# 大数据精准营销中搜狗用户画像挖掘

2016-ccf-data-mining-competition，（用户画像的英语原文应该是 user profile，所以也可以翻译为用户资料、用户概况）

# 竞赛简介

在现代广告投放系统中，多层级成体系的用户画像构建算法是实现精准广告投放的基础技术之一。其中，基于人口属性的广告定向技术（其实就是个性化推荐）是普遍适用于品牌展示广告和精准竞价广告的关键性技术。人口属性包括自然人的性别、年龄、学历等基本属性。

在搜索竞价广告系统中，用户通过在搜索引擎输入具体的查询词来获取相关信息。因此，用户的历史查询词与用户的基本属性及潜在需求有密切的关系。

举例如下：

1、 年龄在19岁至23岁区间的自然人会有较多的搜索行为与大学生活、社交等主题有关

2、 男性相比女性会在军事、汽车等主题有更多的搜索行为

3、 高学历人群会更加倾向于获取社会、经济等主题的信息

本题目提供用户历史一个月的查询词与用户的人口属性标签（包括性别、年龄、学历）做为训练数据，要求参赛人员通过机器学习、数据挖掘技术构建分类算法来对新增用户的人口属性进行判定。

# 数据描述

## 1、数据集

Train.csv 带标注的训练集

Test.csv 测试集

## 2、数据介绍

本数据来源于搜狗搜索数据，ID经过加密，训练集中人口属性数据存在部分未知的情况（该情况为竞赛题目特定设置，需要参赛人员的解决方案能够考虑数据缺失对算法性能的影响）。

## 3、数据所有字段解释

ID 加密后的ID

age 0：未知年龄; 1：0-18岁; 2：19-23岁; 3：24-30岁; 4：31-40岁; 5：41-50岁; 6： 51-999岁

Gender 0：未知1：男性2：女性

Education 0：未知学历; 1：博士; 2：硕士; 3：大学生; 4：高中; 5：初中; 6：小学

Query List 搜索词列表

## 数据示例

train.csv中的一个样本：

00627779E16E7C09B975B2CE13C088CB 4 2 0 钢琴曲欣赏100首 一个月的宝宝眼睫毛那么是黄色 宝宝右眼有眼屎 小儿抽搐怎么办 剖腹产后刀口上有线头 属羊和属鸡的配吗

# 任务描述

本题目提供用户历史一个月的查询词与用户的人口属性标签（包括性别、年龄、学历）做为训练数据，要求参赛人员通过机器学习、数据挖掘技术构建分类算法来对新增用户的人口属性进行判定。即**对test.csv文件中的每条记录进行年龄、性别、学历的判断**。**这样看来，其实和情感分析是一样的，我们要根据段文本，给出3个分类结果。可以用原项目中的方法，最后用LSTM实现，看看性能上是否有所突破。**