更新一次。参赛队伍可在测试集发布之前随时上传验证集的计算结果，管理系统会及时更新各队伍的最新排名情况，验证集榜不计入比赛最终成绩。测试集为最终评估比赛成绩的有效评测集，待测试集发布后，允许参赛队伍每天至多提交一次，总共至多提交五次评测结果。

**所有获奖团队需额外提交相关材料以供资格审查，要求如下**：

1. 测试集结果文件，用result.txt命名
2. 相关代码及说明
3. 方法描述文档

将以上三个文件在任务提交截止日期前发送至邮箱ccks2022\_casd@163.com。邮件的标题为：“ccks2022-参赛队名称”，例如“ccks2022-小虎队”。代码及其文档需打包成一个文件（tar，zip，gzip，rar等均可），作为邮件附件传送，要求提交所有的程序代码及相关的配置说明，程序可以运行且所得结果与result.txt相符。如果方法使用了额外资源，要求说明并提供资源文件或地址。

# 奖励设置

第一名：人民币10000元

第二名：人民币3000元

第三名：人民币2000元

创新奖：人民币5000元

# 时间安排

评测任务发布：4月6日

报名时间：4月6日—7月25日

训练及验证数据发布：暂定4月25日（以最终发布时间为准）

测试数据发布：7月25日

提交测试结果：7月31日

# 任务组织者

曹自强（苏州大学）

王德瑞（百度内容策略部）

徐扬（百度内容策略部）

谢文睿（百度内容策略部）

# 任务联系人

QQ群：471919965

曹自强：zqcao@suda.edu.cn

王德瑞：wangderui@baidu.com

徐扬：xuyang24@baidu.com

谢文睿：xiewenrui01@baidu.com

# 附录

优质文章的标注标准如下：

|  |  |  |  |
| --- | --- | --- | --- |
| 写作 | 切题 | 题目点明主题，文章内容紧扣题目，不能是标题党 | 评分：优、中、差 |
| 文字质量 | 无错别字、病句  逻辑通顺，阅读流畅 | 评分：优、中、差 |
| 直观易懂 | 用大量准确的数字或数据描绘所写的事物，文章通过数据/数字客观的量化事物全貌，使用户更直接的了解事物，文章内容更具准确性  或者  用简洁明确的语言，为所写事物的特征进行概括或下定义，使用户更直接的接收到该事物的特征，观点表达更言简意赅 | 评分：优、中、差 |
| 可信度、深度 | 多方、多角度 | 满足以下任一特征:   1. （议论文适用）引用充分的外部引证(事例、证据、论据等)，对中心思想或个人观点进行多方论证或解析，增强文章说服力，使文章经得起推敲，用户更信服 2. （议论文适用）以严谨的思维逻辑，发现事物的多面性(不借助外部引证和历史背景)，对事物和观点进行多维度阐述，能让用户从不同维度理解内容，并对每个维度进行深层挖掘和表达，使主题或观点清晰深刻 3. （说明文适用）文章专业性强，术语丰富，能学到知识 4. （记叙文适用）文章从时间顺序、空间顺序、逻辑顺序等多角度描述事物 | 评分：优、中、差（不适用文章打分固定为中） |
| 背景信息丰富 | 主题背景信息脉络清楚，由点及线，如时间线、事件发展线等  拒绝平铺直叙的背景或事件罗列 | 评分：优、中、差 |
| 新热性 | 时效/创意 | 满足以下任一特征:  1)内容主题涉及热门事件或事件的延伸进展及热门话题讨论(包含但不限于娱乐、影视类等)  2)近期国内外或知名人物的重大事件，如国内重大时政新闻、国际冲突、国际/两岸关系近期变化、特朗普访华等  3) 作者对常见事物有别出心裁的分析视角，观点新颖少见，表达的内容使用户耳目一新，扩展认知 | 评分：优、中、差 |
| 正能量 | 积极乐观 | 满足以下任一特征:   1. 作者观点激励进取，催人奋进，给予用户内心力量，引人深思 2. 文章氛围温馨，以感化人性，温暖人心的人或事，引发情感共鸣 3. 使用风趣幽默的语言表达，带动用户情绪，内容表达轻松，使用户心情愉悦 | 评分：积极、中性、消极 |

依据以上维度综合判定是否为优质文章的具体规则在数据集发布时一并给出。