# 《大数据分析》课程大纲

**一、课程名称：**大数据分析

**二、课程性质：**选修、理论课

**三、学时与学分：**40 学时，2.5 学分

**四、课程先导课：**高级语言程序设计，大数据导论、Python 语言等

**五、课程简介**

针对大数据集合的 4v 特性，了解如何将数据挖掘原理应用与解剖大型复杂数据集，包括非常大型数据库中的数据集，或通过数据解析来进行挖掘，学习探索和分析的数据模式，了解将数据转化为有价值的可用信息的大数据分析方式和工具。课程将系统地介绍大数据分析的理论、算法，包括初级数据挖掘和高级关系挖掘、协同滤波等经典大数据分析算法，同时就文本大数据分析、知识计算、网络数据挖掘、社会媒体分析等内容进行应用方面的简述。

**六、课程目标**

通过相关教学活动，让学生接触并理解大数据分析的工作原理，掌握常见的大数据分析方法，使学生具有 Python 大数据分析和开发的能力。提升学生数据分析的能力。

课程的具体目标包括：

**目标 1：**熟悉基本的 map-reduce 处理思想，掌握 shuffle 和 combine 过程的原理与意义，以培养学生对大数据问题的基本思考模式。

**目标 2：**了解 PageRank 问题的背景与应用场景，掌握概率转移矩阵的迭代运算方法，了解设置阻尼系数和归一化过程的意义。

**目标 3：**熟悉频繁项集、支持度、关联规则和置信度的概念，掌握 Apriori 和 pcy 算法原理。

**目标 4：**掌握 kmeans 算法核心要点，掌握两种基本的评价指标并且能够可视化两个维度下的聚类效果图。

**目标 5：**了解协同过滤(CF)与基于内容推荐(CB)两种推荐算法的基本思想与应用场景，掌握这两种推荐算法的实现。在此基础上，进一步要求学生掌握 MinHash 算法的基本原理，要求能够运用该算法对效用矩阵进行降维处理。

## 七、课程目标对毕业要求的支撑关系

| 支撑的毕业要求二级指标点 | 对应课程目标 |
| --- | --- |
| **1.3** 能将软硬件知识、相关工程知识和模型方法用于推演和分析计算机复杂工程问题 | 目标 2、3、4、5 |
| **1.4** 能将软硬件知识、相关工程知识和模型方法用于计算机复杂工程问题解决方案进行比较和综合 | 目标 1 |
| **2.1** 能综合运用数学、自然科学、工程科学以及计算机科学的基本原理，识别、判断和表达计算机复杂工程问题的关键环节 | 目标 2、3、4、5 |
| **3.1** 掌握与计算机复杂工程问题有关的工程设计和软硬件产品开发全周期、全流程的基本设计/开发方法和技术，了解影响设计目标和技术方案的多种因素 | 目标 2、3、4、5 |

## 八、教学设计及对课程目标的支持

### 第一章　大数据分析系统与平台

**1.教学目标**

1）了解国际、国内大数据分析平台与工具；

2）理解 MapReduce 编程模型及其核心思想；

3）编写 MapReduce 程序实例--词频统计(Word Count)；

4）掌握 MapReduce 分布式计算框架的基本组成及各部分的主要功能；

5）熟练掌握 MapReduce 框架下常用编程组件与功能模块的使用及实现；

6）熟悉与了解常见的大数据分析系统，包括 Hadoop MapRedcue、Spark 等对大数据分析计算的性能评价指标，理解不同数据分析系统的特点及局限性；

**2.教学重点**

1）MapReduce 编程模型

MapReduce 是一种编程模型，用于大规模数据集（大于 1TB）的并行运算。概念"Map（映射）"和"Reduce（归约）"，是它们的主要思想，都是从函数式编程语言里借来的，还有从矢量编程语言里借来的特性。它极大地方便了编程人员将自己的程序运行在分布式系统上。通过指定一个 Map（映射）函数，用来把一组键值对映射成一组新的键值对，指定并发的 Reduce（归约）函数，用来保证所有映射的键值对中的每一个共享相同的键组。

2）MapReduce 的工作过程

从大数据分析课程的实践需求出发，要求学生能深刻理解 MapReduce 先分后总的分布式数据分析的内涵，明确大数据分析课程要解决什么问题，利用 MapReduce 工具可以解决什么问题。要求学生理解从输入文件分片由 MapTask 处理到 Reduce 端合并输出的分布式处理分析过程。

3）MapTask 工作原理

要求学生能理解 MapTask 的工作机制。具体来说，理解 MapTask 处理流程划分的五个阶段以及每个阶段的具体任务。

4）ReduceTask 工作原理

要求学生能理解 ReduceTask 的工作机制。具体来说，理解 ReduceTask 处理流程划分的三个阶段以及每个阶段的具体任务。

5）Shuffle 工作原理

Shuffle 过程包含在 Map 和 Reduce 两端，即 Map shuffle 和 Reduce shuffl。要求学生掌握 Map 端的 Shuffle 过程包括：对 Map 的结果进行分区、排序、分割，然后将属于同一划分（分区）的输出合并在一起并写在磁盘上，最终得到一个分区有序的文件，分区有序的含义是 map 输出的键值对按分区进行排列，具有相同 partition 值的键值对存储在一起，每个分区里面的键值对又按 key 值进行升序排列。要求学生掌握 Reduce 端的 Shuffle 过程主要包括：复制 Map 输出、排序合并两个阶段。

6）MapReduce 性能优化策略

进行大数据运算，当数据量极其大时，那么对 MapReduce 性能的优化重要性不言而喻，尤其是 Shuffle 过程中的参数配置对作业的总执行时间影响特别大。要求学生了解下列与 MapReduce 相关的性能优化策略和方法，主要包括四个方面：数据输入、Map 阶段、Reduce 阶段、Shuffle 阶段。

**3.教学难点**

1）MapReduce 工作过程

从大数据分析课程的实践需求出发，要求学生能深刻理解 MapReduce 先分后总的分布式数据分析的内涵，明确大数据分析课程要解决什么问题，利用 MapReduce 工具可以解决什么问题。要求学生深入了解从输入文件分片由 MapTask 处理到 Reduce 端合并输出的整个分布式处理分析过程。

**4.教学环节设计**

围绕教学重点和教学难点，综合应用课堂讲授与讨论、编成实践与作业、课外阅读等教学形式。

1）讨论

围绕不同大数据处理系统的性能评价指标及其内涵和局限性等问题展开讨论。

2）编程实践与作业

围绕 MapReduce 的编程运行及优化，要求学生参与一系列经典案例的编程实践，包括：词频统计、倒排索引、数据去重和 TopN 等。

3）课外阅读

推荐学生阅读关于国际大数据分析技术相关的最新资料、报道。

**第二章　链接分析（Link Analysis）与 PageRank**

本章的主要知识点包括 PageRank、随机游走、特定主题游走、图的相似性、Link Spam、TrustRank、HITS 等。PageRank 是谷歌的镇店之宝，一种用来对网络中节点的重要性排序的算法。这个算法最初是用来对网页重要性进行排序。人们对 PageRank 进行个各种改动，基于相关算法在推荐、社会网络分析、自然语言处理等领域推出了很多实用的解决方案。本章内容要求学生围绕着两个基本问题展开学习，即 PageRank 算法是怎么来的呢？怎么计算？

**1.教学目标**

1）了解 PageRank 基本概念,根据网页重要性进行页面排名；

2）掌握基本的 PageRank 算法；

3）理解 PageRank 的矩阵表达；

4）理解并掌握 PageRank 的矩阵表达实例；

5）深刻理解密迭代方法;

6）从马尔可夫角度看待 PageRank;

7）理解进阶版 PageRank 的必要性；

9）理解 Teleport 实列；

10）PageRank 算法效率分析；

11)理解和掌握完整版 PageRank 算法；

12）基于特定主题的 PageRank;

**2.教学重点**

1）PageRank 的基本算法

要求学生理解 PageRank 算法利用网络的图结构来评价网页的重要性，这里的图结构是指指向网页的链接，也就是 Inlink。要求理解 PageRank 算法的两种假设：数量假设（指向该网站的数量越多，重要性越高）和质量假设（指向该网站越权威，重要性越高）；并由此为基准理解 PageRank 基本算法思想及定义。

2）PageRank 的矩阵表达

理解 Column stochastic (列随机) 矩阵，即矩阵 $M$(每一个列上的元素之和为 1，符合上面约束条件还有行随机矩阵和双随机矩阵)，我们假设网页 $j$ 有 $d_j$ 个外链

接，第 *j* 个外链接指向第 *i* 个网页，即可倒出 PageRank 的矩阵表达式。要求学生理解矩阵特征值、特征向量的概念。

3）理解和掌握密迭代计算方法

根据 PageRank 矩阵表达的分析，要找到重要度的向量，就是要求转移矩阵 M 的特征值为 1 的特征值向量。求这个特征向量的方法就是 Power Iteration Method，也就是是求绝对值最大的特征值向量的方法。要求学生理解和掌握密迭代方法的原理和计算。

4）理解什么是 Teleport 及进阶版 PageRank

要求学生理解基本版 PageRank 无法处理的两种情况：“网页只有入度没有出度（Dead End）”以及“网页即使有出度也是指向其本身（Spider Traps）”。Dead End 是一个严重问题；第二种情况 Spider Traps 不会对收敛性产生影响，但收敛到的 PageRank 不是理想目标。以上两种情况的解决方法就是“teleport”（随机跳转）。

5）特定主题游走

了解原始的 PageRank 算法只能提供通用的 importance score 这一局限性。

优化目标：不只是根据 importance score 来评估网页，而是加上该网页离某个主题的距离，例如运动、娱乐、历史等。就是要加入各种权重来重新计算最终的 PageRank 算法计算结果。

**3.教学难点**

1）理解和掌握密迭代计算方法

要求学生理解并熟练掌握 PageRank 矩阵表达的分析，理解要找到重要度的向量，就是要求转移矩阵 M 的特征值为 1 的特征值向量。求这个特征向量的方法就是 Power Iteration Method，也就是是求绝对值最大的特征值向量的方法。

2）深刻理解掌握特定主题游走

深刻理解原始的 PageRank 算法只能提供通用的 importance score 这一局限性。根据不同具体需求设立优化目标，不只是根据单一的 importance score 来评估网页，而是加上该网页离某个主题的距离，例如运动、娱乐、历史等。就是要加入各种权重来重新计算最终的 PageRank 算法计算结果。

**4.教学环节设计**

围绕教学重点和教学难点，综合应用课堂讲授与讨论、编成实践与作业、课外阅读等教学形式。

1）讨论

围绕基于不同特定主题游走算法的性能评价指标及其内涵和局限性等问题展开讨论。

2）编程实践与作业

围绕 PageRank 的编程运行及各个进阶版本，要求学生参与一系列经典案例的编程实践。

3）课外阅读

推荐学生阅读关于链路分析、PageRank、TrustRank 等最近技术的相关资料、报道。

## 第三章　推荐系统

### 1. 教学目标

1）了解不同推荐算法的特点与应用背景；

2）掌握效用矩阵的概念与原理；

3）熟练掌握基于内容(Content-Based)推荐算法的原理及实现；

4）熟练掌握协同过滤(Collaborative Filtering) 推荐算法的原理及实现；

5）了解隐语义模型(Latent Factor Models)推荐算法的原理；

6）分析比对不同推荐算法的优劣势与应用场景。

### 2. 教学重点

1）效用矩阵的概念

要求学生掌握效用矩阵的基本概念与计算方法以及在推荐算法中的地位与作用，包括效用的定义，不同推荐算法中效用矩阵的计算方法，在不同视角下效用矩阵的不同表现形式（例如在协同过滤算法中，效用矩阵可以表现为 user-user 和 item-item 两种形式）等。

2）基于内容推荐算法的原理

要求学生掌握基于内容推荐算法的基本概念与实现原理。包括“基于内容”的意义，TF-IDF 的概念与计算方法，余弦相似度的原理与计算方法以及如何根据相似度进行评分和推荐等。

3）协同过滤推荐算法的原理

要求学生掌握协同过滤推荐算法的基本概念与实现原理。包括“协同”的意义，对 Jaccard 相似度的改进方法以及这么做的原因，user-user 和 item-item 两种过滤方法的异同和效果等。

4）不同推荐算法的比对分析

要求学生根据课堂讲述内容以及实验结果对不同的推荐算法进行分析比对，了解不同推荐算法在不同应用场景下的效果，分析并总结不同推荐算法的特点和适用场景。

**3. 教学难点**

1）TF-IDF 概念的理解

要求学生深刻理解 TF-IDF 的概念和计算方法。TF-IDF 是 Term Frequency - Inverse Document Frequency 的缩写，即“词频-逆文本频率”，由 TF 和 IDF 两部分组成。TF 即词频，是文本中各个词的出现频率统计，并作为文本特征； IDF 即“逆文本频率”，用于反映词的重要性，进而修正词特征值。

2）相似度计算

要求学生熟练掌握各种相似度的特点与计算方法，包括余弦相似度的原理与计算方法，Jaccard 相似度的原理与计算方法，对 Jaccard 相似度的改进手段理解，偏置值的计算方法与意义等。

**4. 教学环节设计**

围绕教学重点和教学难点，综合应用课堂讲授与讨论、编成实践与作业、课外阅读等教学形式。

1）讨论

围绕不同推荐算法的优劣与适用场景展开，鼓励学生根据自身见解与实际体验发表看法。

2）随堂测试

通过基础的计算测试题，考察学生对相似度概念的理解程度以及对计算方法的掌握程度。

3）编程作业

要求学生根据课堂讲述内容亲手实现推荐算法，提高学生的动手能力并强化对相关概念的理解。

4）课外阅读

推荐学生阅读关于推荐算法的最新技术资料和报道。

## 第四章　维度约减

**1.教学目标**

1）了解特征值和特征向量的定义及计算；

2）了解特征值在主成分分析中的应用；

3）掌握奇异值分解（SVD）的矩阵分析方法及实现算法；

4）了解奇异值分解算法在稀疏矩阵中如何运用；

5）掌握 CUR 分解算法的原理及实现。

本章教学支持课程目标 1。

**2.教学重点**

1）SVD 分解算法的原理

理解 SVD 中对高维矩阵进行低维表示的矩阵分析手段，可对任意矩阵进行精确表示，并去除表示中的非重要部分来获得任意维度的近似表示。和特征分解不同，SVD 奇异值分解能够用于任意 m*n 矩阵，并不要求要分解的矩阵为方阵。

2）SVD 的缺陷

首先是可解释性较差：对于 SVD 分解通常的理解应该是，左奇异向量以及右奇异向量分别张成了原始矩阵所在的列空间以及行空间，但是对于原始矩阵而言，并没有较强的可解释性。其次是太过密集：就算原始矩阵是一个稀疏矩阵，该矩阵所分解而成的 U、V 矩阵仍然是高度密集的，这在某些应用场景下是难以接受的。

3）SVD 的应用

a. 奇异值分解可以被用来计算矩阵的广义逆阵（伪逆）。

b. 奇异值分解的另一个应用是给出矩阵的列空间、零空间和秩的表示。对角矩阵的非零对角元素的个数对应于矩阵的秩。与零奇异值对应的右奇异向量生成矩阵的零空间，与非零奇异值对应的左奇异向量则生成矩阵的列空间。在线性代数数值计算中奇异值分解一般用于确定矩阵的有效秩。

c. 奇异值分解在统计中的主要应用为主成分分析（PCA）。数据集的特征值（在 SVD 中用奇异值表征）按照重要性排列，降维的过程就是舍弃不重要的特征向量的过程，而剩下的特征向量张成空间为降维后的空间。

4）CUR 分解算法的原理

针对 SVD 分解的缺陷，CUR 分解的可解释性不仅更强，也更适用于稀疏矩阵的情况。具体来讲，对于一个秩 k 矩阵 A ，如果选择其中 k 列张成矩阵 A 的列空间，选择其中 k 行张成矩阵 A 的行空间，那么也应该能够通过组合这些线性映射来恢复原矩阵。

5）CUR 的应用

与传统的主成分分析的矩阵分解方法相比较，在特征选择方面，CUR 分解方法不仅具有很高的准确度，而且还具有很好的可解释性；在矩阵恢复方面，CUR 矩阵分解方法具有很高的稳定性同时还具有很高的准确度

**3.教学难点**

1）SVD 分解在某些场景下可解释性的问题

SVD 分解得到的结果与原始数据直观的关系就不明确，只是原始数据的线性组合表达，可解释性并不强。因此在某些场景下并不适用。

2）CUR 分解中如何选择 C 和 R

CUR 矩阵分解类似于行选择问题，关于怎么选择行，主要是两种方案：1.随机采样 ；2.自适应采样。适应采样中比较简单的就是根据 1 范数或者 2 范数来做，稍微复杂一点的就比如是杠杆值采样等等。

**4.教学环节设计**

围绕教学重点和教学难点，综合应用课堂讲授与讨论、课外阅读等教学形式。

1）讨论

围绕高维数据的不同降维手段，如何提高降维效率等问题展开。

2）作业

围绕 SVD 分解和 CUR 分解算法等内容布置。

3）课外阅读

阅读有关最新的维度约减的文献。

## 第五章　关联关系挖掘

**1.教学目标**

1）了解频繁项集的定义；

2）掌握关联规则的工作原理、形式；

3）掌握 A-Priori 算法的原理、实现；

4）了解大数据集在内存中处理的困难，掌握 PCY 算法的原理及实现；

5）了解 PCY 算法的一系列改进算法，包括多阶段算法，多哈希算法。

本章教学支持课程目标 1。

**2.教学重点**

1）A-Priori 算法的原理

理解 A-Priori 算法进行多遍扫描的原因，理解完成频繁项发现需要用到的表，掌握算法执行过程。

2）PCY 算法的原理

PCY 算法改进 A-Priori 算法中第一遍扫描过程中大量未用内存空间，理解哈希表在 PCY 算法中的作用，掌握算法执行过程。

**3.教学难点**

1）频繁项发现中最大内存过程

理解频繁项发现方法中最大消耗内存的步骤。

**4.教学环节设计**

围绕教学重点和教学难点，综合应用课堂讲授与讨论、课外阅读等教学形式。

1）讨论

围绕大数据下频繁项发现对内存的消耗现状、多阶段执行等问题展开。

2）作业

围绕频繁项发现等内容布置。

3）课外阅读

阅读有关最新的频繁项发现的文献。

## 第六章　大数据聚类

**1.教学目标**

1）理解点、空间、距离、聚类的定义；

2）掌握层次聚类的基本原理，算法实现，效率分析；

3）掌握 K-means 聚类算法的基本原理，初始化的方法，k 值的选择方法，算法实现；

4）掌握 CURE 算法的基本原理。

本章教学支持的课程目标为目标 2。

**2.教学重点**

1）K-mean 算法及其实现

掌握簇质心和簇中心点的区别和联系，明确簇指标，掌握 K-means 聚类的步骤，理解初始化对最终结果的影响，掌握 K 值的确定的多种方法。

2）CURE 算法及其实现

理解 K-mean 算法的局限性，掌握 CURE 算法对簇形状的要求，CURE 算法的步骤，簇合并的时机的确定。

**3.教学难点**

1）K-mean 算法中 K 值的确定

理解 K 值的不同对 K-means 算法最终结果的影响，掌握确定最优 K 值的方法。

**4.教学环节设计**

围绕教学重点和教学难点，综合应用课堂讨论、课外阅读等教学形式。

1）课堂讨论

围绕不同聚类方法的局限性展开。

2）课外阅读

非欧空间下聚类方法等相关内容。

## 九、教与学

### 1.教学方法

主要的教学环节包括课堂授课、研讨、课后作业等环节。本课程的教学设计特色主要体现在如下两个方面：

1）基于问题的教学方法。将围绕课程教学的重、难点，精心设计若干探究性问题，引导同学深入思考，加深所学重、难点知识的理解和应用。

3）强调动手实践。该课程的教学与独立设置的课程实验相配合，实验内容与理论课程教学进度同步，通过实验加深对所学理论知识的理解，提升学生应用理论知识解决复杂问题的能力，通过实验也可以检验理论课程的学习效果。

### 2.学习方法

“大数据分析”是一门理论性、技术性和实践性都很强的专业选修课程，学习过程中，首先要注重对课程基本理论的钻研，要引导学生积极参与课堂讨论、深刻理解原理和技术本质；其次，要站在系列课程的角度学习，本课程的学习需要Python 语言，大数据导论等前导课的知识和技术支撑；第三，独立完成课程配套开设的独立实验，通过实验，加强对课程理论知识的理解，同时，训练学生发现问题、分析问题和解决问题的能力。

## 十、学时分配

| 序号 | 主要内容 | 学时分配 |
|---|---|---|
| 1 | 第 0 章 大数据分析概论 | 1 |
| 2 | 第一章 Mapreduce 简介 | 2 |
| 3 | 第二章 图数据分析 | 5 |
| 4 | 第三章 推荐系统 | 5 |
| 5 | 第四章 维度约减 | 3 |
| 6 | 第五章 关联关系挖掘 | 4 |
| 7 | 第六章 大数据聚类 | 4 |
| 8 | 课程实验 | 16 |
| 总计 | | 40 |

## 十一、课程考核与成绩评定

### 1.课程成绩构成

课程最终成绩由考勤与作业成绩、实验成绩和课程期末实验报告成绩综合而成，各部分成绩的比例如下：

**1）考勤与作业成绩：20%**。考勤主要检查学生遵守学习相关规定, 按时参加课堂学习的情况。作业将引导学生复习和巩固讲授的内容（基本理论、基本方法、基本理论分析与计算、课外阅读报告等），主要考查作业完成率和质量。

**2）实验成绩：30%**。 主要考察学生在实验课上代码完成率和质量。

**3）期末实验报告成绩：50%**。主要考核学生根据实验代码的项目对应的项目报告的质量。

课程考核成绩评定如表 1 所示。

表 1 大数据分析课程考核与成绩评定

| 课程目标 | 考核与评价方式及成绩比例（约） | | |
|---|---|---|---|
| | 考勤与作业 | 实验成绩 | 期末实验报告 |
| 1 | 4 | 5 | 5 |
| 2 | 4 | 5 | 5 |
| 3 | 4 | 5 | 5 |
| 4 | 4 | 5 | 5 |
| 5 | 4 | 10 | 30 |

2.考核与评价标准

1）作业成绩考核与评价标准

表 2 大数据分析作业考核与成绩评定

| 评价标准 | | | |
|---|---|---|---|
| 优秀 | 良好 | 中-及格 | 不及格 |
| 按时提交作业，概念准确，计算结果正确，分析充分，论述清晰，层次分明。 | 按时提交作业，概念准确，存在少量错误，分析较充分，论述清晰，层次分明。 | 按时提交作业，概念基本准确，计算结果存在一些错误，论述基本清晰。 | 未按时交作业，概念欠准确，计算结果错误较多。 |

2）实验成绩

按实验课考核标准评分。

3）课程考核与成绩评定

根据期末实验报告评分标准进行评定。

大数据课程组

2020 年 6 月制定

2021 年 04 月修订

2022 年 04 月修订