学习任务一　数据采集与清洗

学习目标

1.根据工作任务书，填写工作联系单。

2.能根据相关技术要求及规范，制定数据采集方案，撰写采集步骤表格。

3.能正确完成基本的计算机操作及系统语音、图像操作。

4.能根据方案的具体要求，在规定时间内完成图像数据采集、图像数据清洗、语音数据采集、语音数据清洗及交付验收等操作，确保图像及语音数据集的正确采集。

5.能分析工作中存在问题的原因，并提出解决方案。

6.能对学习与工作中出现的问题进行总结反思，并与他人开展良好合作，进行有效沟通。

建议课时

40 学时。

工作情景描述

学院校企合作单位 XX 公司需要进行人工智能开发，需要采集指定的图像数据及语音数据,要求我院人工智能技术应用专业教师组织同学共同完成数据采集任务。教师接受该订单，带领同学共同完成该任务，该任务要求两周内完成。

技术员(学生)接到主管(教师)下达的调试任务书后，认真阅读任务书，了解客户需求；收集相关后收集相关资料，熟悉数据采集的相关步骤，制定数据采集方案，并对方案进行可行性分析；而后按照方案，完成图像数据采集、图像数据清洗、语音数据采集、语音数据清洗等任务，并进行测试，出具验收报告；按照任务要求向主管(教师)交付验收并进行总结。

工作流程与活动

1. 明确任务（6 学时）
2. 数据采集方案编制（8 学时）
3. 图像数据采集与清洗（10 学时）
4. 语音数据采集与清洗（10 学时）
5. 验收与总结（6 学时）

学习活动 1　　明确任务

学习目标

1. 根据工作任务书，填写工作联系单。

2. 能口述并填写人工智能数据及数据集类型一览表。

3. 能口述并填写人工智能图像数据采集流程。

4. 能口述并填写人工智能语音数据采集流程。

4. 能查阅典型的数据清洗相关知识，分析其主要作用，填写人工智能数据清洗流程一览表。

建议课时

6 学时。

学习过程

一、明确任务内容和要求

（一）阅读任务书

独立阅读工作情境描述，用荧光笔在任务书中画出关键词，并将关键词载录如下，其中需要进一步了解的词用星号标注出来：

（二）解释术语

1. 查阅相关数字资源库或通过网络搜索，解释数据清洗的概念是什么？

（三）填写工作联系单

查阅相关数字资源库或通过网络搜索，根据实际情况填写表 1-1-1 所示的工作联系单。

表 1-1-1　工作联系单

任务名称				接单日期	
工作地点				任务周期	
工作内容					
工具、量具及设备					
工作项目					
项目负责人姓名		联系电话		验收日期	
团队负责人姓名		联系电话		团队名称	
备注					

二、数据及数据集基础认知

在人工智能领域，数据是最主要的信息载体。而数据集，则是由数据所组成的集合。不同的数据场景中，构成数据集的元素都是各不相同，比如在图像处理场景中，图像是图像处理数据集的数据元素，而在语音处理场景中，语音是语音处理数据集的数据元素，但是它们的表现形式则是可以多样化的。在人工智能应用中，数据集往往用于数据的运算，实现数据的分析、归类等，最终用于数据训练。

（一）常见的数据类型

人工智能领域的数据元素是多种多样的，请查阅相关资料，了解较为常见的人工智能数据类型，并填写表 1-1-2 人工智能常见数据类型一览表。

表 1-1-2　人工智能常见数据类型一览表

数据类型名称	数据类型简述及特点

数据类型名称	数据类型简述及特点

（二）常见的数据集

数据集是一类数据的集合，而数据集的形式可以是多样化的，请查阅相关资料，了解当前较为常见的人工智能数据集类型，并填写表 1-1-3 人工智能常见数据集类型一览表。

表 1-1-3　人工智能常见数据集类型一览表

数据集名称	功能及简述

三、数据采集基础认知

人工智能的数据采集，主要的作用任务从大自然中获取图像、声音、文字等信息媒介，通过计算机处理后作为数据进行存储，并组成数据集。不同的信息媒介，甚至同样

类型的信息媒介，数据采集的方式都是多种多样的。

（一）图像数据采集方法

人工智能能算法模型训练的时候，经常需要训练数据，特别是图像识别的算法模型进行训练的时候，更是需要大量的图片进行训练。请查阅相关资料，了解较为常见的人工智能图像数据采集的方法，并填写表 1-1-4 人工智能图像数据采集方式一览表。

表 1-1-4　人工智能图像数据采集方式一览表

采集方式名称	采集方式简述
摄像头采集	在实际场景中使用手机或电脑摄像头采集图像

（二）语音数据采集流程

在人工智能智能语音系统应用中，需要语音数据来训练算法模型。不同的语音产品需要不同程度的、量级的语音数据。语音数据又分为很多不同的类型，常见的类型有语音识别数据（ASR），和语音合成数据(TTS)。请查阅相关资料，了解人工智能语

音数据的采集流程，并补充填写表 1-1-5 人工智能语音数据的采集流程表。

表 1-1-5　人工智能语音数据的采集流程表

基本流程	流程简述
找出用户需要说的内容	从语音数据中确定需要进行语音处理的部分

四、数据清洗基础认知

在采集到的大量数据中，难免存在残缺、错误、重复的数据，或者不符合条件的数据，需要在对数据进行处理前进行数据的筛查，即人工智能的数据清洗。狭义的数据清洗是指发现并纠正数据文件中可识别的错误的最后一道程序，包括检查数据一致性，处理无效值和缺失值等。广义的数据清洗则同时包含人工筛选等对数据进行前处理的步骤。

（一）数据清洗的作用

请查阅相关资料，了解数据清洗的基础，并简述数据清洗的作用及重要性。

（二）数据清洗的流程

从数据清洗的概念就可以得知数据清洗是数据库中的“脏”数据。“脏数据”，即数据库中残缺、错误、重复的数据。数据清洗，旨在提高数据的质量、缩小数据统计过程中的误差值。请查阅相关信息，并补充填写表 1-1-6 数据清洗流程一览表

表 1-1-6　数据清洗流程一览表

基本流程	流程简述
预处理	采用人工方式先对数据进行初步筛选,将明显不符合要求的数据去除
对缺失值进行清洗	