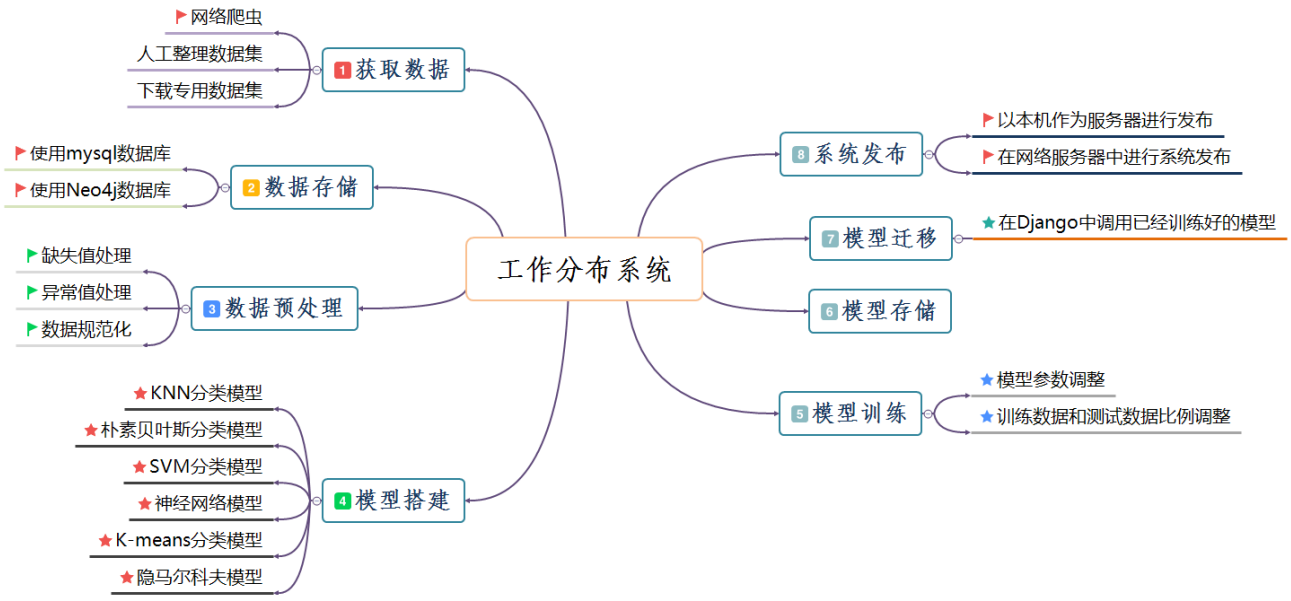
# 大数据技术与应用课程说明

<table>
<tr><td colspan="3">大数据技术开发与应用</td></tr>
<tr><td>课程特点</td><td colspan="2">多技术融合的交叉课程</td></tr>
<tr><td>涉及知识</td><td colspan="2">数据挖掘、网络爬虫、机器学习、Hadoop、数学的相关知识（高等数学、线性代数、概率论）、Python 语言、数据库知识</td></tr>
<tr><td rowspan="4">所学内容</td><td>1、数据采集、存储以及预处理</td><td>数据采集方式：<br>1、网络爬虫；（基本爬虫、反爬虫处理、spider 框架、scrapy 框架）<br>2、人工总结；<br>3、下载已经发布的数据包。<br>数据存储：<br>1、采用 mysql 数据库进行存储（基本操作以及使用 sql 语句进行简单的数据预处理）<br>2、采用 Neo4j 数据库进行存储（如何使用 Cypher 进行数据库的简单操作）<br>数据预处理：<br>1、 数据清洗（缺失值发现与处理、异常值发现与处理）<br>2、 数据变换（规范化、离散化、属性构造）<br>3、 数据规约（属性规约、数值规约）</td></tr>
<tr><td>2、数据分析与挖掘</td><td>1、通过机器学习算法对数据进行分析（knn、朴素贝叶斯、svm、decision tree、人工神经网络、pca 降维、隐马尔科夫模型）</td></tr>
<tr><td>3、数据可视化</td><td>主要是通过 matplotlib 来进行数据的可视化（修改默认样式、创建多维立体图像、创建动态图形）</td></tr>
<tr><td colspan="2">4、使用 hadoop 的分布式并行架构，用于存储海量数据</td></tr>
<tr><td>课程项目</td><td colspan="2">1、 文本分类系统<br>2、 工作分布系统<br>3、 学生自主项目</td></tr>
<tr><td>课程目标</td><td colspan="2">1、 使学生能够掌握独立开发一套完整系统的能力；<br>2、 使学生能够熟练地掌握网络爬虫；<br>3、 使学生能够掌握数据预处理的基本方法；<br>4、 使学生能够将收集到的数据进行可视化展示；<br>5、 使学生能够了解一些机器学习以及数据挖掘的相关概念，并能够掌握机器学习以及数据挖掘的基本算法原理以及实现与应用；<br>6、 使学生能够了解关于 Hadoop 的相关知识。</td></tr>
</table>

## 1、系统思维导图




## 2、学生自主项目

本部分的主要目的是提高学生实践能力以及团队协作能力，具体规划是对上课的学生进行分组。以组为单位进行整体项目的开发，以本课程所讲述的知识点为依据，自主选择自己感兴趣的项目以此项目来作为本课程的结课考核。

## 3、课程总体安排

| 工作分布系统 | |
|---|---|
| 周数 | 上课内容 |
| 第 1 周 | Python 的相关知识点复习 |
| 第 2-4 周 | 网络爬虫 |
| 第 5 周 | 数据库的操作以及数据预处理 |
| 第 6 周 | 几种分类模型的原理与实现 |
| 第 7 周 | Django 网站框架的搭建与实现（flask） |
| 第 8 周 | 在 windows 系统中发布系统 |

| 学生自主项目 | |
|---|---|
| 周数 | 内容 |
| 第 1 周 | 确定学生分组名单以及确定项目题目 |
| 第 2-4 周 | 开始完成自主项目前期工作 |
| 第 5 周 | 完成中期汇报 |
| 第 6-8 周 | 项目成型 |

4、课程考核方式：平时 10%　　课堂项目 40%　　自主项目 50%