《大数据采集与清洗》简答题目复习

## 1.请阐述什么是数据清洗。

数据清洗就是指把“脏数据”彻底洗掉，包括检查数据一致性，处理无效值和缺失值等，从而提高数据质量。

数据清洗可以有多种表述方式，其定义依赖于具体的应用，它的定义在不同的应用领域不完全相同。例如，在数据仓库环境下，数据清洗是抽取转换装载过程的一个重要部分，在数据清洗时要充分考虑数据仓库的集成性与面向主题的需要。

## 2.请阐述数据清洗的常用方法。

处理缺失值：

1. 确定缺失值范围
2. 对于指标重要性高，缺失率也高的缺失值数据，需要和取数人员或业务人员了解，是否有其他渠道可以取到相关数据，必要时进行重新采集。
3. 对于指标重要性低，缺失率也低的缺失值数据，可只进行简单填充或不作处理；对于指标重要性低，缺失率高的缺失值数据，可备份当前数据，然后直接删掉不需要的字段。

逻辑问题数据：

去重处理。去重处理一般放在格式内容清洗之后，原因是格式内容清理之后才能总体发现重复的业务数据。

1. 离群值（异常值）处理。
2. 矛盾内容处理。

## 3.请阐述数据质量的含义。

大数据应用必须建立在质量可靠的数据之上才有意义，建立在低质量甚至错误数据之上的应用有可能与其初心背道而驰。因此，数据质量正是企业应用数据的瓶颈，高质量的数据可以决定数据应用的上限，而低质量的数据则必然拉低数据应用的下限。

## 4.请阐述数据清洗的常用工具有哪些。

Excel、Kettle、DataCleaner、OpenRefine、Python等

## 5.请阐述什么是文件格式。

文件格式是指在计算机中为了存储信息而使用的对信息的特殊编码方式，是用于识别内部储存的资料。如文本文件、视频文件、图像文件等。

## 6.大数据中常见的文本文件格式有哪些？

txt格式、doc格式、xls格式、pdf格式、XML格式、JSON格式、HTML格式、tar格式、dmg格式、py格式

## 7.请阐述什么是数据抽取。

数据抽取指把数据从数据源读出来，一般用于从源文件和源数据库中获取相关的数据，也可以从Web数据库中获取相关数据。

## 8.请阐述什么是Web数据抽取。

就是根据用户或公司的需要，从一个或多个网站提取各种数据集

## 9.简述如何使用Kettle进行Web数据抽取。

在Kettle中要实现Web数据抽取可以在“查询”选项中利用“HTTP client”、“HTTP post”、“REST client”以及“Web 服务查

询”等多个组件来完成。

## 10.请阐述什么是网络爬虫？

网络爬虫又称之为网络机器人、网络蜘蛛，是一种通过既定规则，能够自动提取网页信息的程序。

## 11.请阐述如何使用requests爬取网页内容。

（比于urllib库，Requests库非常简洁。作答时只需要写代码，以下图标不需要写）

|  |  |
| --- | --- |
| 方法 | 说明 |
| requests.request() | 构造一个请求，支撑以下各方法的基础方法 |
| requests.get() | 获取HTML网页的主要方法，对应HTTP的GET |
| requests.head() | 获取HTML网页头的信息方法，对应HTTP的HEAD |
| requests.post() | 向HTML网页提交POST请求方法，对应HTTP的POST |
| requests.put() | 向HTML网页提交PUT请求的方法，对应HTTP的PUT |
| requests.patch() | 向HTML网页提交局部修改请求，对应于HTTP的PATCH |
| requests.delete() | 向HTML页面提交删除请求,对应HTTP的DELETE |

import requests

url="http://www.163.com"

strhtml=requests.get(url)

print(strhtml.text)

## 12.请阐述如何使用BeautifulSoup爬取网页内容。

from bs4 import BeautifulSoup

file = open('4-1.html', 'rb')

html = file.read()

bs = BeautifulSoup(html,"html.parser") # 缩进格式

print(bs.prettify()) # 格式化html结构

print(bs.title) # 获取title标签的名称

print(bs.title.name) # 获取title的name

print(bs.title.string) # 获取head标签的所有内容

print(bs.head)

## 13.请阐述Kettle数据清洗的常见步骤。

1. 计算器
2. 字符串替换
3. 字符串操作
4. 值映射
5. 字段选择
6. 去除重复记录
7. 增加常量
8. 排序记录
9. 拆分字段
10. 列拆分为多行
11. 将字段值设置为常量
12. 增加序列
13. 剪切字符串

## 14.请阐述在Kettle中什么是值映射。

使用一个标准的值来替换字段里的其他值。

## 15.请阐述在Kettle中什么是采样。

在“统计”中选择“数据采样”选项来实现

## 16.请阐述在Kettle中什么是计算器，以及如何使用计算器。

计算器：对一个或多个字段进行计算的组件，该步骤提供了很多预定义的函数来处理输入字段，并且随着版本的更新还在不断增多。

## 17.请阐述在Kettle中如何对重复数据进行清洗。

在“转换”中选择“去除重复记录”选项

## 18.请阐述什么是数据迁移。

数据迁移是一种将离线存储与在线存储融合的技术，是数据系统整合中保证系统平滑升级和更新的关键部分。

## 19.请阐述数据迁移有哪些过程。

数据迁移的实现可以分为3个阶段：数据迁移前的准备、数据迁移的实施和数据迁移后的校验。

## 20.数据迁移有哪些方法？

1.基于主机的迁移方式

2.备份恢复的方式

3.基于文件系统的迁移方式

4.基于数据库的迁移方式

## 21.请阐述如何使用Kettle实现数据迁移。

成功运行kettle后在菜单栏单击文件，在“新建”中选择“转换”选项，在“输入”中选择“表输入”选项，在“输出”中选择“文本文件输出”选项。双击“表输入”选项，在打开的对话框中单击“新建”按钮，该操作用于连接数据库。

在弹出的数据库连接对话框中进行基本设置，执行操作。

## 22.请阐述如何使用urllib爬取网页内容。

import urllib.request

urllib.request.urlopen(url,data=None,[timeout,]\*, cafile=None, capath=None, cadefault=False, context=None)

response = urllib.request.urlopen('https://www.python.org')

print(response.read().decode('utf-8'))

## 23.请阐述什么是中文分词。

中文分词也称为切分，是将中文文本分割成若干个独立、有意义的基本单位的过程。

## 24.请阐述现有的中文分词算法可分为哪几大类。

1.基于字符串匹配的分词方法。2.基于理解的分词方法和基于统计的分词方法。3.按照是否与词性标注过程相结合，又可以分为单纯分词方法和分词与标注相结合的一体化方法。

## 25.请阐述jieba有哪几种分词模式。

jieba支持的三种分词模式包括：

1.精确模式：试图将句子最精确地切开，适合文本分析，并且不存在冗余；

2.全模式：把句子中所有的可以成词的词语都扫描出来, 速度非常快，但是不能解决歧义问题，存在冗余； 3.搜索引擎模式：在精确模式的基础上，对长词再次切分，提高召回率，适合用于搜索引擎分词，有冗余。

## 26.请阐述NumPy的特点。

Numpy库具有以下特征：

1.numpy库中最核心的部分是ndarray对象。它封装了同构数据类型的n维数组，它的功能将通过演示代码的形式呈现。 在数组中所有元素的类型必须一致，且在内存中占有相同的大小。

2.数组元素可以使用索引来描述，索引序号从0开始。

3.numpy数组的维数称为秩，一维数组的秩为1，二维数组的秩为2，以此类推。在Numpy中，每一个线性的数组称为是一个轴，秩其实是描述轴的数量。

## 27.请阐述Pandas的特点。

Pandas特点：

1.带有标签的数据结构，Pandas库主要围绕Series类型和DataFrame类型这两种数据结构。允许简单索引和多级索引.

2.轻松处理浮点数据中的丢失数据（以NaN表示）以及非浮点数据。

3.功能强大，灵活的按组功能来执行对数据集拆分申请，联合行动，对于聚合和转换数据。

4.可以轻松地将其他Python和NumPy数据结构中的不同索引的数据转换为DataFrame对象。

5.基于智能标签的切片，花式索引和大型数据集的子集。

6.直观的合并和连接数据集。

7.数据集的灵活的重塑和旋转。

## 28.请阐述如何使用Pandas进行数据清洗。

在DataFrame中的数据分析方法包含有：数据计算、数据扩充、数据索引、数据丢弃、数据排序和数据汇总等。可使用其中功能实现数据清洗。

## 29.请阐述如何填充缺失值。

1. 用常数填充。如果想在pandas中填充数据缺失值，可以使用fillna方法来实现，该方法会将缺失值更换为一个指定的常数值，如fillna（n）。
2. 用字典填充。在填充缺失值时，也可以以列为单位来进行
3. 用method方法填充。在填充缺失值时，也可以使用前面出现的值来填充后面同列中的缺失值
4. 用inplace=True直接修改原始值。在填充缺失值时，可以使用inplace=True来直接改变原对象的值

## 30.请阐述什么是DataCleaner。

DataCleaner是一个简单，易于使用的数据质量的应用工具，旨在分析，比较，验证和监控数据。它能够将凌乱的半结构化数据集转换为所有可视化软件，并可以读取的干净可读的数据集。此外，DataCleaner还提供数据仓库和数据管理服务。

## 31.请阐述如何安装与运行DataCleaner。

DataCleaner环境要求如下：

1. 一台有图形界面的计算机
2. 已经安装了java7 或以上版本。

要使用DataCleaner需首先在网上下载该软件，再直接解压运行即可。如已经安装成功，则在安装目录下直接双击DataCleaner图标即可运行

## 32.请阐述如何使用DataCleaner进行数据分析与清洗。

首先导入数据，并查看数据，在Customers|Analysis jobs界面中选中Analyze，使用其中的各种功能来实现。

## 33.简述如何使用Kettle进行不同格式的数据转换。

使用Kettle可以对不同格式的文件进行相互的转换。在Kettle中“核心对象”下的“输入”选项和“输出”选项即可看到多种不同格式的文件。

## 34.大数据中常见的文本文件格式有哪些？

txt格式、doc格式、xls格式、pdf格式、XML格式、JSON格式、HTML格式、tar格式、dmg格式、py格式

## 35.请描述什么是元数据。

元数据，又称中介数据、中继数据，是描述数据的数据，一般来讲，元数据主要用来描述数据属性的信息，例如记录数据仓库中模型的定义、各层级间的映射关系、监控数据仓库的数据状态及 ETL 的任务运行状态等。

## 36.请描述结构化数据和非结构化的不同存储方式。

结构化数据可以存储在传统的关系型数据库中。关系型数据库在处理事务、及时响应、保证数据的一致性方面有天然的优势。非结构化数据可以存储在新型的分布式存储中，比如Hadoop的HDFS。分布式存储在系统的横向扩展性、降低存储成本、提高文件读取速度方面有着独特的优势。

## 37.请描述什么是MP4格式。

MP4是一套用于音频、视频信息的压缩编码标准，由国际标准化组织（ISO）和国际电工委员会（IEC）下属的“动态图像专家组”制定。MP4格式的主要用于网上 流、光盘、语音发送，以及电视广播等。

## 38.请描述什么是Ogg格式。

Ogg是一种音频压缩格式，类似于MP3等的音乐格式。从商业推广上看，Ogg是完全免费、开放和没有专利限制的。在播放质量中，这种文件格式可以不断地进行 大小和音质的改良，而不影响旧有的编码器或播放器。

## 39.目前Web数据抽取的实现有哪几步。

第一步，确立采集目标，即由用户选择目标网站。

第二步：提取特征信息，即根据目标网站的网页格式，提取出采集目标数据的通性。第三步：网络信息获取，即利用工具自动的把页面数据把存到数据库。

## 40.请描述数据抽取的两个常用抽取方式。

数据抽取的两个常用抽取方式分别是全量抽取和增量抽取。其中全量抽取类似于数据迁移或数据复制，它将数据源中的表或视图的数据原封不动的从数据库中抽取出来，并转换成自己的ETL 工具可以识别的格式，全量抽取通常比较简单。而增量抽取只抽取自上次抽取以来数据库中要抽取的表中新增或修改的数据。在ETL 使用过程中，增量抽取较全量抽取应用更广。

## 41.请描述什么是HTTP协议。

HTTP也叫作超文本传输协议，它是一个基于请求与响应，无状态的，应用层的协议，常基于TCP/IP协议传输数据，是互联网上应用最为广泛的一种网络协议，所有的WWW文件都必须遵守这个标准。

## 42.请描述网络爬虫的特点。

网络爬虫在信息搜索和数据挖掘过程中扮演着重要的角色，对爬虫的研究开始于20世纪，目前爬虫技术已趋于成熟。网络爬虫通过自动提取网页的方式完成下载网页的工作，实现大规模数据的下载，省去诸多人工繁琐的工作。

## 43.请描述在Kettle中字符串操作有哪些功能。

字符串操作：该步骤提供了很多常规的字符串操作，如大小写转换、字符填充、移除空白字符等。

## 44.请描述使用Kettle可以完成数据清洗与数据转换的哪些工作。

使用Kettle可以完成数据仓库中的数据清洗与数据转换工作，常见的有：数据值的修改与映射、数据排序、重复数据的清洗、超出范围的数据清洗、日志的写入、

Javascript代码数据清洗、正则表达式数据清洗、数据值的过滤以及随机值的运算等。

## 45.请描述什么是MySql。

MySql是一个小型的关系数据库管理系统，由于该软件体积小、运行速度快、操作方便等优点，目前被广泛的应用于Web上的中小企业网站的后台数据库中。

## 46.请描述异构数据库的数据迁移特点。

异构数据库的数据迁移一般使用第三方软件实现，这种方法适用于纯数据库迁移，并且不需要关注具体的存储过程。如今第三方软件大多都提供了不同数据库转换的解决方案，不过无论哪种解决方案均须度数据库迁移后的各种数据进行测试。异构数据库的数据迁移不限操作系统以及数据库平台，不过需要大量的时间和费 用，特别是专门的定制开发，更需要长时间的测试才能真正投入使用。

## 47.请描述现有的中文分词算法分类。

现有的中文分词算法可分为三大类：基于字符串匹配的分词方法、基于理解的分词方法和基于统计的分词方法。按照是否与词性标注过程相结合，又可以分为单纯分词方法和分词与标注相结合的一体化方法。

## 48.请描述jieba支持的三种分词模式。

精确模式：试图将句子最精确地切开，适合文本分析，并且不存在冗余；全模式：把句子中所有的可以成词的词语都扫描出来, 速度非常快，但是不能解决歧义问题，存在冗余；搜索引擎模式：在精确模式的基础上，对长词再次切分，提高召回率，适合用于搜索引擎分词，有冗余。

## 49.请描述matplotlib库。

matplotlib库是Python下著名的绘图库，也是Python可视化库的基础库， matplotlib库的功能十分强大。为了方便快速绘图，matplotlib通过pyplot模块提供了一套和Matlab类似的绘图API，将众多绘图对象所构成的复杂结构隐藏在这套API内部。因此，只需要调用pyplot模块所提供的函数就可以实现快速绘图以及设 置图表的各种细节。

## 50.请描述柱状图。

柱状图也叫作条形图，是一种以长方形的长度为变量的表达图形的统计报告图，由一系列高度不等的纵向条纹表示数据分布的情况，用来比较两个或以上的数值。

## 51.请描述什么是DataCleaner。

DataCleaner是一个简单，易于使用的数据质量的应用工具，旨在分析，比较，验证和监控数据。它能够将凌乱的半结构化数据集转换为所有可视化软件，并可以读取的干净可读的数据集。此外，DataCleaner还提供数据仓库和数据管理服务。

## 52.请描述DataCleaner的特点。

可以访问多种不同类型的数据存储，如ORACLE，MYSQL，MS CSV文件等。DataCleaner还可以作为引擎来清理、转换和统一来自多个数据存储的数据，并将其统一到主数据的单一视图中。