**CCKS 2020 技术评测任务书**

# 基于标题的大规模商品实体检索

# 任务描述

本评测任务为基于标题的大规模商品实体检索。即对于给定的一个商品标题，参赛系统需要匹配到该标题在给定商品库中的对应商品实体。 基于标题的大规模商品实体检索存在的挑战如下：（1）输入文本中可能无法识别出实体指代词；（2）商品标题中存在很多变异指代，没有给定的指代映射表；（3）商品标题一般较短，上下文语境不丰富，须对上下文语境进行精准理解。

**输入：**

 输入文件包括若干行商品标题。

**输出：**

 输出文本每一行包括此标题对应的商品实体，即给定知识库中商品ID，只返回最相关的1个结果。

**示例：**

 输入：

 [转卖]现货初中学生巨文库 武志红2016年新作 浙江出版社传

 俄罗斯代购印度睫毛增长液3ml,断货王！如果有需要请咨询客服

 泰国正品DC套装yh产品定居曼谷提供视频支持直发

…………

 输出：

 1

 231

 467

…………

说明：存在输入的商品标题不指向任何商品实体的可能，若判断商品标题不指向任何实体，输出N/A即可。

# 数据描述

## 2.1. 商品知识库

该任务的商品知识库包括来自淘宝网站的约40万个商品实体。知识库中的每个实体都包含一个subject-ID，一个字符串名称，及与此实体相关的一系列三元组<subject，predicate，object>信息形式。知识库中每行代表知识库的一条记录，每条记录的格式为一个json格式。predicate-id和object-id的值都为0，subject-id的值为一个正整数，说明subject总是对应知识库中的一个实体。

例如，

{

 "subject\_id": "1",

 "subject": "巨婴国",

 "type": ["Publication"],

 "data": [

 {

 "predicate": "出版社",

 "object": "浙江人民出版社"

 },

 {

 "predicate": "书名",

 "object": "巨婴国"

 },

 {

 "predicate": "作者",

 "object": "武志红"

 }，

 {

 "predicate": "出版时间",

 "object": "2016年12月"

 }，

 {

 "predicate": "装帧",

 "object": "平装"

 }，

 ]

}

## 2.2 标注数据集

标注数据集由训练集、开发集和评估集组成。其中训练集中包括9万条短文本标注数据，开发集和评估集分别包括1万条短文本数据，3份数据均通过众包标注生成。

标注数据集中每条数据的格式为：

{

 "text\_id": "1",

 "text": "[转卖]现货初中学生巨文库 武志红2016年新作 浙江出版社传",

 "implicit\_entity": [

 {

 "subject": "巨婴国",

 "subject\_id": "1"

 }

 ]

}

标注数据集主要来自于：真实的电商网站数据，是商家发布的商品信息，商品标题平均长度为23中文字符。

标注文本对象的示例数据如下：

|  |
| --- |
| **标注文本对象** |
| 龙飞凤舞彩色双头龙双头仿真阳具同志拉拉零号超爱 |
| 俄罗斯代购印度睫毛增长液3ml,断货王！如果有需要请咨询客服 |
| 泰国正品DC套装yh产品定居曼谷提供视频支持直发 |
| 原装舒服太原厂质量保证 |
| 巨ying国 武志红big baby |

# 评价指标

对于给定的商品标题，参赛系统输出的结果为商品知识库中的商品ID。我们通过将输出结果与人工标注集合S进行比较得到准确率(Precision)作为评价指标：

给定M个商品标题输入$T=\{t\_{1},t\_{2},t\_{3},…,t\_{M}\}$，系统输出结果中与标注结果相同的数量为N，则准确率为：

$$Precision=\frac{N}{M}$$

# 任务提交

本次任务将采取刷榜的方式，验证集发布后，允许参赛队伍多次向平台提交结果，参赛队伍可在评测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；

测试集发布后，允许参赛队伍多次提交测试集结果文件。

**最终提交文件要求：**每一个参赛队需提交的材料如下：

1. 测试集结果文件，用result.txt命名（UTF-8格式），格式与任务描述中的示例输出相同；
2. 相关代码及说明
3. 方法描述文档（非评测论文，评测论文撰写要求见CCKS 2020官网）

以上三个文件需在任务提交截止日期前发送至邮箱ccks2020\_er@163.com。邮件的标题为：“CCKS-ER-参赛队名称”，例如“CCKS-ER-XX队”。

代码及其文档需打包成一个文件（tar，zip，gzip，rar等均可），用code.xxx命名，要求提交所有的程序代码及相关的配置说明，**程序应当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。

# 时间安排

• 评测任务发布：3月20日

• 报名时间：3月20日—7月10日

• 训练及验证数据发布：3月20日

• 测试数据发布：7月10日

• 提交测试结果：7月20日

• 评测论文提交：8月5日

• CCKS会议日期(评测报告及颁奖)：8月15日—18日

# 参赛奖励

CCKS专委会为每个任务前三名设立奖金并颁发证书。

第一名：5000元

第二名：3000元

第三名：2000元

同时组委会将针对不同任务特别评选“创新技术奖”（每一任务至多一项，奖金：5000元），专门用于鼓励创新性技术的使用。

# 组织者

黄龙涛： kaiyang.hlt@alibaba-inc.com

袁博： qiufu.yb@alibaba-inc.com