**项目需求**

该项目旨在实践大数据应用和分析。应建立推荐服务系统，允许用户浏览大量的数据，并根据推荐做出决策。该系统在后台管理所有数据，并支持在互联网上浏览和搜索界面。这些功能可以通过集成工具来提供。

每个团队将根据团队决策选择主题并分配工作负载。

项目的所有部分都应该单独完成或由一组学生完成。团队将在团队任务开始之前建立。每队可有4-8人。由于团队合作是实际IT服务世界的基石，因此在这种协作学习环境中，作为虚拟团队的一员体验非常重要。

编程不是这个项目的重点，它应该直接与用户的观点/体验或面向服务的任务相联系。换句话说，只要系统是可用的、有用的和有效的，使用任何编程技巧或语言都无关紧要。

截止时间：目标到最后一节课的午夜。

**项目部分**

这个项目可以分为五个部分：

第1部分：项目计划（15%）

项目计划应该包含IT项目所需的信息，如项目目标、动机、资源估算、进度安排等。

第2部分：数据转换（15%）

应收集并格式化大量数据。这些数据应该作为一个数据集进行检查和设计，以便进一步利用大数据分析。

第三部分：推荐制度（30%）

应设计和建立一个推荐服务系统。服务系统是主要的开发和测试平台。一旦服务系统建立并运行，用户将能够与后端交互，以获得基于大数据分析的建议。

第4部分：实现服务接口（15%）

应该提供一个服务接口，允许用户与推荐系统进行交互。该服务应支持对数字属性（如出生日期）和文本属性（如技能/个性描述）的查询，并允许对文本属性进行简单的基于关键字的搜索。

第5部分：文件（25%）

最后，应该生成文档来总结项目并指导用户完成培训过程。以下项目可能会有所帮助：

a） 系统架构图和配置。

b） 服务系统的用户指南或操作菜单。

c） 潜在的培训过程/材料，如演示、屏幕截图或PowerPoint演示文稿。

d） 源代码和可执行程序包。

e） 项目团队工作报告，包括

1）列出团队成员名单；2）每个人的任务分配；3）每个角色贡献的工作量。

**Notes**

**合作伙伴的选择**

学生可以单独实施项目，也可以4-8人一组实施。选择权由每个学生决定，但在选择项目合作伙伴时，请记住以下规则：

预计工作量相同，个人和团队项目采用相同的评分标准。教职员工的经验表明，一般来说，团队合作并不一定更容易或更有成效——这主要取决于个人偏好和工作风格。

如果你在一个团队里工作，那么就要仔细选择你的搭档。团队可以在课程期间的任何时候“离婚”（由于不相容、一方放弃课程或任何其他原因），并且随着项目的进展，个别学生可以选择组队，但是离婚团队的学生不能组建新团队或加入其他团队。换句话说，如果一个学生作为一个团队的一部分提交项目的任何部分，那么项目的每个后续部分都必须单独提交或者作为同一个团队的一部分提交。

对于“我做了所有的工作，我的搭档什么也没做”的投诉，我们将不予受理！

如果你在一个团队中工作，你的工作必须作为一个提交文件一起提交。如果我们单独收到一个小组的学生提交的材料，他们将被罚10%的折扣。此外，如果以上分开提交的材料不同，则以得分较低的提交材料为准。作为个人作品提交的团队合作将被视为剽窃，并通过大学官方渠道进行处理。

**学术诚信**

你们每个人都要提交自己的原创作品，或者项目团队的原创作品。在许多情况下，向其他人（讲师、助教或其他学生）寻求提示或调试帮助，或者一般性地谈论编程策略是很有用的。这样的活动是可以接受和鼓励的，但您必须指出您收到的任何帮助（人力、在线或其他）。任何未被适当引用的协助将被视为剽窃。在任何情况下，您都有责任编码、理解并能够自己或作为一个团队解释您提交的所有项目工作。

我们将积极追查所有涉嫌剽窃的案件，并通过大学官方渠道进行处理。如果您对本政策或我们将在多大程度上追究违反学术诚信的行为有任何疑问，请立即与课程导师讨论您的问题。

**项目主题样本（美国大学生使用）**

**项目方案1**

**股票推荐制度**

动机

帮助股票购买者做出更明智的选择

寻找善于从股票中获利的人（专家）

使用基于用户的协同推荐，寻找买家和专家之间的相似度

从最相似的专家那里向股票购买者推荐一些股票。

减轻股票购买者的重担，看完成千上万的股票，做出明智的选择。

数据集

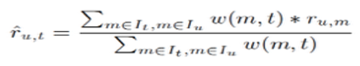
纳斯达克证券交易所数据

●雅虎金融历史价格数据集

●100000用户的模拟交易记录

算法

●基于用户的推理标签协同过滤



工具

•Eclipse J2EE、Mahout、Maven、MySQL、PHP服务器

进展

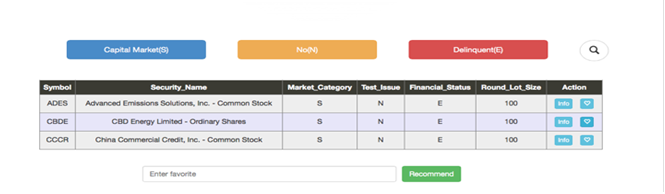
○实现的UI和数据库

○根据推断的标签等级确定基于用户的协同过滤算法。

预期捐款

○利用推断的标签评级实现改进的协同过滤算法。

○根据专家的购买选择分析，向客户提供合理的股票建议。



**方案2项目**

**电影探索**

动机

观众偏好：

我喜欢喜剧；我喜欢大片；我喜欢小罗伯特唐尼；

我想和我女儿一起看电影，我。。。。。。

生产商指导：

顾客想看什么样的电影？

哪个演员最适合这个角色？

有什么特质可以让电影令人难忘？

数据集：在线搜索

工具：灵活

算法：

Pig：过程用户配置文件

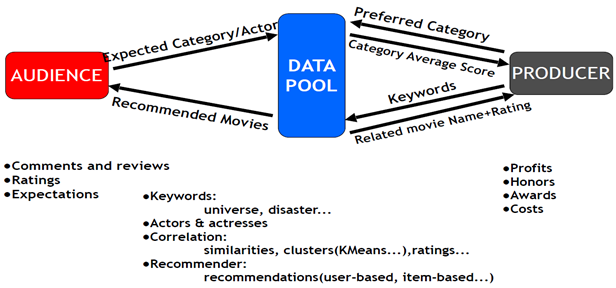
分类（&C）：

在观众中进行关联

建议：

根据关键字、提取和首选项进行断言

预期捐款



**项目方案3**

**梦幻篮球-制胜策略！**

动机

在美国，体育运动在人们中扮演着重要的角色。因此，人们把更多的时间花在看NBA，MLB，NFL…等比赛上作为他们的休闲时间。在过去的五年里，ESPN和雅虎发布了一款所有职业体育联盟的在线体育游戏，让人们有机会在网上组建一个模拟团队，与其他玩家竞争。在线球队的排名是基于他们为球队挑选的球员的真实行为。因此，选秀成为比赛胜利的关键因素。我们希望为起草工作提供一个成功的策略。同时，希望能为真正的NBA球队选择球员提供一个切实可行的动态策略。

目标：

每轮为在线游戏玩家推荐一名玩家

-将所有球员分为不同的小组，包括得分、篮板等。

在每个分组内对玩家进行排名

-根据在线游戏玩家的现状提出建议

数据集

-NBA球员统计

http://downloads.nbastuffer.com/nba-player-data-sets

或者在底部

https://www.nbastuffer.com/

算法

-聚类

Kmeans聚类、f-Kmeans聚类等

PLA算法、线性回归等

-分类、建议

工具

Mahout、Matlab和课堂上提供的所有资源

进展

-梦幻篮球得分体系研究综述

-搜索所有可能对我们的建议有用的数据集

-熟悉我们可能使用的所有工具和算法

预期捐款

-为游戏建立动态推荐系统

-帮助玩家为游戏建立一个平衡的冠军团队

**项目选项4**

**如何给你的新生儿取名？**

动机

在这个项目中，一些（相对）简单，但在回答一个现实问题方面非常重要的事情可以进行。

你应该给你的新生儿取什么名字？我会查看历史上流行的男女宝宝名字，并使用各种推荐方法，以及分类和/或聚类技术来帮助预期的父母做出正确的名字选择，这样他们的孩子就不会抱怨自己的下半生没有那么酷的名字了。

数据集

从社会保障管理局（socialsecurity Administration）可以追溯到19世纪的流行男女宝宝名字。

算法

基于用户和基于项目的推荐、聚类和/或分类

工具

R

HDFS/Hive/PIG/HBase

行家

预期捐款

为用户建立动态“姓名”推荐系统