# 想法1

（这种方法超级麻烦，不具有普遍适用性，我连考虑都不会考虑）

分成147（3\*7\*7）类， 后来觉得这样效果不好，后来看了看竞赛要求的也是分别预测，分别评分，而不是一次就把3类的标签都给出

所有后来我们改进了当时的想法，决定对年龄，性别，学历进行分别预测

# 想法2

我们先对所有的单词进行分类，分成比如体育，经济，教育等等，一些大类别，然后看看每个用户搜索的关键词属于哪一类。作为特征

后来，因为无法确定分为多少类，，所以否定了这个想法。

# 想法3

Step1:进行文本分词处理，提取出搜索词中的关键词；

Step2:建立向量空间模型 （1）权值计算（布尔权值，词频权值,TF/IDF，TFC，ITC等方法）

(2) 向量相似度量（内积，绝对值距离，切比雪夫距离等） ；

Step3:对文本进行分类（概率分类器，决策树分类器，神经网络分类器等）

Step4:进行数据测试，根据给定的查询词，首先也对它先分词，提取关键词，然后和一个大类进行相关性测量，

VSM中的（两个向量的夹角越小说明关联度越大），然后决定出它是属于哪一个类，

之后再和这个类别下的关键词进行相关性对比，然后逐个决定出人物属性（年龄，性别，学历）

确定了思路之后，我用skleran 这个机器学习包进行了实现。

## 初赛思路v1

一 Preprocessing

1分词 采用结巴分词

2数据清洗，删掉缺失数据，例如数据中的未知（标签为0）

二 Feature extraction

 1 countvector：计算词频

2 tfidfvector: 计算tfidf作为权重值

3 hashvector: 利用hash 算法将单词映射到向量空间

三 Feature selection 

选择特征的数量，也就是数据矩阵的维度。v1没有使用算法，直接指定特征的维度。

四 Feature union

初赛中没有使用特征融合

五 Model selection

KNN

SVM 效果最好

贝叶斯

六 Model Evaluation and Optimization

利用准确率衡量分类的结果

## 初赛思路v2

在特征提取的时候，我们担心数据过拟合，所以进行特征选择

我们的特征选择的方法主要是卡方跟LDA主题模型

但是经过测试，LDA+TFIDF的效果不如单纯的tfidf

TFIDF+卡方的效果稍好一点点

特征提取我们也试过n-gram效果也不是很好

## 最终初赛思路（v2)

1.jieba分词

2.特征提取：tfidf

3.特征选择：卡方

4.分类：SVM