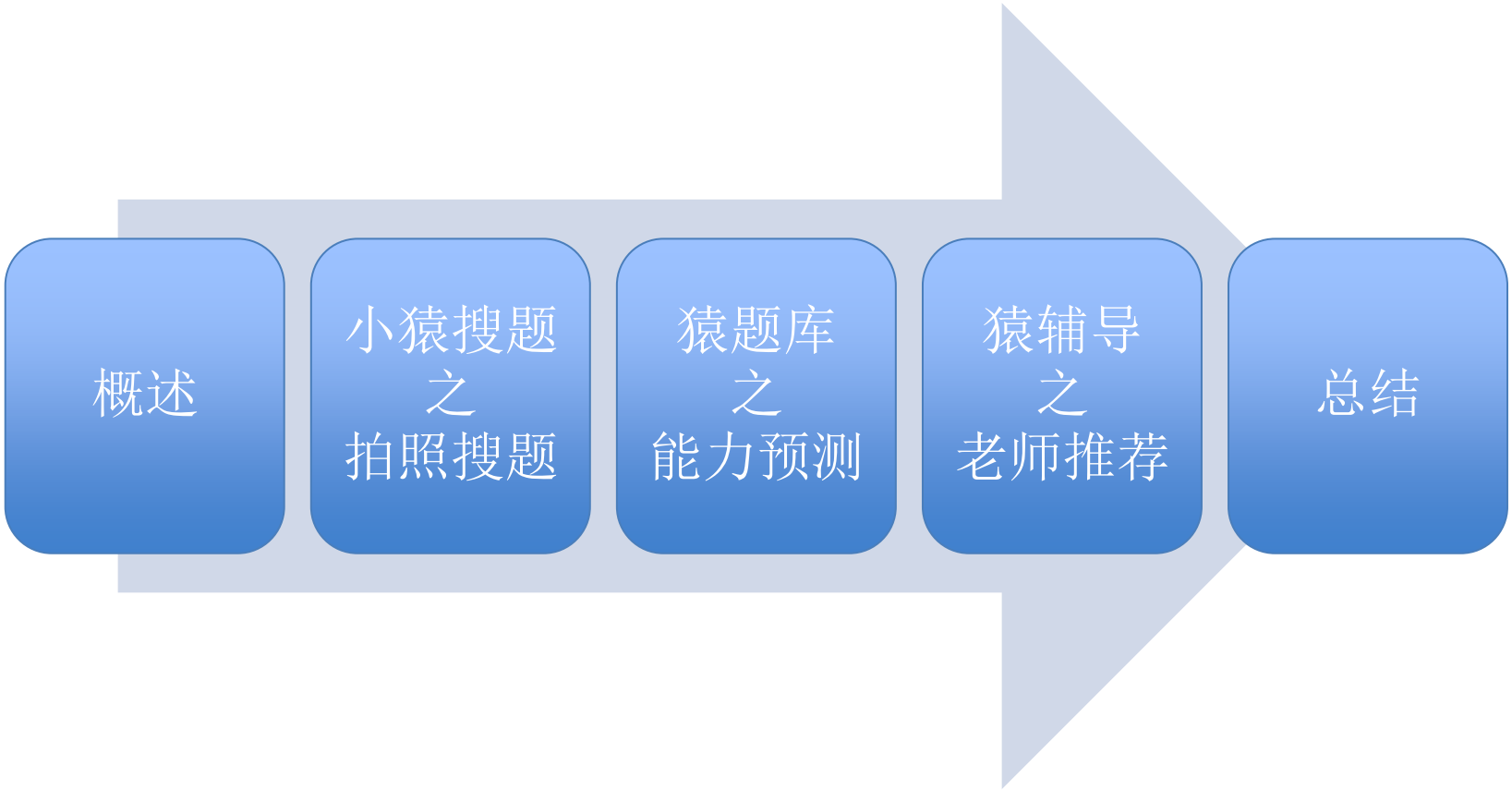
在线教育领域的机器学习应用

邓澍军
dengsj@yuantiku.com
2015.04.25

猿题库

提纲



机器学习

50-60年代
感知机

70-80年代
神经网络

90-00年代
SVM
Boosting
随机森林

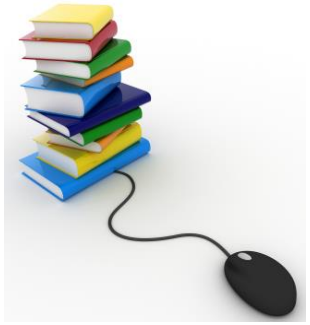
10年代
Deep Learning

90年代
网校：远程教育

00年代
传统教育转战线上

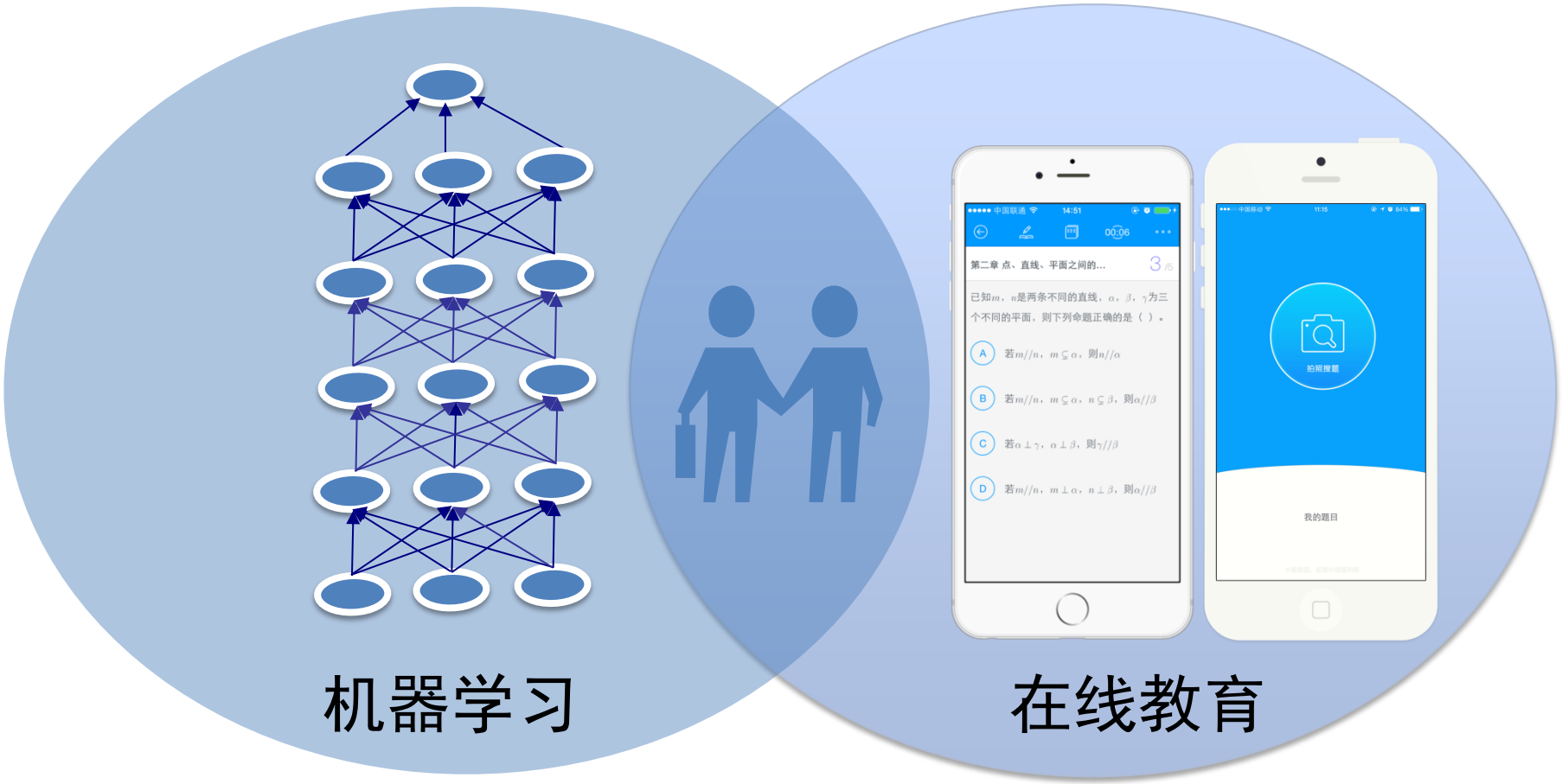
10年代
互联网公司进军在线教育

2012-
移动互联网在线教育



在线教育

机器学习邂逅在线教育

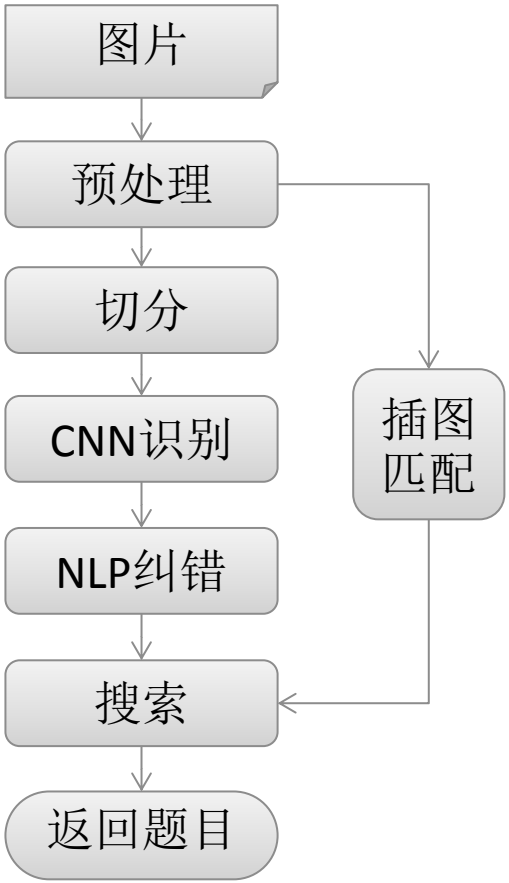

机器学习
在线教育
第二章 点、直线、平面之间的...
3/5
已知m，n是两条不同的直线，α，β，γ为三个不同的平面，则下列命题正确的是（ ）。
A 若m//n，m⊆α，则n//α
B 若m//n，m⊆α，n⊆β，则α//β
C 若α⊥γ，α⊥β，则γ//β
D 若m//n，m⊥α，n⊥β，则α//β
拍照搜题
我的题目

应用之一：小猿搜题之拍照搜题

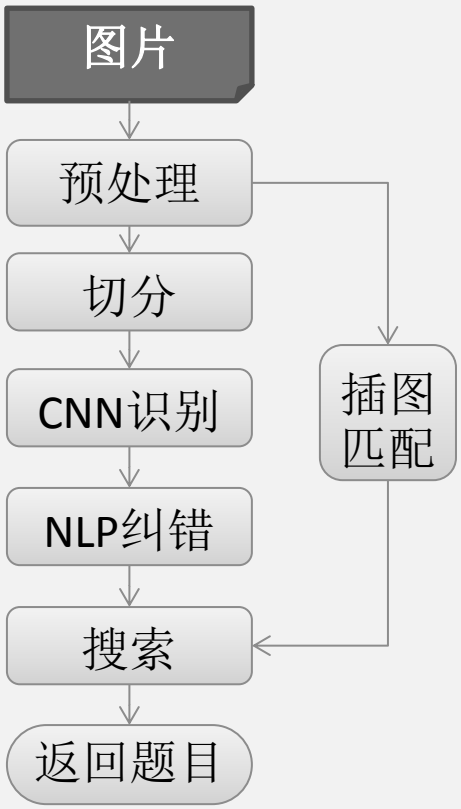
小猿搜题之拍照搜题


中国移动
11:31
31%
小猿搜题
(Ⅱ)从该小组同学中任选2人，求选
都在[18.5,23.9)中的概率.
D. $\frac{6\sqrt{7}}{7}$
的右焦点的连线交C于
渐近线，则 $p=$ (　)
D. $\frac{4\sqrt{3}}{3}$
最小值时，$x+2y-z$ 的
(　)
D. $\frac{9}{4}$
18.(本小题满分12分)
设函数 $f(x)=\frac{\sqrt{3}}{2}-\sqrt{3}\sin^2\omega x-\sin$
心到最近的对称轴的距离为 $\frac{\pi}{4}$.
(Ⅰ)求 ω 的值；
(Ⅱ)求 $f(x)$ 在区间 $\left[\pi,\frac{3\pi}{2}\right]$ 上的

小猿搜题技术框架

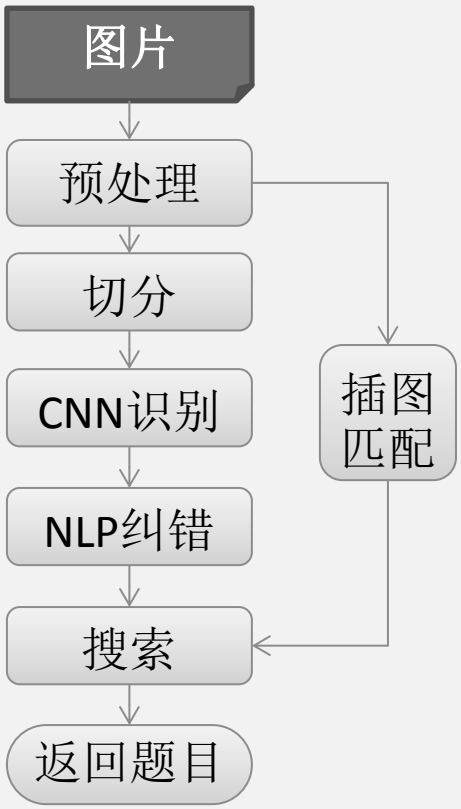


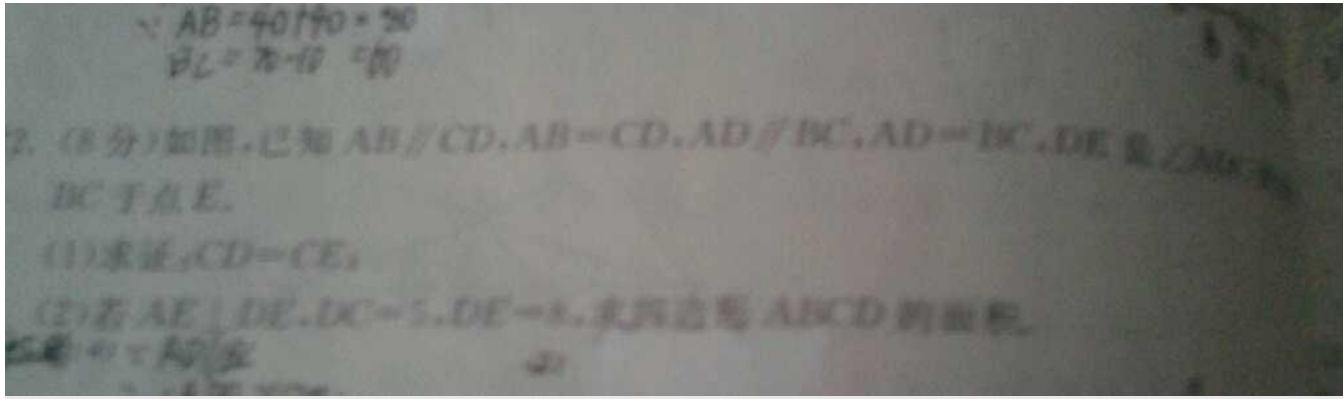
图片类别



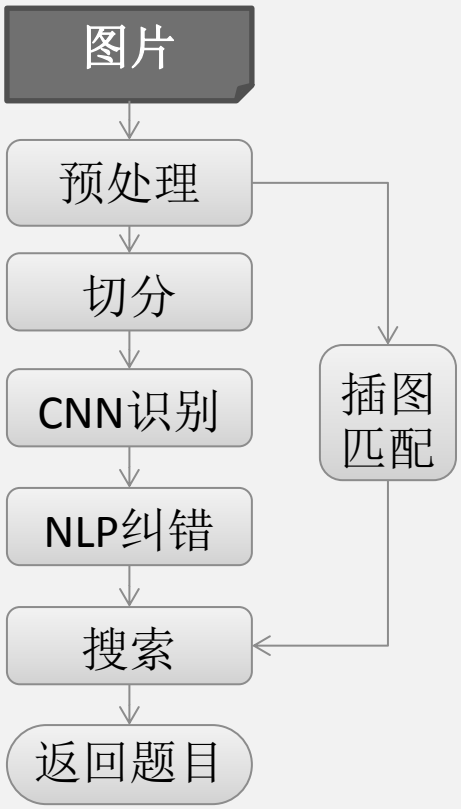
- ✓ 图片特征
  - ➢ 内容多样
    - 语数英等10来个科目
  - ➢ 模糊图很多，占**30%+**
    - 光照、扭曲、抖动等
  - ➢ 含有大量公式，数学占**50%**
    - 上下标、分式、根号等

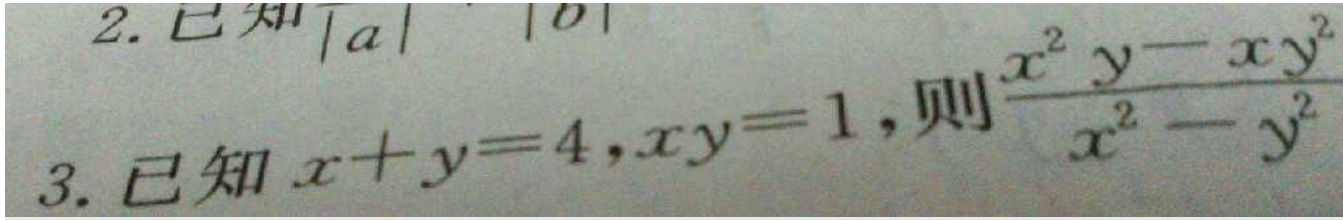
模糊图片

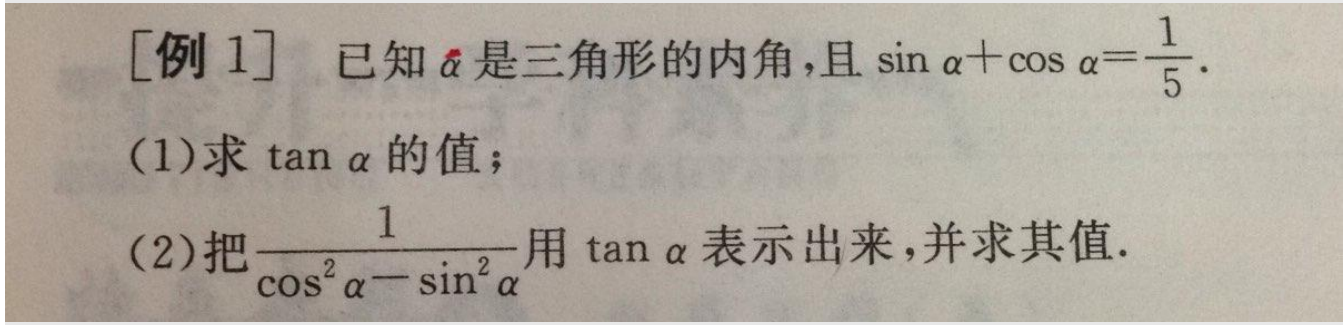


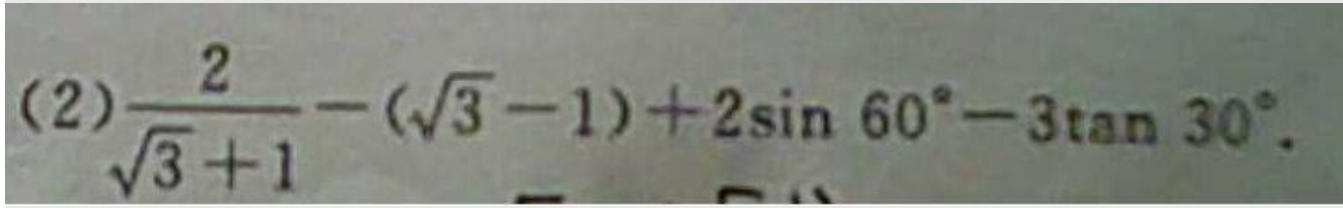


公式图片

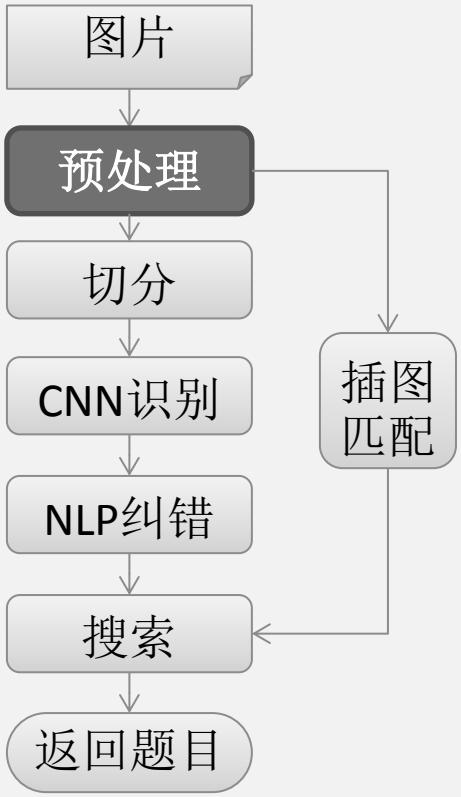

图片
预处理
切分
CNN识别
NLP纠错
搜索
返回题目
插图匹配

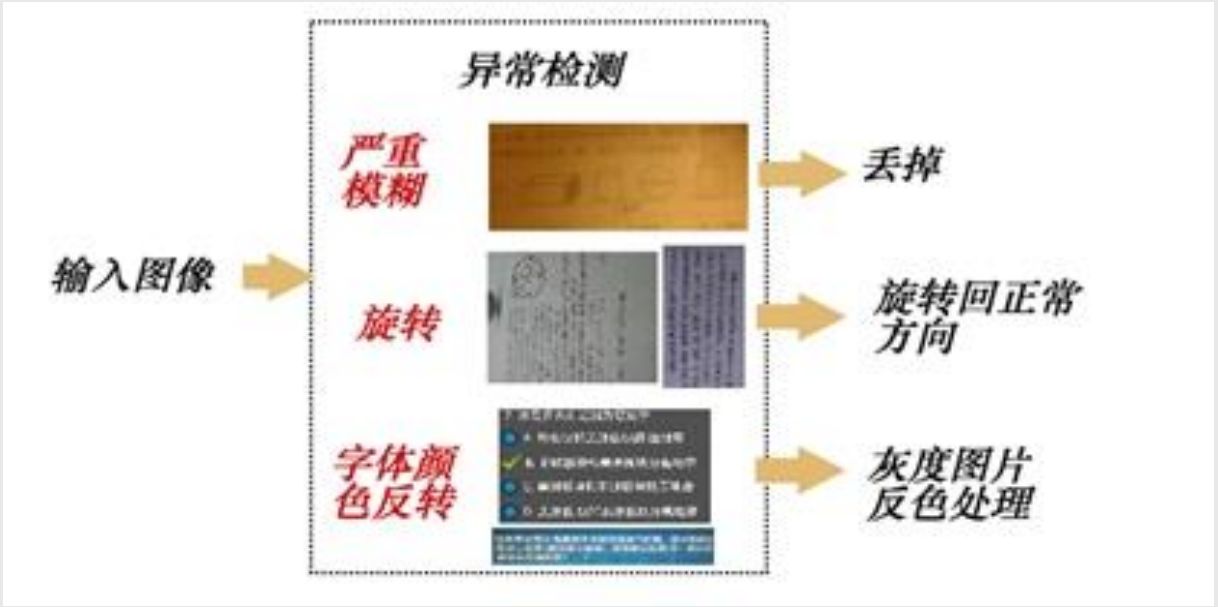

3. 已知 $x+y=4, xy=1$，则 $\frac{x^2y-xy^2}{x^2-y^2}$


[例 1] 已知 α 是三角形的内角，且 $\sin\alpha+\cos\alpha=\frac{1}{5}$.
(1)求 $\tan\alpha$ 的值；
(2)把 $\frac{1}{\cos^2\alpha-\sin^2\alpha}$ 用 $\tan\alpha$ 表示出来，并求其值.

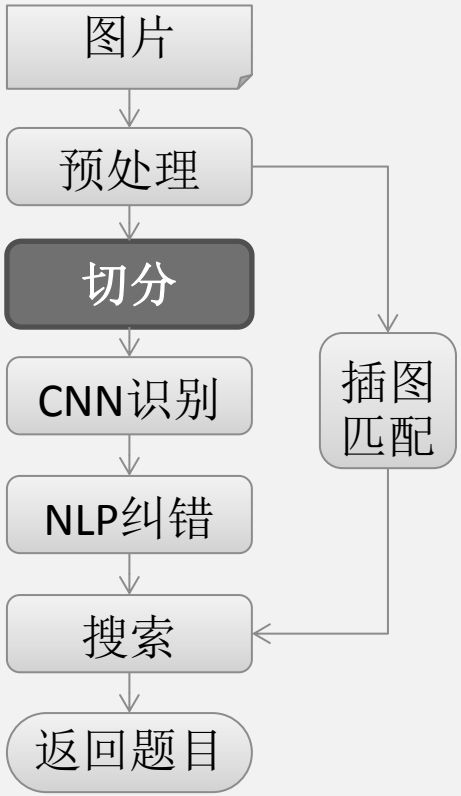

(2) $\frac{2}{\sqrt{3}+1}-(\sqrt{3}-1)+2\sin 60^\circ-3\tan 30^\circ$.

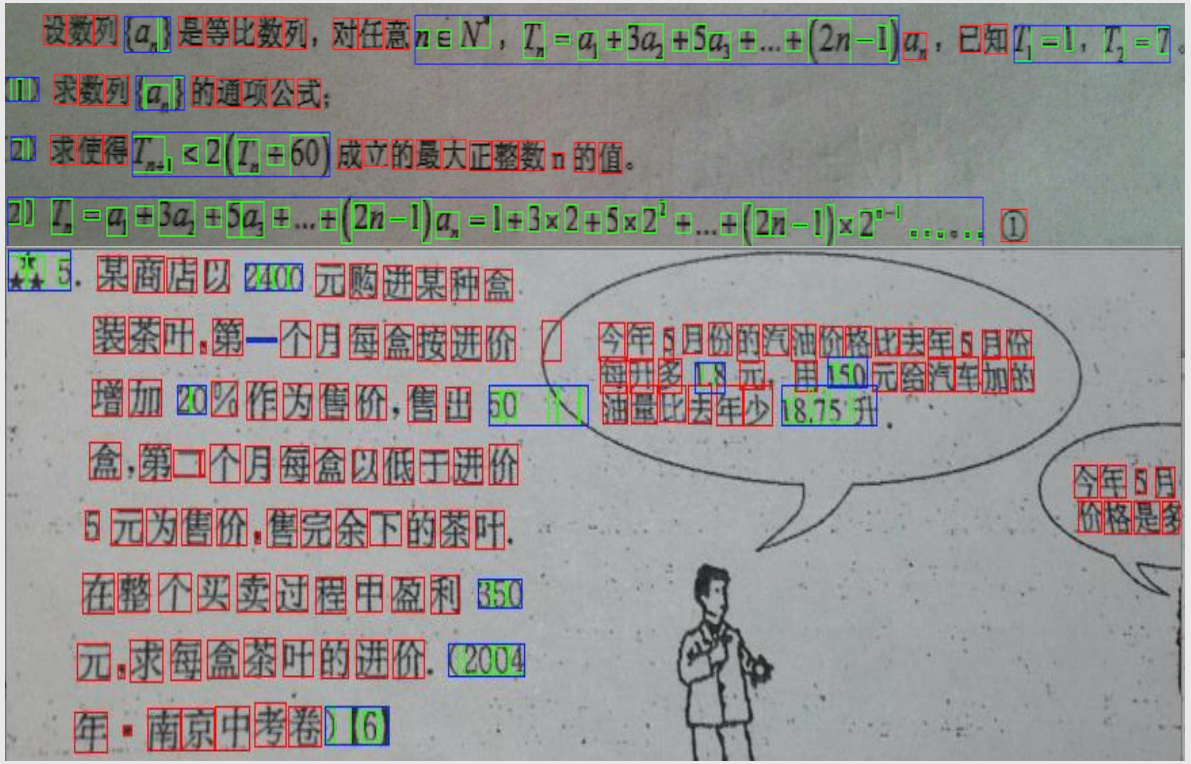
图片预处理


图片
预处理
切分
CNN识别
NLP纠错
搜索
返回题目
插图匹配

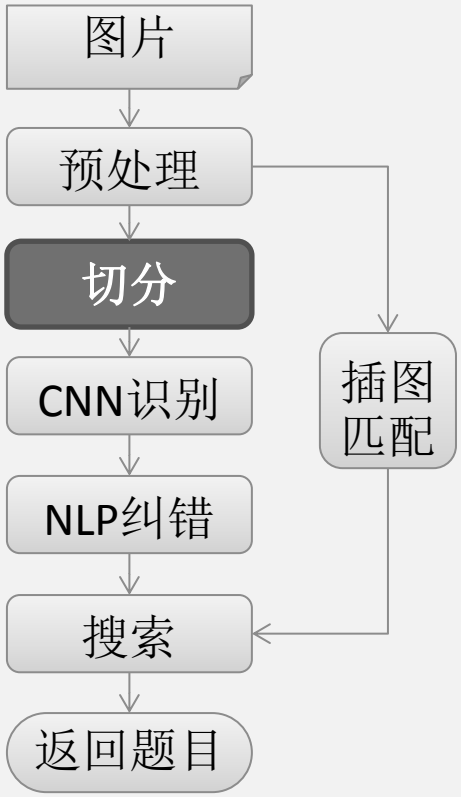

异常检测
输入图像
严重模糊
旋转
字体颜色反转
丢掉
旋转回正常方向
灰度图片反色处理

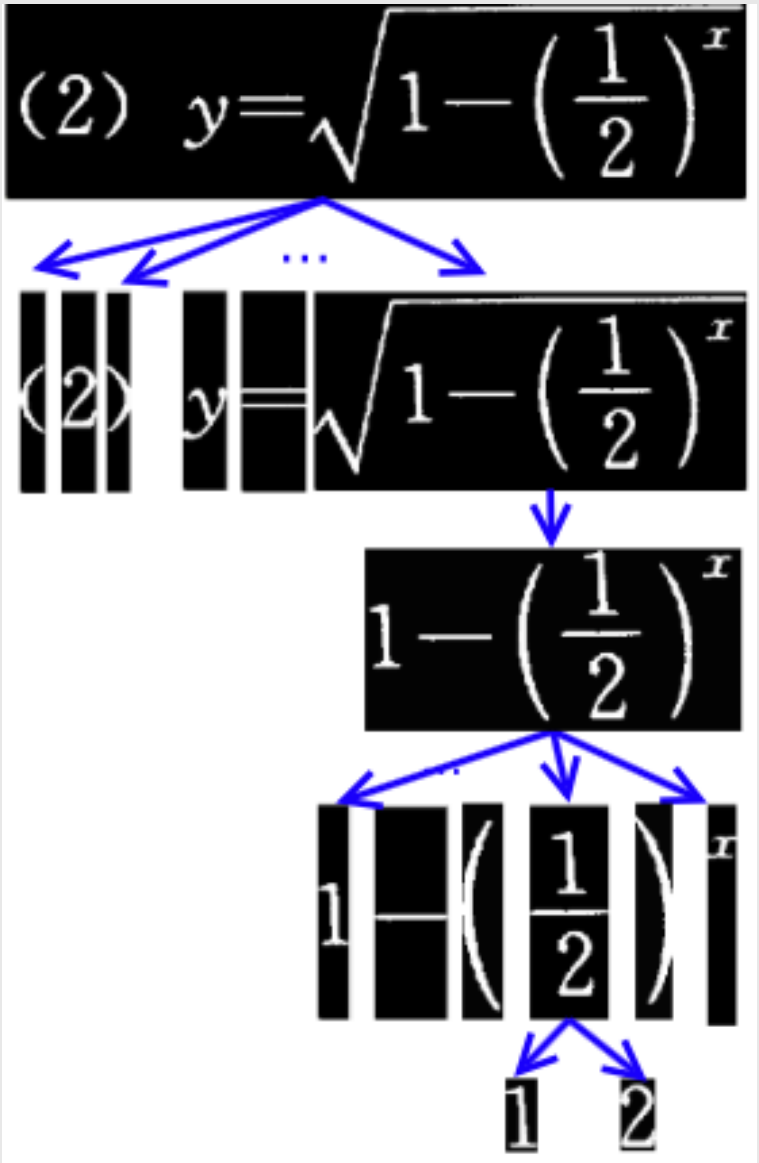
字符切分


图片
预处理
切分
CNN识别
NLP纠错
搜索
返回题目
插图匹配

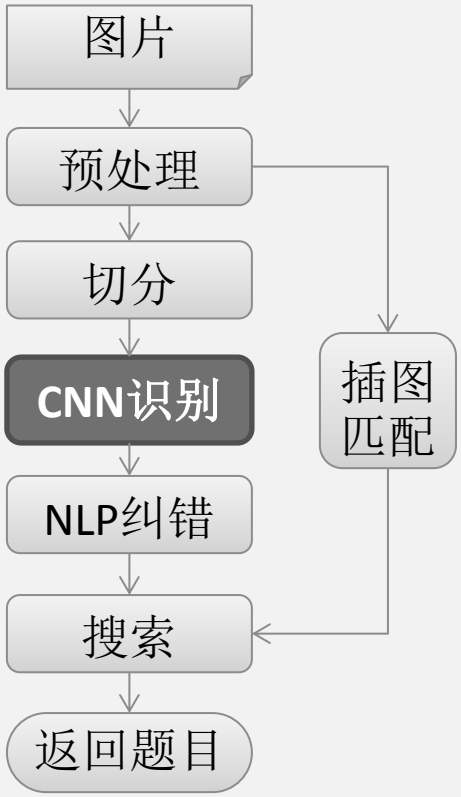

设数列{a_n}是等比数列，对任意n∈N*，T_n=a_1+3a_2+5a_3+...+(2n-1)a_n，已知T_1=1，T_2=7。
(I) 求数列{a_n}的通项公式；
(II) 求使得T_{n+1}<2(T_n+60)成立的最大正整数n的值。
(II) T_n=a_1+3a_2+5a_3+...+(2n-1)a_n=1+3×2+5×2^2+...+(2n-1)×2^{n-1} ..... ①
5. 某商店以2400元购进某种盒装茶叶，第一个月每盒按进价增加20%作为售价，售出50个盒，第二个月每盒以低于进价5元为售价，售完余下的茶叶，在整个买卖过程中盈利350元，求每盒茶叶的进价.（2004年·南京中考卷）[6]
今年5月份的汽油价格比去年5月份每升多1.5元，用150元给汽车加的油量比去年少18.75升.
今年5月
价格是多

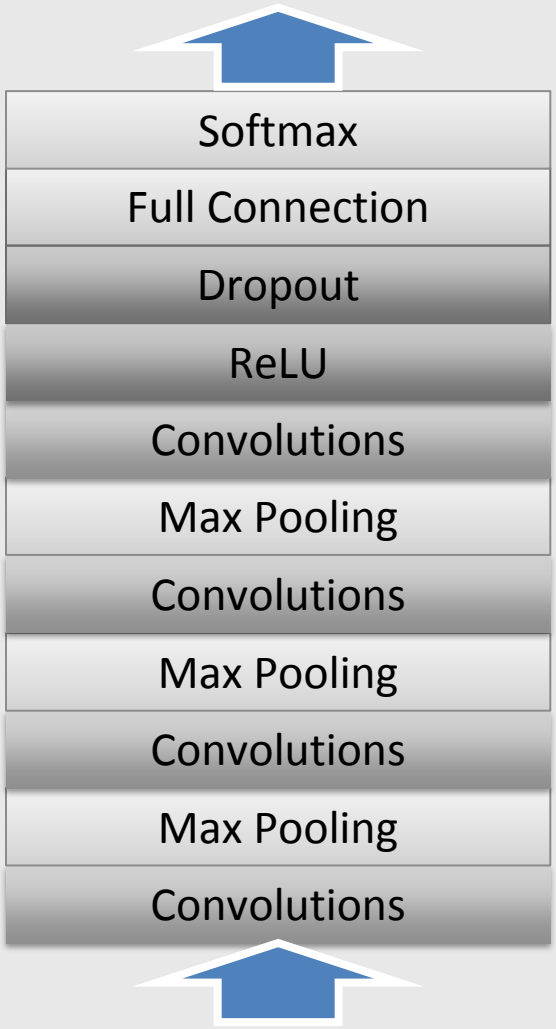
复杂公式切分


图片
预处理
切分
CNN识别
NLP纠错
搜索
返回题目
插图
匹配

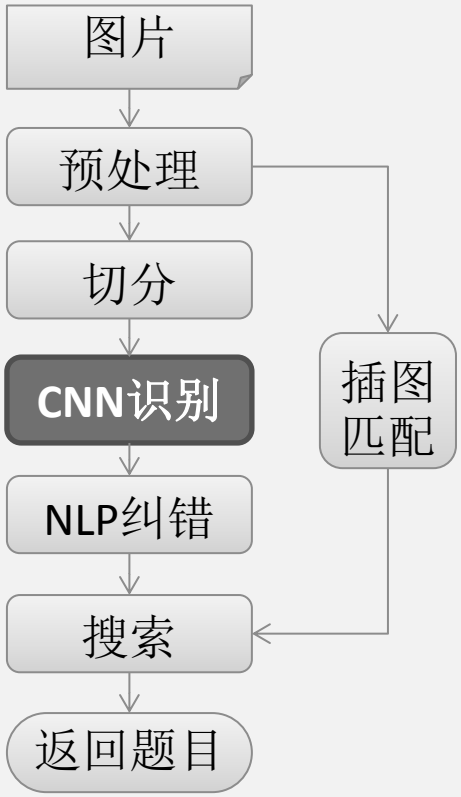

(2) $y=\sqrt{1-\left(\frac{1}{2}\right)^x}$
$1-\left(\frac{1}{2}\right)^x$
1
2

卷积神经网络模型

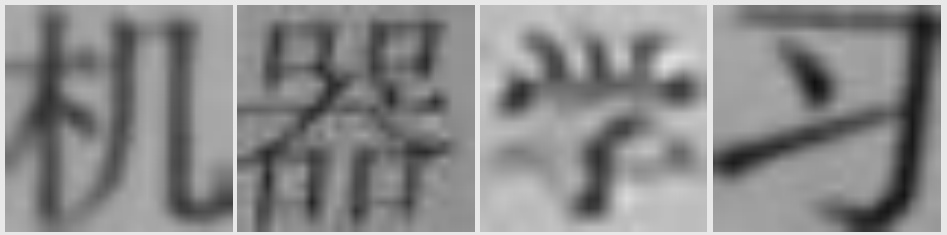

图片
预处理
切分
CNN识别
NLP纠错
搜索
返回题目
插图匹配


Softmax
Full Connection
Dropout
ReLU
Convolutions
Max Pooling
Convolutions
Max Pooling
Convolutions
Max Pooling
Convolutions

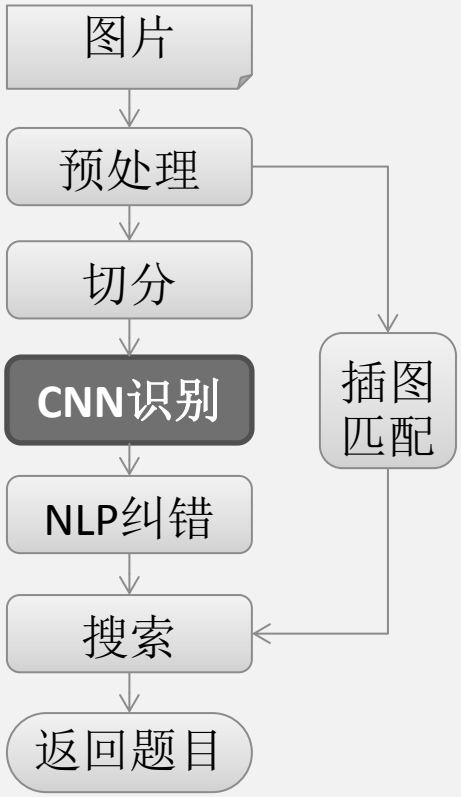
自动生成标注数据



- ✓ 标注数据自动生成
  - ➢ 场景融合
  - ➢ 旋转、拉伸等

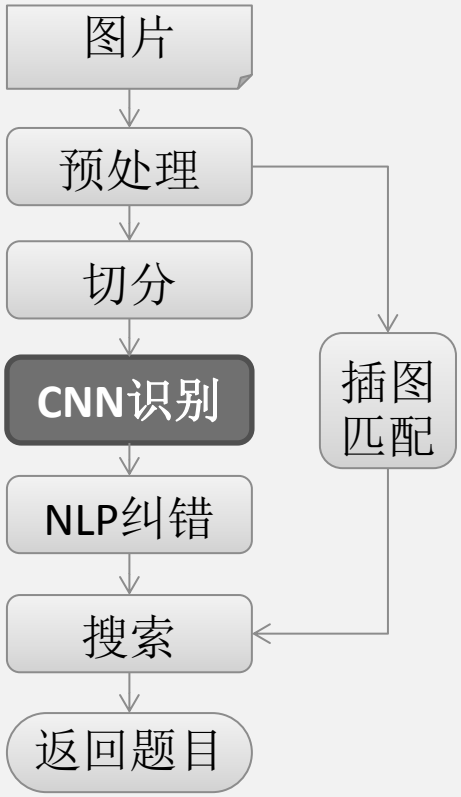


Deep Learning优化

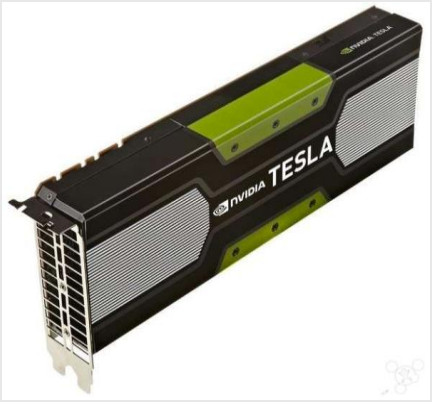


- ✓ Deep Learning的优化算法多种多样，模型最终的效果也不尽相同
  - ➢ 小猿搜题中尝试了多种不同优化算法
  - ➢ 一般来说，先SGD再采用Gauss-Newton能够在更短时间内收敛

Deep Learning加速



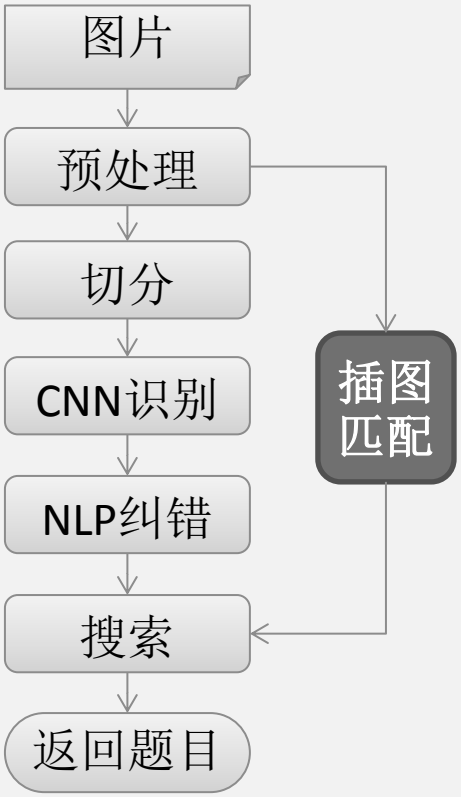
- ✓ 用GPU K40训练相比CPU模式速度能够提升**5-6**倍
- ✓ GPU K40线上预测速度能够提升**2-4**倍



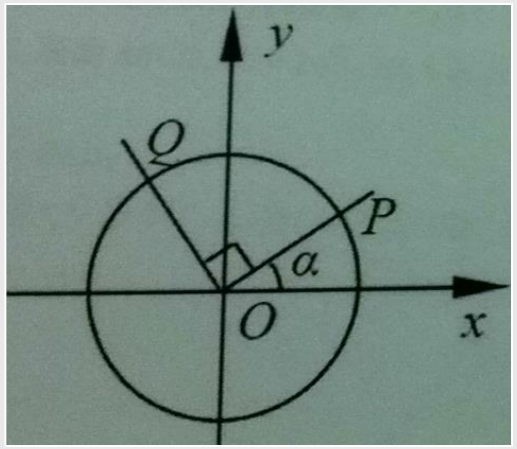
辅助策略——NLP纠错

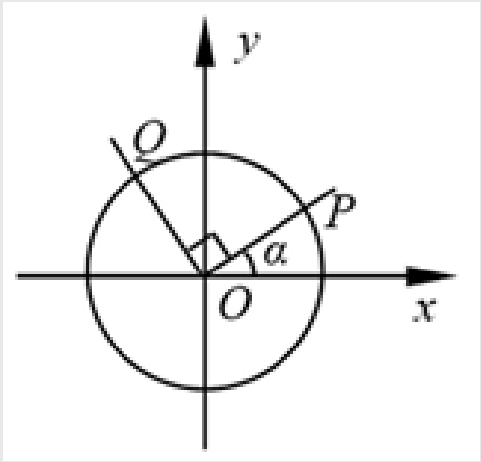
- ✓ 利用语言模型进行纠错
  - ➢ 平行回边形 → 平行四边形
  - ➢ 电灯炮 → 电灯泡
  - ➢ 入 ←→ 人
  - ➢ l ←→ 1

辅助策略——插图匹配

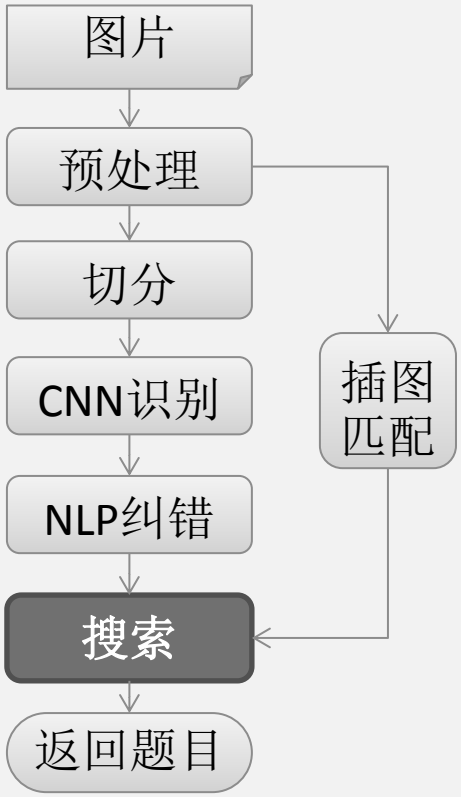


- ✓ 如果题目中的插图匹配(SIFT等特征)，则为加分项





搜索

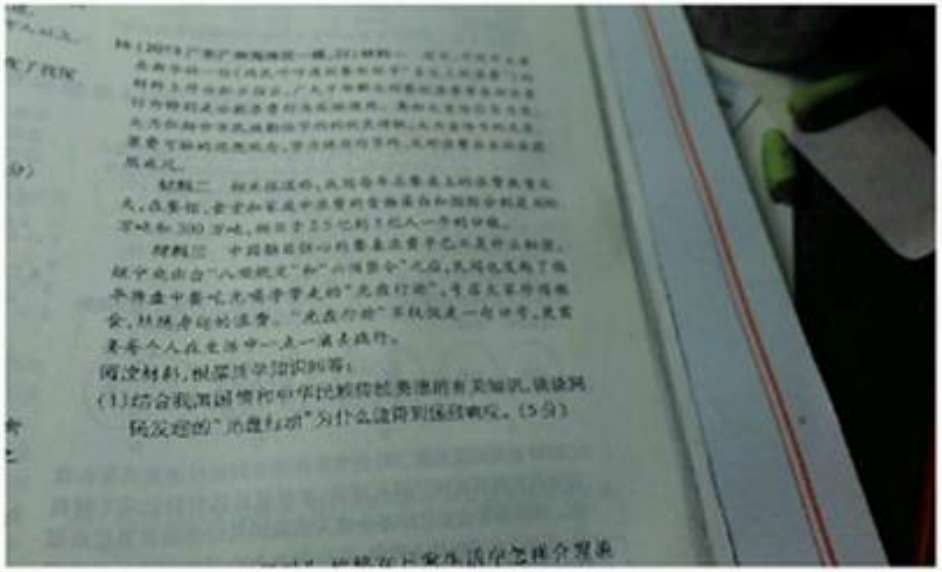


- ✓ 搜索主要模块
  - ➢ 分词
  - ➢ 倒排索引
  - ➢ 排序
    - Learning to Rank: GBRT

搜索示例

小猿搜题

初高中拍照搜题利器

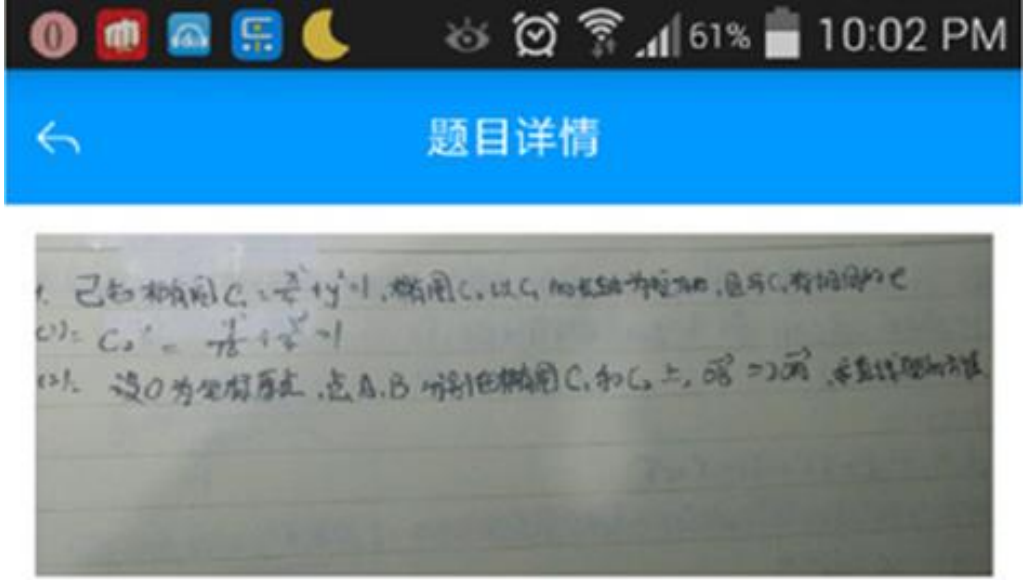


【题文】材料一：近日，习近平主席在新华社一份《网民呼吁遏制餐饮环节"舌尖上的浪费"》的材料商作出批示指出，广大干部群众对餐饮浪费等各种浪费行为特别是公款浪费行为反映强烈。要加大宣传引导力度，大力弘扬勤俭节约优秀传统。

材料二：相关报道称，我国每年在餐桌上的浪费数量巨大。在餐桌上浪费的事务相当于2.5亿到3亿人一年的口粮。

材料三：中国触目惊心的餐桌浪费早已不是秘密。一些网民倡导将盘中餐吃光喝净带走的"光盘行动"。号召大家珍惜粮食，杜绝浪费。"光盘行动"不仅仅是一句口号，更需要去践行。

搜索示例



2014年12月30日 22:01

搜索结果

题干

已知椭圆C_1：$\frac{x^2}{4}+y^2=1$，椭圆C_2以C_1的长轴为短轴，且与C_1有相同的离心率．

（1）求椭圆C_2的方程；

（2）设O为坐标原点，点A，B分别在椭圆C_1和C_2上，$\vec{OB}=2\vec{OA}$，求直线AB的方程．

小结

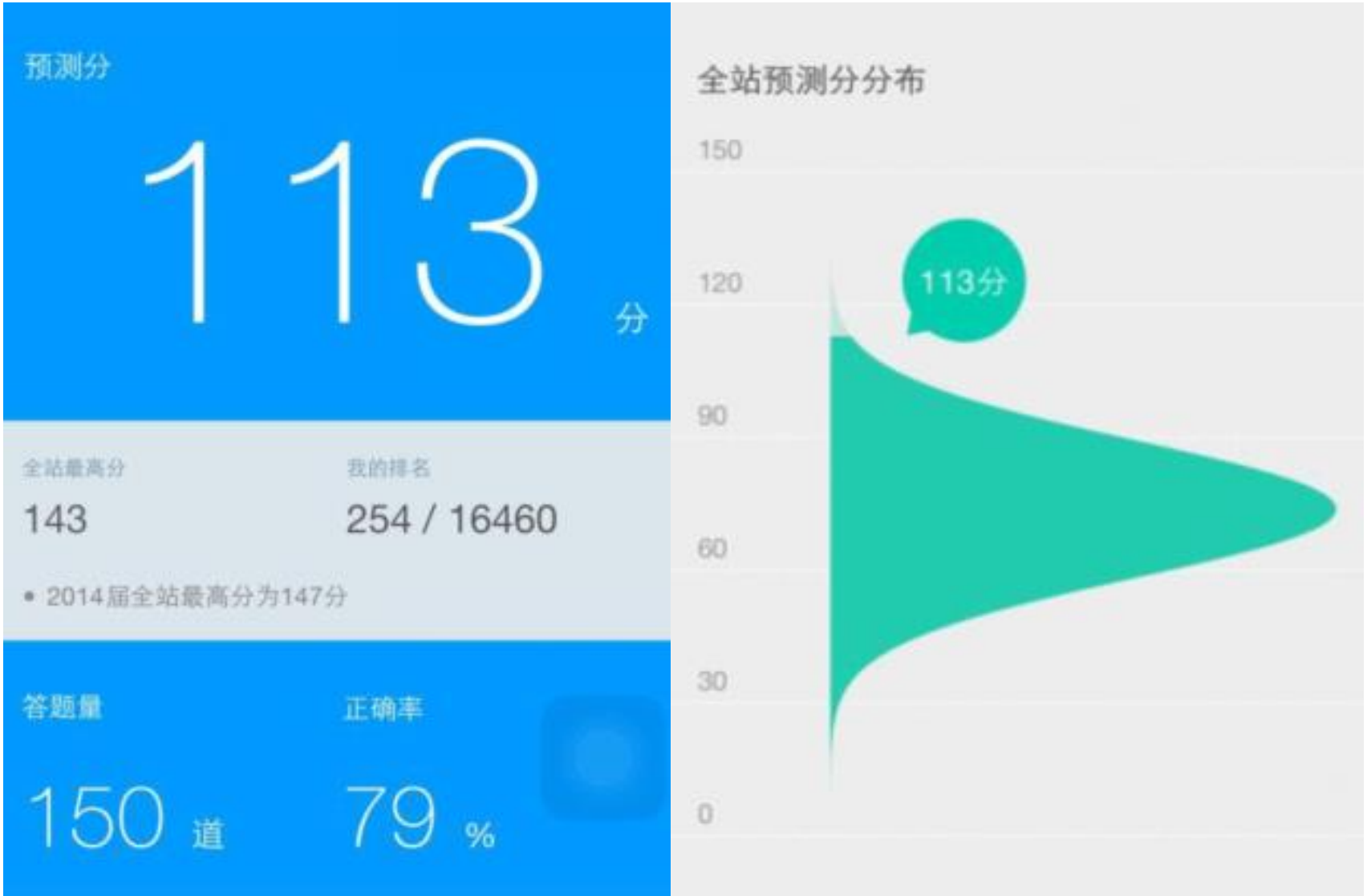
- ✓ 小猿搜题之拍照搜题
  - ➢ Computer Vision
    - 预处理
    - 切分
    - 训练数据自动生成
  - ➢ Deep Learning
    - 识别
  - ➢ NLP
    - 纠错
  - ➢ Learning to Rank
    - 排序

应用之二：猿题库学生能力预测

猿题库学生能力预测


中国联通
15:49
65%
英语
预测分
113
分
距离高考
48
答题量
150
练习天数
6
听力
0/605
语言知识
130/6373
完形填空
0/10975
阅读理解

猿题库学生能力预测(续)


预测分
113
分
全站最高分
143
我的排名
254 / 16460
• 2014届全站最高分为147分
答题量
150
道
正确率
79
%
全站预测分分布
150
120
90
60
30
0
113分

传统教育模型

- ✓ 项目反应理论(IRT)
  - ➢ 最简单的IRT模型

$$p_{irt}(y=1|\theta,b)=\frac{1}{1+e^{-(\theta-b)}}$$

  - ➢ 题目难度b
    - 标注
  - ➢ 学生能力θ
    - 模型参数，优化得到

机器学习模型

- ✓ 机器学习模型

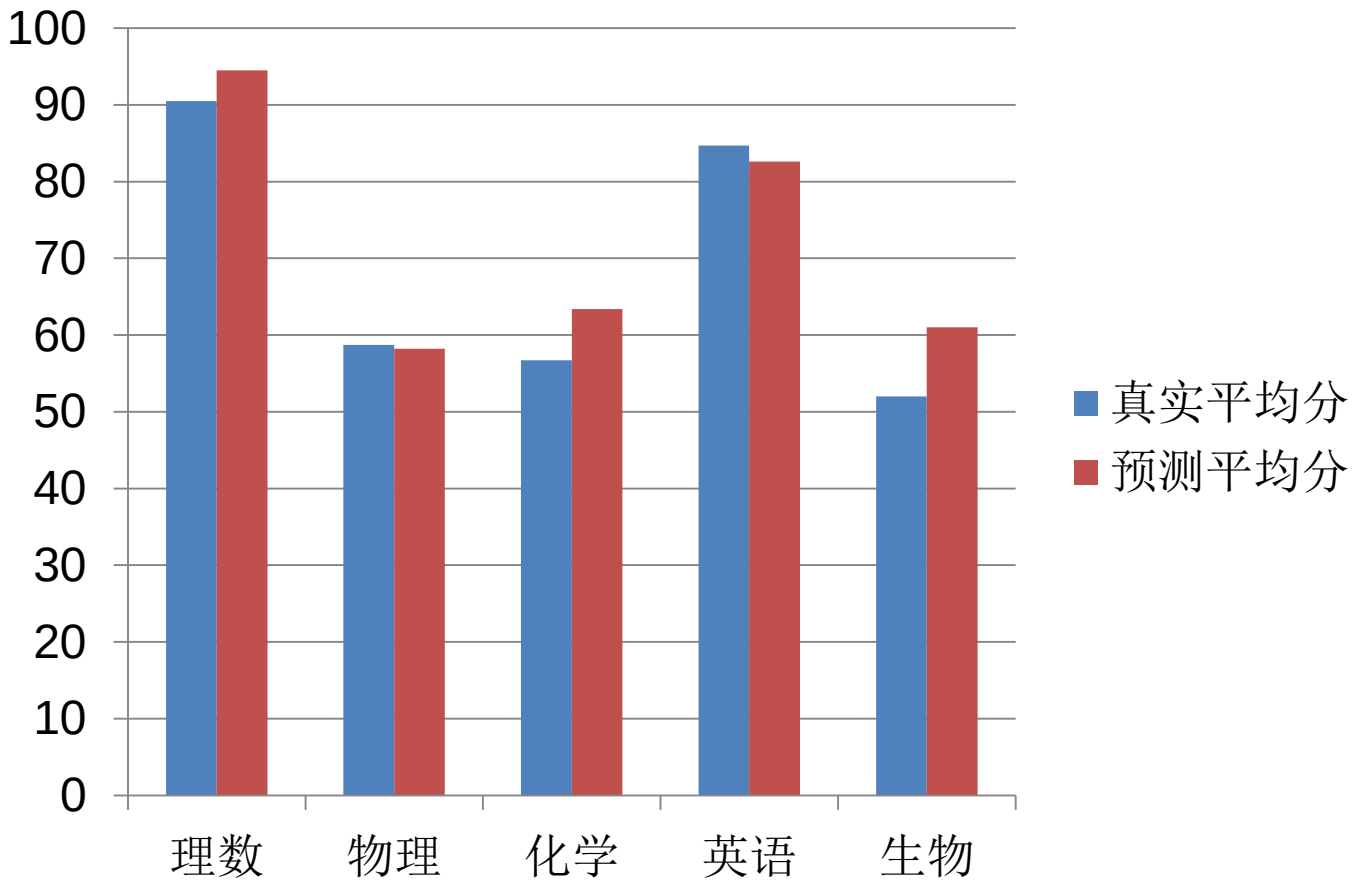
$$p(y=1|w,x)=\frac{1}{1+e^{-w'x}}$$

  - ➢ Offline model
    - Logistic Regression
  - ➢ Online model
    - Follow-the-Regularized-Leader

特征

- ✓ 所用特征
  - ➢ 用户相关特征
    - 学校，地区，目标考试，……
  - ➢ 题目相关特征
    - 知识点，关键词，难度，……
  - ➢ 时序特征
    - 距离高考时间，……
  - ➢ 组合特征

预测分评估



2014年广东省理科高考

小结

- ✓ 传统教育模型
  - ➢ 项目反应理论（Item Response Theory）
- ✓ 机器学习模型
  - ➢ Offline model
    - LR (Logistic Regression)
  - ➢ Online model
    - FTRL (Follow-The-Regularized-Leader)

应用之三：猿辅导老师推荐

猿辅导老师推荐


无练习 不辅导
90%
约课
课前练习
在线上课
听课报告
筛选老师
默认排序
150元/课时
数学 - 七/八/九年级
清华大学 3年教龄
共授7课时
好评率100%
150元/课时
数学 - 七/八/九年级
清华大学 6年教龄 学科带头人
共授8课时
好评率100%
150元/课时
数学 - 七/八/九年级
清华大学 6年教龄 竞赛辅导员
共授10课时
好评率100%
150元/课时
数学 - 七/八/九年级
7年教龄
共授3课时
暂无评价

猿辅导老师推荐(续)

- ✓ 推荐系统
  - ➢ 冷启动：Content-Based
  - ➢ Item-Based Collaborative Filtering
- ✓ 机器学习
  - ➢ Logistic Factorization Machine
  - ➢ Exploitation and Exploration(E&E)

总结

- ✓ 小猿搜题之拍照搜题
  - Deep Learning
  - Computer Vision
  - Learning to Rank
- ✓ 猿题库学生能力预测
  - 传统教育领域的项目反应理论（IRT）
  - 计算广告点击率预测模型LR、FTRL
- ✓ 猿辅导老师推荐
  - 推荐系统
  - LFM，E&E

未来

- ✓ 教育领域知识图谱
  - ➢ 学生的最优能力成长之路
- ✓ 手写识别
  - ➢ 手写拍照搜题
  - ➢ 解答题
  - ➢ 自动判卷
- ✓ 高考机器人
  - ➢ 机器自动出题
  - ➢ 机器自动做题
- ✓ 智能芯片
- ✓ ……

Q&A？

Thanks!

dengsj@yuantiku.com

猿题库