选择

1. 在 DataFrame 中的常见扩充是对每一列或者()

A 每一行

B 每一个字符

1. 在数据管理以及数据仓库管理中最重要的部分就是（）
2. 数据映射
3. 数据转换

3.<!DOCTYPEhtml>表示（）

A.CSS

B.HTML5

C.JAVA

4.一般而言，HTML文件的结构包括头部（Head），（）两大部分。

A.标题

B.主体（Body）

5.（）是描述数据的数据，是数据仓库的重要构件，是数据仓库的导航图，在数据源抽取，数据仓库应用开发，业务分析以及数据仓库服务等过程中都发挥着重要的作用。

A.主数据

B.元数据

6.Web数据抽取可以依靠各种编程或（）来实现

A.开源软件

B.网络

7.在Pandas中可使用（）方法来处理数据缺失值。

A.dropna

B.cumsum

8.Apache框架Hadoop是一个越来越通用的分布式计算环境，主要用来处理（）

A.数据集

B.软件

9.（）是一个用来将Hadoop和关系型数据库中的数据相互转换的工具。

A.Sqoop

B.Hive

10.Liquibase不依赖于特定的（）

A.数据

B.数据库

11.UDDI（Universal Description Discovery and Integration）是一种用于描述、发现、集成Web Service的技术，它是（）的一个重要成分

A.Web Service协议栈

B.HTTP协议栈

12.Pattern finder 表示（）

A.对字符串进行提取

B.对字符串进行匹配

13.sort\_index表示（）

A.对数据值查找

B.对数据值排序

14.MP4是一套用于音频，视频信息的（）

A.压缩编码标准

B.数据转换标准

15.后缀丢弃算法可以丢弃一个（）的后缀部分。

A.单词

B.词组

16.idxmax表示（）。

A.统计个数

B.统计最大值的索引

填空

1. 数据清洗就是指把（“脏东西”）彻底洗掉
2. 数据清洗的常用方法有缺失值处理方法和（异常值处理）处理方法
3. 数据清洗的评估中，数据对用户必须是可信的以及是（可用）的
4. 在微观层面，数据清洗的对象分为（模式层）和实例层对象
5. 数据迁移的实现可以分为3个阶段：数据迁移前的（准备阶段）、数据迁移的（完成阶段）和数据迁移后的校验
6. 在进行文本挖掘的时候，首先要做的预处理是（分词）
7. 现有的中文分词方法可以分为三大类：基于（字符串）的分词方法，基于理解的分词方法和基于（统计）的分词方法
8. 错误数据包含（格式）问题数据和逻辑问题数据两类
9. 精确性描述数据是否与其对应的客观实体的特征相一致，（唯一性）描述数据是否存在重复数据
10. 在描述性统计中，（均值）也叫做平均数，中位数又称为中值
11. 文本分词是将（文本/字符串）划分为有意义的单词的过程
12. 精确性描述数据是否与其对应的客观实体的特征相一致，（完整性）描述数据是否存在缺失记录或缺失字段

判断

（√）1.在具体实现中SOAP用来描述传递信息的格式，WSDL用来描述如何访问具体的接口，UDDI用来管理，分发，查询Web Service。

（×）2.最大熵模型认为，学习概率模型时，在所有可能的概率模型中，熵最大的模型是最好的模型

（×）3.动态网页的内容是通过纯粹的HTML代码来书写，包括一些资源文件如图片，视频等

（√）4.离群值是指采集数据时可能因为技术或物理原因，数据取值超过数据值域范围

（√）5.BeautifulSoup提供一些简单的，Python式的函数来处理导航，搜索，修改分析树等功能

简答

1. 在已学的Kettle实例操作中，去除重复数据有两种方法，关键步骤是什么？

答：（1）唯一行（哈希值）：可以不事先对数据进行排序，它是在内存中对数据进行去重操作，它是根据哈希值进行比较的。

（2）去除重复记录：需要对数据进行排序，它是根据相邻两行数据是否一致进行比较的。

1. 在Pandas中，isnull（），fillna（），dropma（），cleaned（）方法有什么用？（书P209）

答：isnull（）用于返回一个布尔值，标明哪些是缺失值

fillna（）用于填充缺失值数据

dropma（）用于根据条件过滤缺失值

cleaned（）用于清除所有缺失值数据

1. Jieba的分词模式有哪些？（书P180）

答：三种模式

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

编程：

请阐述如何使用requests爬取网页内容。

import requests

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

strhtml=requests.get(url)

print(strhtml.text)

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)

## 请阐述如何使用urllib爬取网页内容。

import urllib.request  
fp = urllib.request.urlopen(r'http://www.baidu.com')  
print(fp.read().decode())  
fp.close()