## 1. 项目需求

该项目旨在实践大数据应用程序和分析。应该建立推荐服务系统，以允许用户浏览大量数据，并根据推荐做出决策。该系统在后端管理所有数据，并支持Internet上的浏览和搜索界面。这些功能可以通过集成工具来提供。

每个团队将选择主题并根据团队决策分配工作量。

项目的所有部分都应单独完成或由一组学生完成。在开始分配团队之前，将建立团队。每个团队可能有4-8人。由于团队合作是实际IT服务世界的基石，因此在此协作式学习环境中体验成为虚拟团队的成员非常重要。

编程不是该项目的重点，应直接与用户的观点/经验或面向服务的任务联系起来。换句话说，只要系统可用，有用且有效，那么使用任何编程技能或语言都无关紧要。

提交时间：最后一节课当天零点

## 2. 项目组成

暂时，项目可以分为5个部分：

1. 项目计划（15%）

应该生成一个项目计划，包含项目所需的信息，例如项目目标、动机、资源估计、进度安排等

1. 数据转换（15%）

应收集并格式化大量数据，检查数据并将其设计为数据集，以进一步利用大数据分析

1. 推荐系统（30%）

应设计和建立推荐服务系统，服务系统是主要的开发和测试平台。服务系统设置并运行后，用户将能够与后端进行交互，以基于大数据分析获取建议。

1. 实现服务接口（15%）

服务界面应可用，以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/性格描述）的查询，并允许对文本属性进行基于关键字的简单搜索。

1. 文件（25%）

最后，应生成文档以包装项目并指导用户进行培训。以下各项可能会有所帮助：

* 系统架构图和配置
* 服务系统的用户指南或操作菜单
* 潜在的培训过程/材料，例如演示、屏幕截图或PowerPoint演示文稿
* 源代码和/或可执行程序包
* 项目组工作报告，包括：
  + 名单组成员
  + 每个人的任务分配
  + 每个角色贡献的工作量

## 3. 注意事项

### 3.1 选择合作伙伴

学生可以单独实施项目，也可以4-8人为一组。选择权取决于每个学生，但是在选择项目合作伙伴时请牢记以下规则：

预期每个人有相同的工作量，并且个人和团队项目使用相同的评分等级。教师的经验表明，一般而言，团队合作不一定会更容易或更有效率-这很大程度上取决于每个人的个人喜好和工作风格。

如果您在团队中工作，请仔细选择合作伙伴。团队可以在课程中的任何时候“离婚”（由于不兼容，一个合作伙伴放弃课程或任何其他原因），并且个别学生可以选择随项目进展而组队，但是离婚团队的学生则不能组建新团队或加入其他团队。换句话说，如果学生以团队的形式上交项目的任何部分，则项目的每个后续部分都必须单独或以同一团队的形式上交。

我们不会以“我完成了所有工作，而我的伙伴却什么也不做”的形式提出任何投诉。 仔细选择您的伙伴！

如果您在团队中工作，您的工作必须作为一个提交文件一起上交。 如果我们分别收到小组中学生的意见书，他们将获得10％的罚款。此外，如果上述分开提交的内容不同，则以分数较低的提交为准。 作为个人工作投入的团队合作将被视为剽窃，并通过大学的官方渠道进行处理。

### 3.2 学术诚信

你们每个人都应提交自己的原创作品，或团队中项目的原创作品。 在很多情况下，向其他人（教师，TA或其他学生）寻求提示或调试帮助，或大体谈论编程策略是很有用的。 此类活动既可以接受也可以鼓励，但您必须指出收到的任何帮助（人工，在线或其他方式）。 任何未得到适当引用的帮助都将被视为窃。 无论如何，您负责编码，理解并能够自己或以团队的方式解释您提交的所有项目工作。

我们将积极处理所有涉嫌剽窃的案件，并将通过大学的官方渠道进行处理。 如果您对此政策或我们对违反学术诚信的程度有任何疑问，请立即与课程讲师讨论您的疑虑。

## 4. 项目选择

### 4.1 股票推荐系统

#### 动机

* 帮助股票购买者做出更加明智的选择
* 寻找那些非常善于从股票中获利的人（专家）
* 使用基于用户的协同推荐来发现买家和专家之间的相似性
* 向股票购买者推荐最相似专家的一些股票
* 减轻股票购买者的负担，避免他们浏览成千上万的股票并作出明智的选择

#### 数据集

数据集来自纳斯达克证券交易所数据

* YaHoo Finance历史数据集
* 100,000个用户的模拟交易记录

#### 算法

基于用户的带推断标签的协同过滤

#### 工具

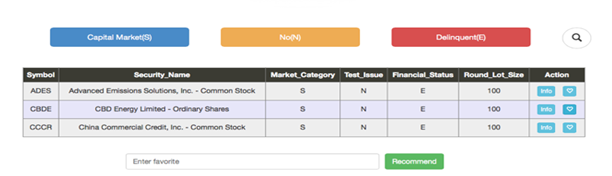
* Eclipse
* J2EE
* Mahout
* Maven
* MySQL
* PHP server

#### 过程

* 实现UI和数据库
* 确定算法-具有推断标记等级的基于用户的协同过滤。

#### 预期效果

* 通过推断的标签等级实现改进的协作过滤算法
* 根据专家的购买选择分析为客户提供合理的库存建议



### 4.2 电影探索

#### 动机

受众群体偏好

我喜欢喜剧；我更喜欢大片；我喜欢小罗伯特·唐尼；  
我想和女儿一起看电影； 我……

#### 生产者说明

* 客户想看什么样的电影？
* 哪个演员最适合这个角色？
* 他们有什么特点可以拍一部令人难忘的电影？

#### 数据集

在线搜索

#### 工具

灵活选用即可

#### 算法

* Pig：处理用户个人资料

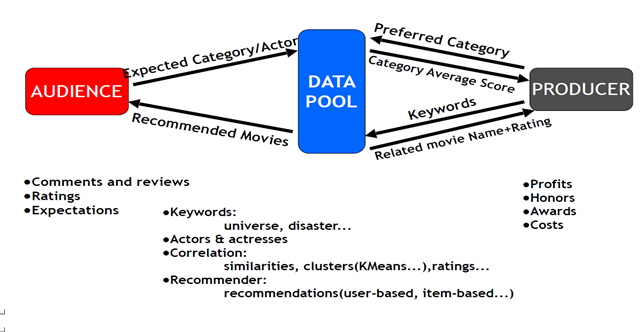
聚类和分类：

* 在受众之间建立关联

推荐：

* 根据关键字，提取内容和偏好进行预测

#### 预期效果：



### 4.3 篮球——取胜策略

#### 动机

体育在美国人民中起着重要作用。结果，人们将更多时间花在观看NBA，MLB，NFL等比赛上。为他们的休闲时间。在过去的五年中，ESPN和Yahoo发布了所有职业体育联盟的在线体育游戏，人们实际上有机会在网上组建模拟团队并与其他游戏玩家竞争。在线团队的排名基于他们为团队选择的球员的真实行为。因此，赛前草案成为比赛胜利的关键因素。我们希望提供一种制胜法则。此外，我们希望为真正的NBA球队选择球员提供一种实用的动态策略。

#### 目标

* 每轮推荐一名在线玩家
* 将所有球员分为得分、篮板等不同功能的不同组
* 对每个分组组内的球员进行排名
* 基于在线玩家的状态提出建议

#### 数据集

NBA球员统计信息

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

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、推荐

#### 工具

* Mahout
* Matlab
* 任何课程中提供的资源

#### 过程

* Fantasy Basketballl积分体系的研究
* 搜索所有可能对我们进行推荐有用的数据集
* 熟悉所有我们可能使用的工具和算法

#### 预期效果

* 建立游戏动态推荐系统
* 帮助玩家为游戏建立一支平衡的冠军团队

### 4.4 如何给新生儿取名

#### 动机

在这个项目中，可以（相对）简单但在回答现实生活中非常重要的事情上工作。

您应该给新生婴儿起什么名字？我会查看历史上流行的男/女婴儿名字，并使用各种建议书以及分类和/或聚类技术来帮助准父母选择正确的名字，这样他们的孩子就不会抱怨自己的名字一点都不酷。

#### 数据集

流行的男/女婴儿名字可以追溯到19世纪的社会保障局（Social Security Administration）

#### 算法

基于用户和基于项目的推荐、聚类和分类

#### 工具

* R lang
* HDFS/Hive/PIG/HBase
* Mahout

#### 预期效果

为用户建立动态的“名称”推荐系统

### 4.5 汇率查询和分析

#### 动机

随着国际贸易和贸易变得越来越重要，不仅特定行业群体而且我们所有人每天都需要获得更新后的汇率的必要性。  
希望获得由我们的项目提供的汇率信息的便利性更高，人们拥有更多的相关知识，可以对货币市场做出更好的决策。

#### 数据集

Bloomberg的即时汇率

来自http://www.livecharts.co.uk/historicaldata.php的历史汇率

#### 算法

我们将在发达市场中预测某些目标国家/地区的货币对美元的汇率，并将可伸缩模型应用于实时预测。 而且我们想使用RMSE来衡量我们预测的可靠性和准确性。 此外，通过显示统计显着性（例如P值），将证明结果的可重复性。

#### 工具

* Eclipse
* Tomcat
* Apache

#### 预期效果

我们的项目最初旨在为用户提供以下内容：

* 大多数世界货币的远期和交叉汇率
* 即时和历史数据
* 汇率的基本分析，包括回归和K均值
* 有关汇率的最新消息

影响汇率的关键因素

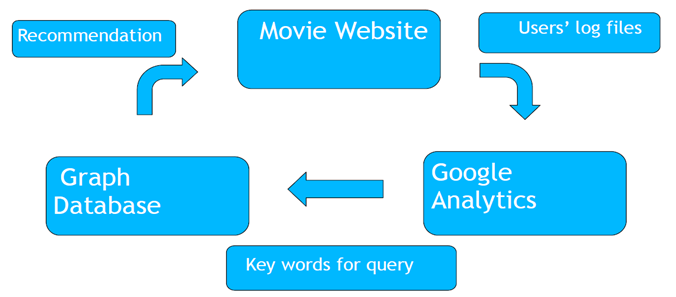
* PPP（购买力平价）
* INT（利率差）—如Libor
* GDP（GDP增长率之差）
* IGP（收入增长率）
* 相对经济实力-我们可能会使用GDP和IGP等因素对其进行定量测量

### 4.6 基于Google Analytics和图的在线电影推荐系统

#### 动机

建立我们自己的电影网站并实施电影推荐功能

#### 过程流



#### 数据集

MovieLens 1M 数据集，其中包含6000个用户对4000部电影的100万个评价

#### 算法

各种协同过滤算法，例如基于用户的推荐，基于项目的推荐等

#### 工具

* 网页设计——Dreamweaver/CoffeeCup
* 用户日志文件分析——Google Analytics
* 图形数据库——Gremlin/Neo4j
* 其他——Mahout/Eclipse

#### 时间线

现在——11/27/14：电影网站的设计和发布

11/28——12/04/14：在网站中加入Google Analytics并分析用户日志文件

12/05——12/11/14：生成图形数据库查询并更新网站上推荐的电影

#### 预期效果

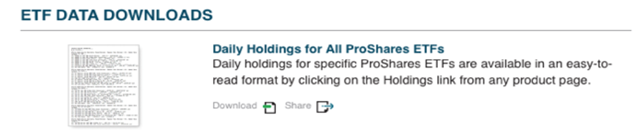
* 建立我们自己的在线电影网站
* 实现电影推荐功能
* 使用Google Analytics进行网站分析

### 4.7 使用Hadoop、MapReduce的股票预测

#### 动机

* 股票市场具有高利润，高风险的特点，对股票市场的分析预测研究已引起人们的重视。股票价格趋势是复杂的非线性函数，因此价格具有一定的可预测性
* Hadoop MapReduce是专门设计用于处理分布式源上的大型数据集的最新框架。Apache的Hadoop是MapReduce的实现

#### 数据集



#### 算法

* Pearson Correlation Similarity
* Euclidean Distance Similarity
* Stochastic Gradient Descent(SDG)

#### 工具

* Hadoop
* Mahout
* Hbase

#### 预期效果

我们将分析名为“Daily Holdings for All ProShares ETFs”的数据集，其中包含从证券交易所市场收集的大量信息。  
第一步是检查数据并提供可能增加的库存。 使用这些筛选出的库存，向某个用户建议她/他可能感兴趣的潜在库存。

### 4.8 客户投诉分析：对困扰银行业的问题的见解

#### 动机

* 银行面临的最大挑战之一是最大程度地降低客户流失率，这直接取决于客户满意度。
* 客户倾向于选择值得信赖的银行。
* 由于缺乏可扩展的解决方案，银行根据数据子集做出决策

**在此项目中，应使用可扩展的设计来解决上述问题！**

#### 数据集

消费者投诉数据库：The dataset contains

向银行和金融机构发出的零售消费者投诉（由消费者金融保护局提供）

http://www.consumerfinance.gov/complaintdatabase/

#### 算法

各种聚类和分类算法

#### 工具

* Hadoop
* Mahout
* Java
* Python

#### 预期效果

* 各州的主要零售银行业务问题，并根据地理或社会经济等级进行匹配分析。
* 各州消费者最关注的问题。
* 客户满意度或对机构投诉的不满对业务产生的影响。
* 针对类似性质的未来投诉提出可能被视为“第一反应”的解决方案。
* 假设绩效指标适用于所有投诉–可用于根据解决时间对投诉进行优先排序