机器学习的术语：

样本集：训练数据（训练集）+测试数据（测试集）

样本的特征：样本的属性（除结果外每一列）

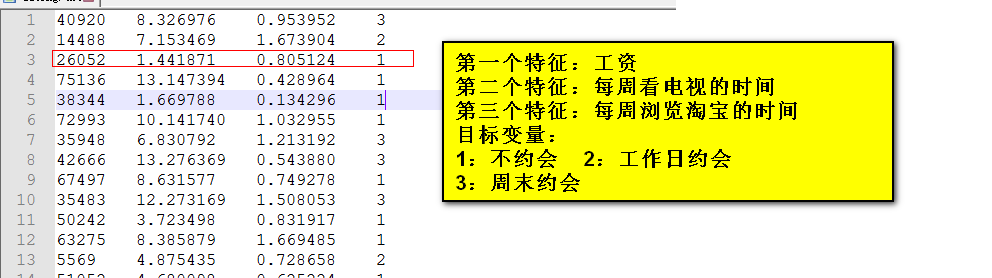
目标变量：结果

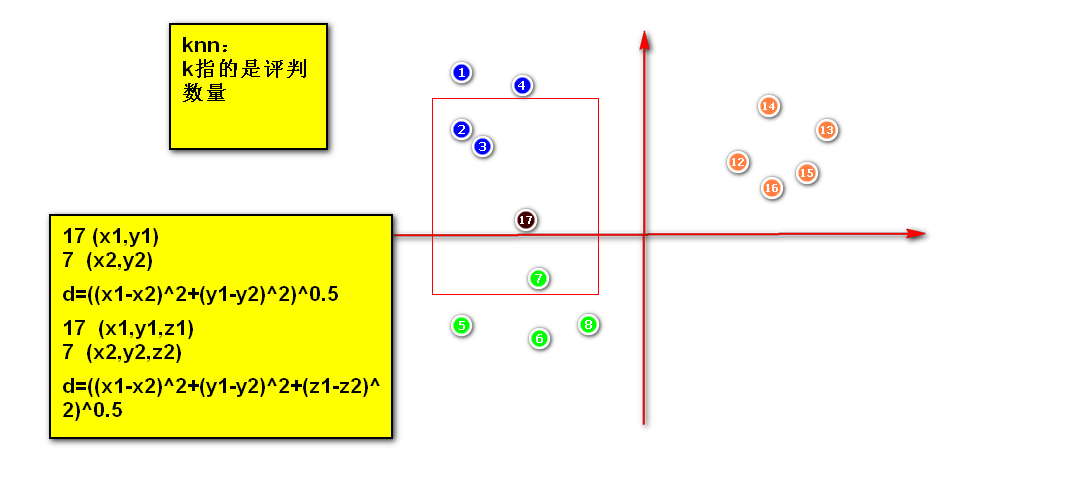
机器学习的目的是为了找样本的特征与目标变量之间的关系

KNN算法

核心思想：物理类聚 人以群分 谁离我近就是谁（by hbw）

栗子：





使用knn算法解决赵老师找对象的问题

1. 加载数据集 分开特征矩阵和目标向量
2. 处理特征矩阵，将所有元素归一化到0-1之间 (除以每一列最大值)
3. 给定一个新样本 计算新样本到每个样本之间的距离 保存到距离数组中
4. 将距离排序，取出前k个距离最小(近)的样本
5. 统计取出样本的每个类别出现的次数
6. 把出现次数最多的那个类别作为新数据的类别

代码实现一定要按照思路、

医院排行榜

<http://www.fudanmed.com/institute/news2017-3.aspx>

球探网

<http://live.win007.com/>

中国东方航空网

<http://www.ceair.com/booking/pek-sha-190301_CNY.html>

智能聊天机器人

需求：

1. 语音交流

语音识别 自然语言处理 语音合成

1. 支持表情包（带文字信息）

文字提取 自然语言处理 语音合成

Restful风格

<http://www.baidu.com?w=1&q=2>

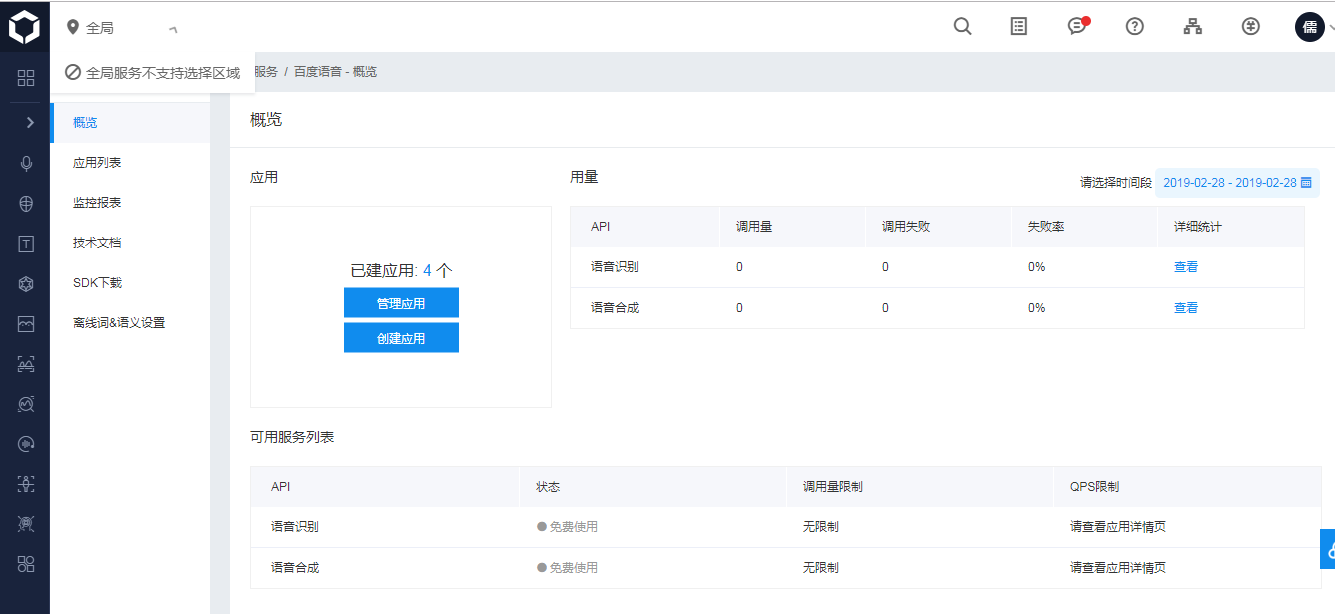
<http://www.baidu.com/1/2> django实现restful

百度AI接口

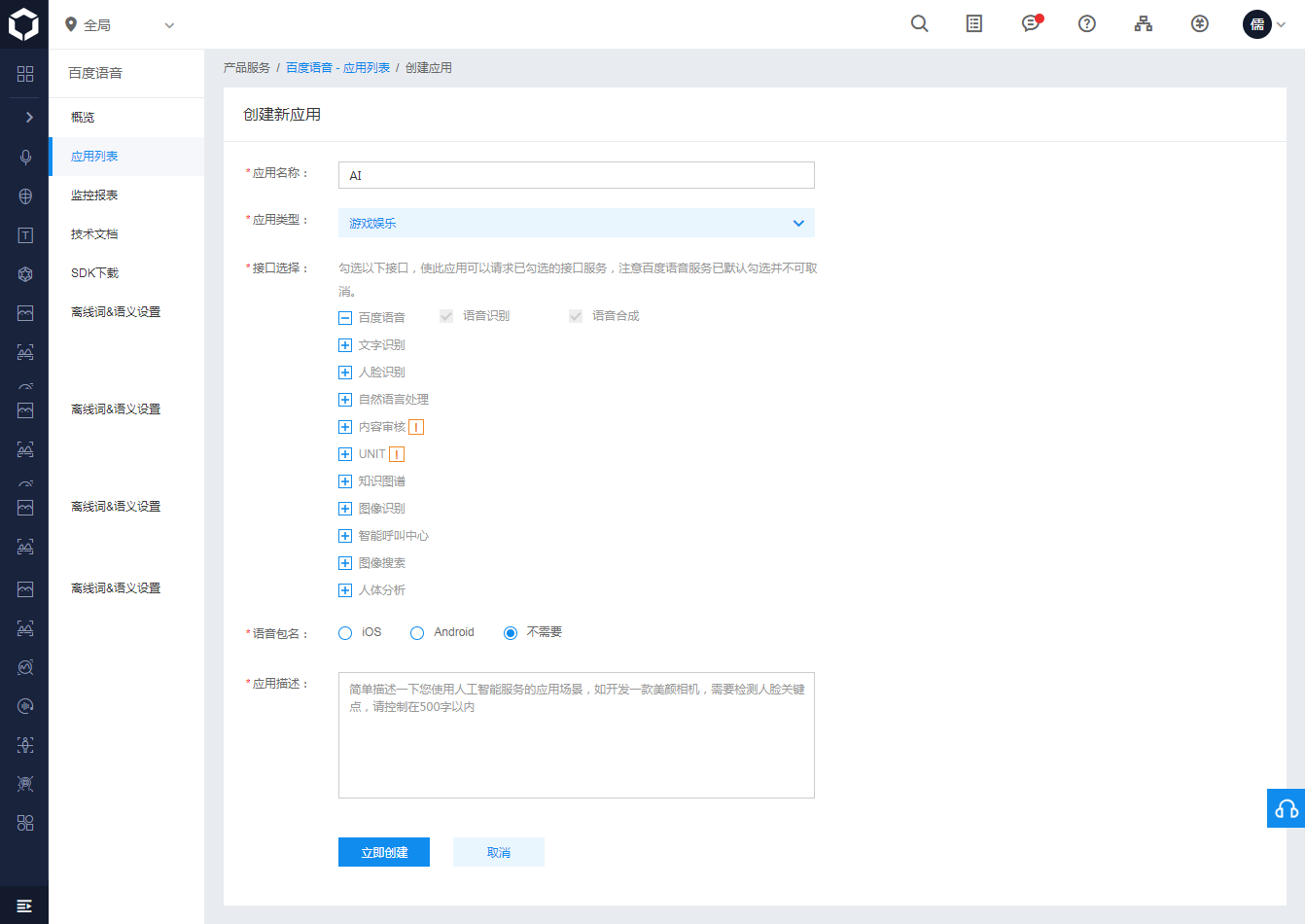
<http://ai.baidu.com/?track=cp:ainsem|pf:pc|pp:tongyong-pinpai|pu:pinpai-baiduAIkaifangpingtai|ci:|kw:10003817>

在主页点击语音技术，跳转登录

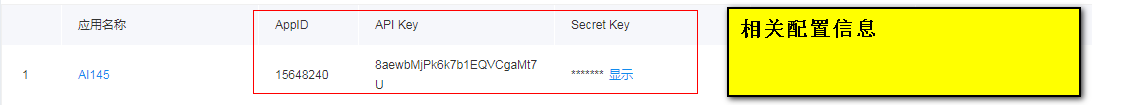
登录以后



点击创建应用



输入相关信息以后，点击立即创建，点击返回应用列表

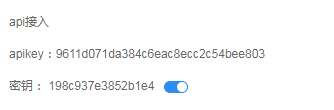


查看技术文档 安装sdk

pip install baidu-aip

语音识别代码看程序

图灵机器人创建完毕以后



用途：

1. 智能助手
2. 智能点歌台 智能点播视频
3. 其他行业的人机交互
4. 拍笔记
5. 违规车辆识别
6. 图像过滤
7. 人体行为检测
8. 识别验证码

简历的技术描述：

精通 熟读源码

熟练掌握 用的熟练

熟悉 使用次数不多

了解 听说过 大概知道

1. Python基本语法 多进程 线程 协程
2. Html css js jq bs ajax
3. Django MTV 反爬 日志 echarts
4. 搭建 redis集群 缓存 持久化
5. 搭建过负载均衡 处理高并发
6. 熟练掌握urllb requests scrapy 并研读过部分源码
7. 熟练掌握常见反爬手段 如
8. 熟练掌握分布式爬虫的设计与实现
9. 熟练掌握搭建HA集群的Hdfs做数据存储
10. 熟练掌握MR的清洗过程
11. 熟练 hbase zk 高可用
12. 熟练掌握hive 做数据二次清洗
13. 熟练sqoop于mysql做数据交互
14. 熟练掌握word，excel等离线数据的清洗
15. 熟练掌握数据结构
16. AI算法