

基本情况

- 占总成绩的20%
- 在如下两个任务中选择一个
 1. 实现一个新闻推荐系统，在给定的数据集上验证效果
 2. 实现一个垃圾短信识别系统，在给定的数据集上验证效果

要求

- 基本要求：
 - 4—5人组队（确认组长1人）
 - 至少实现**两种**推荐（或分类）算法
 - 有算法对比试验及结果
 - 将全部研究过程写成报告提交
 - Introduction, related work, approach, experiments, etc.
 - 讨论问题与其它任务相比(如：电影推荐、图像分类等)的异同点
 - 讨论算法选择的理由、优缺点及可能的改进方案
- 加分项：
 - 做出demo的系统进行演示
 - 线上系统提交url和使用说明文档
 - 非线上系统把演示流程制作成PPT一并提交

数据及算法验证说明（新闻推荐）

- 1万名国内某著名财经新闻网站的用户
 - 一个月的全部浏览记录（2014年3月，11万+条记录）
- 数据格式（每行5个field）：

```
5218791 100648598 1394463264 消失前的马航370 【财新网】（实习记者葛菁）据新华社消息，马来西亚航空公司表示，与一架由吉隆坡
5218791 100648802 1394463205 马航代表与乘客家属见面 3月9日，马来西亚航空公司代表在北京与马航客机失联事件的乘客家属见面。沈伯
5218791 100648830 1394463196 马航召开新闻发布会通报失联航班最新情况 3月9日下午三点，马航在首都国际机场旁边的国都大饭店召开发
```

- 用户编号（已做匿名化处理）
 - 新闻编号
 - 访问页面时间
 - 新闻题目
 - 新闻正文
- 数据集分割
 - 按时间分割，前20天数据作为训练集，后10天数据作为测试集
- 算法验证
 - 精度：依据用户点击文档进行计算
 - 训练/在线推荐 时间对比

数据及算法验证说明（垃圾短信分类）

标签	短信内容
0	商业秘密的秘密性那是维系其商业价值和垄断地位的前提条件之一
1	南口阿玛施新春第一批限量春装到店啦 春暖花开淑女裙、冰蓝色公主衫 气质粉小西装、冰丝女王长半裙
0	带给我们大常州一场壮观的视觉盛宴
0	有原因不明的泌尿系统结石等

- 短信数据

- 带标签数据（用于模型训练和测试）

- 标签域：1表示垃圾短信/0表示正常短信
 - 文本域：短信源文本（进行了一些处理）

- 不带标签数据（用于线上模拟）

- 带标签数据集分割

- 随机分割，5-fold cross validation

- 算法验证

- 精度：Precision/Recall/F1

- 速度：线上预测时间

坐12个小时飞机身体已经疲惫不堪
为什么不能是你③以多数人的努力程度
地址位于天津市滨海新区响罗湾旷世国际大厦A座1801室

提交与评分

- 提交方式及内容：
 - 所有提交，由组长发送邮件至aox@ics.ict.ac.cn
 - 组队后，提交团队成员名单
 - 大作业完成后，提交团队报告，演示demo及团队分完成工
- Deadline：
 - 12月31号
- 评分标准：
 - 1. 实验设计正确性/严谨性，系统设计水平
 - 2. 报告写作水平
 - 3. 团队成员参与度
 - 大作业评分按百分制、最终按照比例折合到总成绩中

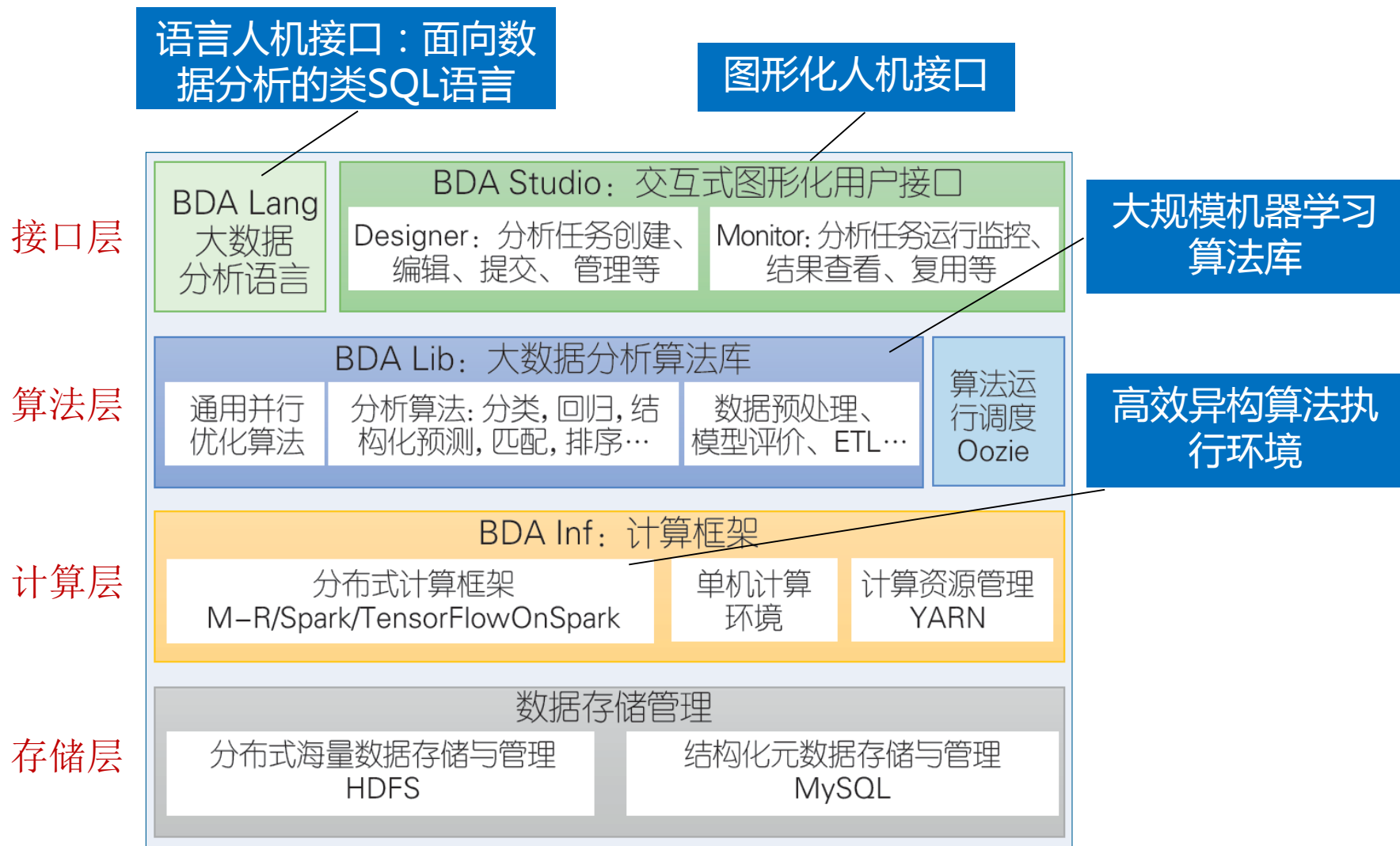
推荐算法总结

- 内容过滤
 - 用户、文档都表达为相同维度的向量（词向量），与用户最为相似的文档被优先推荐
 - 代表模型：KNN
 - 缺点：未能考虑用户与用户、文档与文档之间的相似关系
- 协同过滤
 - 推荐中考虑用户相似性、文档相似性（相似的用户喜欢相似的文档）
 - 代表模型：非负矩阵分解（NMF）
 - 缺点：冷启动，未能考虑用户画像信息和文档内容信息
- 基于特征的推荐模型
 - 综合考虑内容过滤和协同过滤，克服上述缺点
 - 代表模型：SVDFeature（KDD CUP 2012年冠军模型）、LibFM
- 分布式推荐算法实现
 - Spark MLlib、Mahout等提供了协同过滤算法NMF的并行/分布式实现

分类算法总结

- 线性分类器
 - Perceptron、SVM、Logistic Regression
- 非线性分类器
 - 决策树
 - Boosting（如树与Boosting结合：GBDT）
- 分布式算法实现
 - Spark MLlib

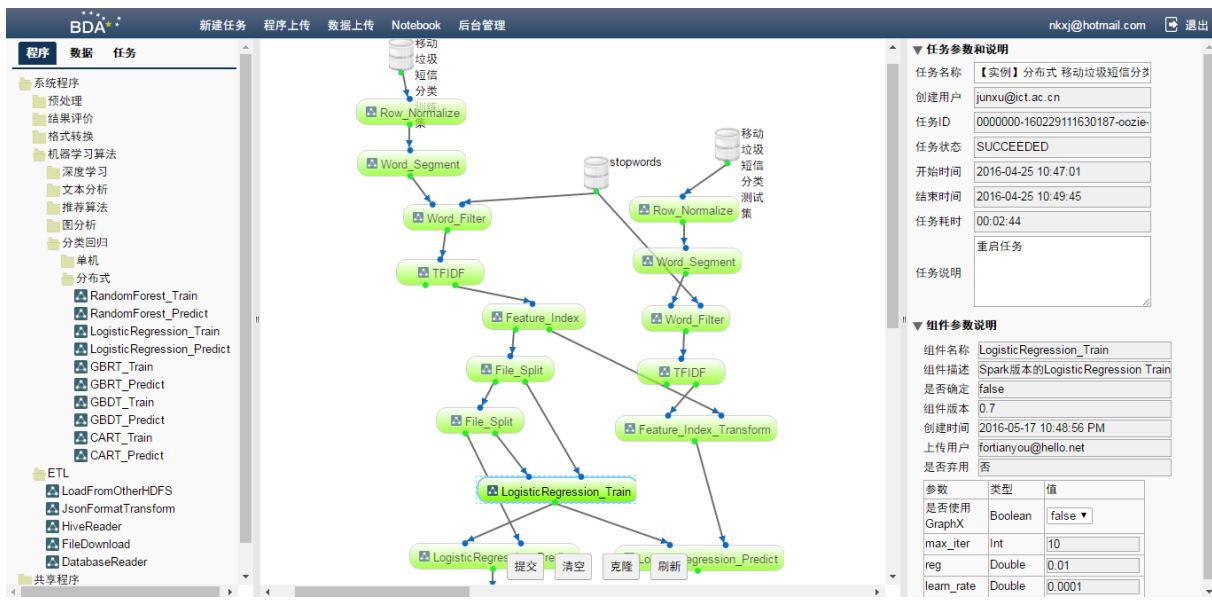
推荐使用计算所开源的 Easy Machine Learning



开放大数据分析云服务平台

<http://159.226.40.104:18080/dev/>

- 分析业务抽象：有向无环数据流图(dataflow DAG)
- 系统特性
 - **开放**：对数据以及功能开放，支持系统和用户自定义资源
 - **高效**：快速分析，结果复用，便捷展示；一站式分析服务
 - **易用**：显著降低学习成本和开发难度
 - **直观**：交互式分析，拖拽式编程
 - **兼容**：整合异构资源，适用于各种不同的计算与存储资源
 - **协作**：支持多人多平台协作开发



开源项目 Easy Machine Learning



<https://github.com/ICT-BDA/EasyML>

- Github Trending中连续一星期**最火**Java项目
- 获得**1400+** stars

ICT-BDA / EasyML

Unwatch 121 Unstar 1,418 Fork 278

Code Issues 11 Pull requests 0 Projects 0 Wiki Settings Insights

Easy Machine Learning is a general-purpose dataflow-based system for easing the process of applying machine learning algorithms to real world tasks.

machine-learning-studio Manage topics

72 commits 1 branch 0 releases 5 contributors Apache-2.0

Branch: master New pull request Create new file Upload files Find file Clone or download

sinllychen Modify international resource encoding mode. Latest commit 6718d7d 8 days ago

.settings	Remove local respository jar and update some settings	4 months ago
img	check images	3 months ago

Docker Toolbox 搭建EasyML系统

□ 搭建流程说明

- ✓ <https://github.com/ICT-BDA/EasyML/wiki/Docker-Toolbox-%E6%90%AD%E5%BB%BAEasyML-%E9%9B%86%E7%BE%A4>

□ 系统要求

- ✓ Windows 64位系统或Mac 64位系统
- ✓ CPU需支持VT-X/AMD-v功能（验证：通过Bios查看是否有Virtualization Technology）
- ✓ 系统内存大小至少4G及以上，硬盘容量至少10G及以上

□ 版本说明

- ✓ 简化版：伪分布式版本，集群仅包含一个master容器。（适合内存4G~8G机器，请**尽量使用8G机器运行**）。下载链接：<https://pan.baidu.com/s/1b3kMAU>
- ✓ 完整版：分布式版本，集群包含1个master容器，2个slave容器。（适合内存8G以上的机器，请**尽量使用8G以上**的机器运行）。下载链接：<https://pan.baidu.com/s/1miFXzVq>



微信交流群

谢谢！