# 第 1 章 概述

导读：大数据技术与我们日常生活密切相关。数据是大数据的前提，原始数据存在大量不完整、不一致、有异常的情况，严重影响到数据利用，甚至可能导致结果的偏差。因此，数据预处理便应运而生。本章首先做数据预处理的概述，使读者对其有个整体认识。然后介绍 Python 数据预处理的开发工具与运行环境，达到工欲善其事必先利其器的效果；最后综合中文分词的实战案例，让读者入门数据预处理。

## 1.1 Python 数据预处理

数据预处理：大数据与人工智能时代离不开海量的原始数据做支撑，这些原始数据存在大量的不完整、不一致、异常值等问题，很难得到高质量是数据建模，甚至可能导致工程应用的偏差，因此，要对原始数据做一定的处理。这种从原始数据到挖掘数据之间，对数据进行的操作叫做数据预处理。数据预处理通常包括数据清理、数据集成、数据归约、数据变换、数据降维等步骤，其目的让数据更好的适应技术或算法，挖掘其应用价值和社会价值。总结：原始数据存在不完整、偏态、噪声、特征比重、特征维度、缺失值、错误值等问题；数据预处理后的数据存在完整、正态、干净、特征比重合适、特征维度合理、无缺失值等优点。

早期互联网时代数据量较少，主要存储在数据库、文件系统等介质中。其数据分析也主要靠人工统计完成。随着计算能力和硬件设施的提升，先前的算法理论（如，神经网络等）便有了用武之地。加之网络普及化，海量数据应运而生。依旧采用人工统计方法对数据处理已不合时宜。于是，来到了大数据与人工智能的时代。而在未来的一段时间，不管是无人驾驶还是智能机器人，亦或是其他应用。主要还是在有监督学习下进行的，这里的监督学习即需要有参考意义的历史数据做基础。当然，这些数据不仅仅是数据库文件、文本文件，还包括音视频、语音、网页等各种介质的数据。这些数据存在形式多样，我们将其称之为异源数据，顾名思义指的是来自不同数据源的数据。

## 1.2 开发工具与环境

Anaconda 是一个用于科学计算的 Python 发行版，支持 Linux, Mac, Windows 系统，提供了包管理与环境管理的功能，可以很方便地解决多版本 python 并存、切换以及各种第三方包安装问题。Anaconda 利用工具/命令 conda 来进行 package 和 environment 的管理，并且已经包含了 Python 和相关的配套工具。这里先解释下 conda、anaconda 这些概念的差别。conda 可以理解为一个工具，也是一个可执行命令，其核心功能是包管理与环境管理。包管理与 pip 的使用类似，环境管理则允许用户方便地安装不同版本的 python 并可以快速切换。Anaconda 则是一个打包的集合，里面预装好了 conda、某个版本的 python、众多 packages、科学计算工具等等，所以也称为 Python 的一种发行版。其实还有 Miniconda，顾名思义，它只包含最基本的内容——python 与 conda，以及相关的必须依赖项，对于空间要求严格的用户，Miniconda 是一个不错的选择。其有以下优点：

Sublime Text 是一套跨平台的文本编辑器，支持基于 Python 的插件。Sublime Text 是专有软件，可通过包 Package 扩充本身的功能。大多数的包使用自由软件授权发布，并由社

区建设维护。Sublime Text 是由程序员 Jon Skinner 于 2008 年 1 月份所开发出来，它最初被设计为一个具有丰富扩展功能的 Vim。其具有漂亮的用户界面和强大的功能，例如代码缩略图，Python 的插件，代码段等。还可自定义键绑定，菜单和工具栏。Sublime Text 的主要功能包括：拼写检查，书签，完整的 Python API ， Goto 功能，即时项目切换，多选择，多窗口等等。Sublime Text 是一个跨平台的编辑器，同时支持 Windows、Linux、Mac OS X 等操作系统。

## 1.3 实战：第一个中文分词程序

中文分词指的是将一个汉字序列切分成一个一个单独的词。分词就是将连续的字序列按照一定的规范重新组合成词序列的过程。我们知道，在英文的行文中，单词之间是以空格作为自然分界符的，而中文只是句段能通过明显的分界符来简单划界，而词是没有一个形式上的分界符的。虽然英文也同样存在短语的划分问题，不过在词这一层上，中文比之英文要复杂得多、困难得多。

## 1.4 源码获取说明

本书的源码支持 GitHUb 下载 https://github.com/bainingchao/PyDataPreprocessing，源码下载编排如下：
PyDataPreprocessing：本书源代码的根目标
Chapter+数字：分别代表对应章节的源码
Corpus：本书所有的训练语料
Files： 所有文件文档
Packages：本书所需要下载的工具包

## 1.5 小结

源码请进【机器学习和自然语言 QQ 群：436303759】文件下载：