贝贝人工智能挑战赛：商品类目预测

赛题描述

商品类目预测为电商平台商家上架商品时选择合适的类目提供了重要的参考信息，可以减少商品关联上错误类目的概率，大幅降低运营人员审核商品的成本。由于平台初期缺少商品类目预测模块，加上审核人员的疏忽，导致已经上架的商品中有不少商品关联在错误的类目下，比如“绿植”关联在“玩具”的类目下。类目错挂问题给电商平台的商品检索相关性、推荐系统的精准度等方面都带来了不小的挑战。

在这样的背景下，如何通过算法的能力快速找到错挂类目的商品，并为这些商品输出建议的正确类目成为我们研究的重要课题之一。商品类目预测模块需要根据商品的标题和描述信息输出商品的正确类目（共三级类目）。本次比赛旨在发现实现商品类目预测模块的高效算法。

组队规则

晋级队伍复赛结果提交截止时间为11月1日22:00:00，前十名会在11月2日收到答辩通知，届时提交结果及报告。

赛程安排

初赛 2018年10月8日-2018年10月23日

1) 平台会在10月8日晚开幕式结束后开放训练数据集A、验证数据集A、测试集A，参赛选手可以自行下载数据，在本地进行算法设计、模型训练及评估。

2) 参赛选手在本地使用先前训练的模型进行预测，生成预测结果并提交至平台。结果提交后，系统会按照评测指标实时反馈分数，并更新榜单排名。

3) 每队每天最多可提交3次。榜单以所有参赛队伍的历史最优成绩进行排名。当有团队提交新的预测结果之后，榜单将实时更新客观指标。

复赛 2018年10月25日-2018年11月1日

1) 平台会在10月23日冻结初赛榜单，同时开放训练数据集B、验证数据集B、测试集B，参赛选手可以自行下载数据，在本地进行算法设计、模型训练及评估。复赛另开榜单，初赛成绩不带入复赛阶段。下载链接:https://pan.baidu.com/s/1BdhibDbMonWdlqJZcKb3og 密码:g131

2) 参赛选手在本地使用先前训练的模型进行预测，生成预测结果并提交至平台。结果提交后，系统会按照评测指标实时反馈分数，并更新复赛榜单排名。

3) 每队每天最多可提交2次。榜单以所有参赛队伍的历史最优成绩进行排名。当有团队提交新的预测结果之后，榜单将实时更新客观指标。

答辩 2018年11月4日

1) 平台中最后一次复赛榜单成绩排名前十的队伍将受邀来进行现场答辩，受邀答辩的队伍将在11月2日收到答辩通知。

2) 参赛队伍应提前准备答辩材料，包括且不限于PPT简报文档、算法代码等。

3) 模型决赛的预测结果得分和答辩成绩的加权总成绩将决出大赛最终的大奖。

注：大赛主办方可以根据实际时间安排，对赛程进行更新。

数据说明

本次比赛我们分别为初赛和复赛各提供了一份高质量的海量数据集，每份数据集都包括训练集、评估集、测试集三个部分。训练、评估和测试数据集的格式如下，所有字段都经过编码脱敏处理，字段之间用Tab健（'\t'）分隔。

字段名	字段说明	是否为预测目标
item_id	商品ID	否
title_characters	商品标题的字符序列	否
title_words	商品标题的分词序列	否
description_characters	商品描述的字符序列	否
description_words	商品描述的分词序列	否
cate1_id	商品一级类目ID	是（测试集中该字段内容缺失）
cate2_id	商品二级类目ID	是（测试集中该字段内容缺失）
cate3_id	商品三级类目ID	是（测试集中该字段内容缺失）

数据示例如下：

原始数据：

> 商品ID：0 商品标题：欧锐仿真触摸屏儿童玩具手机 商品描述：宝宝能够通过滑屏智能手机的数字、字母、符号、颜色提高宝宝的辨识能力；滑屏智能手机具有：9首故 事、9首儿歌、9首唐诗，让宝宝在娱乐的轻松氛围中学到更多的知识;宝宝可以在手机上弹钢琴，仿真的琴声，还有好几首快乐的舞曲。 商品一级类目ID：3 商品二级类目ID：23 商品三级类目ID：89

选手看到的数据：

item_id	0
title_characters	c532,c2439,c913,c306,c849,c1764,c1372,c15,c7,c266,c207,c97,c192
title_words	w41226,w1287,w10038,w11,w127,w792
description_characters	c4,c4,c151,c1211,c164,c229,c197,c1372,c503,c151,c97,c192,c3,c485,c267,c61,c267,c461,c61,c1641,c483,c61,c344,c36,c273,c70,c4,c4,c3,c2549,c1186,c151,c160,c237,c197,c1372,c503,c151,c97,c192,c207,c52,c225,c
description_words	w6,w1814,w1579,w2515,w10714,w764,w792,w2,w1527,w13,w285,w13,w15677,w13,w145,w1048,w4386,w129,w2,w23753,w516,w59,w2515,w10714,w764,w792,w649,w38,w430,w1707,w9023,w6764,w13,w430,w
cate1_id	3
cate2_id	23
cate3_id	89

备注：

- 初赛阶段使用数据集A包含10个一级类目、64个二级类目、125个三级类目。
- 复赛阶段使用数据集B包含20个一级类目、135个二级类目、265个三级类目。

结果提交格式说明：

字段名	字段说明
item_id	商品ID
cate1_id	商品一级类目
cate2_id	商品二级类目
cate3_id	商品三级类目

只需要提交测试集中的商品ID和模型预测的一级、二级、三级类目ID即可。

上传格式要求：大赛官网上传文本文件xxx.txt，字段（列）之间使用分隔符（'\t'）分隔，换行使用（'\n'）分隔

评价标准

我们将以三个类目层次下F1值的加权平均值作为本次比赛结果的评价指标，具体的计算方式如下：
$rank\_score = 0.1 \times F_{1\_cate1} + 0.3 \times F_{1\_cate2} + 0.6 \times F_{1\_cate3}$

其中， $F_{1\_cate1} = (\sum_{c \in C1} 2 \times \frac{precision_{(c)} \times recall_{(c)}}{precision_{(c)} + recall_{(c)}}) / |C1|$

$F_{1\_cate2} = (\sum_{c \in C2} 2 \times \frac{precision_{(c)} \times recall_{(c)}}{precision_{(c)} + recall_{(c)}}) / |C2|$

$F_{1\_cate3} = (\sum_{c \in C3} 2 \times \frac{precision_{(c)} \times recall_{(c)}}{precision_{(c)} + recall_{(c)}}) / |C3|$

$precision_{(c)} = \frac{TP_{(c)}}{TP_{(c)} + FP_{(c)}}$

$recall_{(c)} = \frac{TP_{(c)}}{TP_{(c)} + FN_{(c)}}$

□	预测值(Predictive label): 正	预测值(Predictive label): 负
实际值(Actual label): 正	TP	FN
实际值(Actual label): 负	FP	TN

即将n分类的评价拆成n个二分类的评价，根据每个二分类评价的TP、FP 、FN 计算出准确率和召回率，再由准确率和召回率计算得到F1值。

最终提交说明

提交内容：最高分数测试样例文件、代码及比赛报告，格式如下：

- seedCup2018 初/复赛-xxx队.zip

    - answer.txt（要求测试最优的结果）
    - src（源文件目录）
    - xxx.pdf
- 比赛报告内容
    - 使用语言以及运行环境
    - 提供代码相应的接口并指明运行需要用到的变量含义，以便裁判组进行测试
    - 数据特征提取思路
    - 预测模型选取（包括对于规则的描述和最终模型的选择）
    - 对于模型参数的选择与优化思路
    - 报告内容不限于以上所述内容

初赛提交截止时间：2018年10月23日00:00，报告及源码提交截止10月23日22:00，未提交最终内容的队伍视为放弃比赛。

复赛提交截止时间：2018年11月1日22:00，未提交最终内容的队伍视为放弃比赛。

作品提交到大赛公邮 seedcup@dian.org.cn

注意事项

1. 本次比赛不允许使用外部数据，但可以使用开源的已有算法和工具
2. 如果发现参赛队伍有造假、作弊、雷同等行为，将取消该队伍的参赛资格及奖励
3. 比赛过程中，大赛评审组可能会根据比赛情况，对比赛内容和评分标准进行调整
4. 比赛最终解释权归大赛评审组所有