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中最重要的部分就是（）
2. 数据映射
3. 数据转换

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（）彻底洗掉
2. 数据清洗的常用方法有缺失值处理方法和（）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（）的
4. 在微观层面，数据清洗的对象分为（）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（）、数据迁移的（）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（）
7. 现有的中文分词方法可以分为三大类：基于（）的分词方法，基于理解的分词方法和基于（）的分词方法
8. 错误数据包含（）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，（）描述数据是否存在重复数据
10. 在描述性统计中，（）也叫做平均数，中位数又称为中值
11. 文本分词是将（）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，（）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？

1.哈希值：不需要提前对数据进行排序，他是在内存对数据进行去除重复值操作，他是根据哈希值来比较的

2.去除重复记录：需要对数据进行排序，他是根据相邻的两行数据是否相同进行比较。

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

返回布尔值，表明哪些是缺失值

填充缺失值

根据条件进行对缺失值过滤

清除所有缺失值

1. Jieba的分词模式有哪些？（书P180）

1.精确模式；对句子进行最精确的切分，适合文本分析，不存在冗余。

2.全模式：对句子所有可以成词的词语全部扫描出来，速度非常块，存在冗余，无法处理歧义问题。

3.搜索引擎模式：在于精确模式的基础上，对长词进行再次切分，提高召回率，存在冗余，适合搜索引擎分词

编程：

请阐述如何使用requests爬取网页内容。

请阐述如何使用BeautifulSoup爬取网页内容

## 请阐述如何使用urllib爬取网页内容。