大数据分析——作业1

一、填空题

1. 一台计算机拥有4个CPU，每个CPU具备2个线程，该计算机可同时运行8个Process。
2. 利用超性能计算机的100个核并行计算1000 000次加法，不考虑通讯时间，共需做**10100=1000 000/100+100+**次加法运算。
   提示：每台计算机分别计算1000 000/100=10000次，主程序再计算100次加法。
3. 二维矩阵的最大值和最小值分别为0和10000，利用8位灰度图显示，色表由黑到白逐渐增加，那么，矩阵的8000在灰度图中多大的数值表示**$205=8000/10000*2^8#。**

二、选择题

1. 以下哪类数据可以定义为大数据？[B D]
   A 大学城采集访问系统的人脸数据 B 淘宝的商品、消费者特征、价格等信息
   C 大型望远镜拍摄的天文学图像 D 信用卡的交易数据

2. 下列哪项不是大数据分析必备技能 [D],
   A 数学与统计学 B 计算机与编程 C 行业知识 D 办公软件
3. 数据之间的相似性通过什么定义 [D]
   A 向量数据的值 B 数据点所代表向量的平行关系
   C 数据之间的距离 D 数据之间的相似度
4. 下列哪些数据不属于定类数据 [D]
   A 血型：{A，B，AB, O} B 材料属性：{木质，铁质，铝制}
   C 颜色种类：{红，绿，蓝} D 教育水平：{中学，大学本科，研究生}
5. 物体的质量属于哪类数据？ [D]
   A 定类数据 B 定序数据 C 定距数据 D 定比数据
6. 某银行的信用卡分为普通卡，银卡，金卡，白金卡，钻石卡，该数据属于什么数据？[B]
   A 定类数据 B 定序数据 C 定距数据 D 定比数据
7. 高频词属于文本分析中常见的词语，很难反映文本的风格或语义，以下那种不属于高频词。 [C]
   A “The” B “however” C “negative” D “and”

```
## 三、问答题
```

1. 阐释数据主权的概念？（5分）

答（参考）：数据主权是与领土、领海和领空权等相当，是信息和大数据时代的新权利范畴。他是指网络空间中的国家主权，是一个国家对本国数据进行管理和利用的独立自主性，不受他国干涉和侵扰的自由权，体现了国家作为控制数据权的主体地位。数据主权包括数据所有权和数据管辖权两个方面，所有权是国家对于本国数据排他性占有的权利，管辖权是国家对其本国数据享有的管理和利用的权利。

2. 列举中央处理器（CPU）的主要参数，并解释这些参数对于计算机的影响。（5分） 答（参考）：

主频：CPU内核的时钟频率，即CPU运算时的工作频率。虽然CPU的主频不代表CPU的速度，但提高主频对于提高CPU的运算速度非常重要，可以减少运算的时钟周期占用时间。

核数：芯片上芯片，用来完成所有的计算、接受/存储命令、处理数据等，是数字处理核心。

线程：CPU线程数是指逻辑上的处理单元数，即模拟出的CPU核心数，在Intel超线程技术下，一个CPU可以模拟出多于核数的线程数。线程数越多，CPU能同时并行处理的任务数越多，能够提高处理器运算部件的利用率和运算效率。

架构：CPU的架构简单来说是CPU核心的设计方案，就如房屋的布局一样，架构的设计对内存/缓存访问，各核心间的数据交换等都有影响。架构越先进，相同频率下CPU的处理效率就越高。

缓存：CPU缓存是用于减少处理器访问内存所需平均时间的部件，其容量小于内存但速度接近处理器的频率。由于缓存的运行效率极高，缓存容量的增大，可以大幅度提升CPU内部读取数据的命中率，而不用再到内存或者硬盘上寻找，以此提高系统性能。

3. 解释计算机的核与线程的概念。（5分）

物理cpu个数（physical cpu） 指主板上实际插入的cpu硬件个数。但是这一概念经常被泛泛的说成是cpu数，容易与核数和线程数等概念混淆，所以此处强调是物理 cpu 数）。

由于在主板上引入多个cpu插槽需要更复杂的硬件支持，即连接不同插槽的cpu到内存和其他资源，通常只会在服务器上才这样做。普通个人电脑的主板一般只有一个cpu插槽。

核心（core)

早期的cpu只有一个核（core），对操作系统而言，也就是只能同时运行线程。 为了提高性能，cpu厂商开始在单个物理cpu上增加核数，所以，就出现了双核（dual-core cpu）和多核cpu（multiple-cores），这样的cpu就可以同时运行运行两个或更多线程，有几个core就可并行运行同样数量的线程。 核数=CPU*核数/CPU,比如，一台服务器有10个8核CPU，则总核数为80，可并行运行80个线程。* *线程（threads or processes）**

多线程技术（simultaneous multithreading）和超线程技术（hyper–threading），运行中线程在等待的时候运行其他线程，这样提高CPU的并行性能。本质一样，是为了提高单个core同时执行多线程数的技术（充分利用单个core的计算能力，尽量让其“一刻也不得闲”。所以可以这样说：某款采用 SMT 技术的 4 核心 AMD cpu 提供了8线程同时执行的能力；某款采用 HT 技术的2核心 Intel cpu 提供了4线程同时执行的能力。

4. 将以下文本资料矩阵化 （20分，须给出详细分析过程）

Doc Id	Words
1	Jenifer likes burger and cheese .
2	Tom prefers beef burger and chicken.
3	Dave likes chichen burger, but hates beef.

答：文本数据矩阵化的过程中，列数等于所有文本资源所出现的单词集合，集合数据须合并相同的词，也可以排除高频词汇（在本次作业中不做此操作）。
行数等于文本或字符串的个数。

DocID	Jenifer	likes	burger	and	cheese	Tom	prefers	beef	chicken	Dave	but	hates
1	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	1.0	1.0	0.0	1.0	1.0	1.0	1.0	0.	0.0	0.0
3	0.0	1.0	1.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0	1.0	1.0

5. 将以下关系图矩阵化 （20分，须给出详细分析过程）

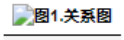
图1.关系图

图1：关系图

答：据观测，上图的关系图，存在一下联系；

	1	2	3	4	5
1	0	1	1	0	0
2	1	0	1	0	0
3	1	1	0	1	0
4	0	0	1	0	1
5	0	0	0	1	0

```python
import numpy as np
# 初始化邻接矩阵
linked_graph = np.zeros((5,5))
# 增加边
def add_link(M,d1,d2):
    d1 -=1
    d2 -=1
    M[d1,d2],M[d2,d1] = 1,1
```

```python
# 根据图像完成矩阵化
add_link(linked_graph,1,2)
add_link(linked_graph,1,3)
add_link(linked_graph,2,3)
add_link(linked_graph,3,4)
add_link(linked_graph,4,5)
print(linked_graph)
```

```
[[0. 1. 1. 0. 0.]
 [1. 0. 1. 0. 0.]
 [1. 1. 0. 1. 0.]
 [0. 0. 1. 0. 1.]
 [0. 0. 0. 1. 0.]]
```

6.→为什么样本总会存在偏差，如何减少偏差让样本统计量逼近总体统计量？（5分）

答：在采样过程中，由于采样方案的设计或者样本数不够大，很难完全反映总体的特征；亦或总体的定义本身就不可

7. 利用长度、密度和速度作为自由变量，对流体动力学方程做无量纲化处理。（20分） 答：Navier-Stokes方程如下，

$$\frac{\partial \mathbf{u}}{\partial t} + (\mathbf{u} \cdot \nabla \mathbf{u}) = \frac{-\nabla p}{\rho} + \nu \nabla^2 \mathbf{u} + \mathbf{g} \tag{1}$$

其中，ρ，p和$\mathbf{u}$分别为流体的密度、压强和速度向量;
ν是流体的运动粘度，$\mathbf{g}$为重力场;
t为时间，∇为梯度算符。

采用长度L_0,密度ρ_0和速率u_0为自由变量的常数，无量纲化的变化如下：
长度量：$\nabla = L_0^{-1}\hat{\nabla}$
密度：$\rho = \rho_0\hat{\rho}$
速度：$\mathbf{u} = u_0\hat{\mathbf{u}}$
时间：$t = t_0\hat{t}$, 其中，$t_0 = L_0/u_0$
压强：$p = p_0\hat{p}$, 其中，$p_0 = \rho_0 u_0^2$
运动粘度: $\nu = \nu_0\hat{\nu}$
重力场：$\mathbf{g} = g_0\hat{\mathbf{g}}$,

方程(1)的左右两边同时乘以L_0/u_0^2，则，右侧第二项$\hat{\nu}\hat{\nabla}^2\hat{\mathbf{u}}$的系数变为变为：

$$\frac{L_0}{u_0^2}\nu_0 L_0^{-2}u_0 = \frac{\nu_0}{L_0 u_0},$$

这样，将无量纲的运动粘度$\hat{\nu}$融入雷诺数，可得为$\mathrm{Re} = \frac{L_0 u_0}{\nu_0\hat{\nu}} = \frac{L_0 u_0}{\nu}$.
注意：ν, ν_0和$\hat{\nu}$的区别

右侧第三项($\hat{\mathbf{g}}$)的系数变为,

$$\frac{g_0 L_0}{u_0^2} = \frac{1}{\mathrm{Fr}^2}$$

由此，可得福禄数$\mathrm{Fr} = \frac{u0}{L0*g0}$.

无量纲化之后，方程(1)变化为：

$$\frac{\partial \hat{\mathbf{u}}}{\partial \hat{t}} + (\hat{\mathbf{u}} \cdot \hat{\nabla}\hat{\mathbf{u}}) = \frac{-\hat{\nabla}\hat{p}}{\hat{\rho}} + \frac{1}{\mathrm{Re}}\hat{\nabla}^2\hat{\mathbf{u}} + \frac{\hat{\mathbf{g}}}{\mathrm{Fr}^2} \tag{2}$$

对于，不可压缩流体，即ρ为常数ρ_0，此时无量纲的密度为$\hat{\rho} = 1$，方程(2)变为：

$$\frac{\partial \hat{\mathbf{u}}}{\partial \hat{t}} + (\hat{\mathbf{u}} \cdot \hat{\nabla}\hat{\mathbf{u}}) = -\hat{\nabla}\hat{p} + \frac{1}{\mathrm{Re}}\hat{\nabla}^2\hat{\mathbf{u}} + \frac{\hat{\mathbf{g}}}{\mathrm{Fr}^2}. \tag{3}$$

作业2-答案

1、 利用正则表达式搜索一下模式；

(1) 代表can, man, fan ； 排除：dan, ran, pan。
答案： [cmf]an 或者[^drp]an
(2) 代表Ana, Bob, Cpc；排除：aax, bby, ccz。
答案： [A-Z][a-z]{2}
(3) 代表 “1. abc”, “2. abc ”, “3. abc” ； 排除4.abc。
答案：^\d.\s+abc 或者^[0-9].\s+abc
(4) 代表file_record_transcript.pdf, file_07241999.pdf, 排除：testfile_fake.pdf.tmp 提示：利用( )，提取文件名，排除拓展名.pdf等。
答案：^(file.+).pdf$

2. 利用Python语言，交换一下变量的值 A,B=15,6

In [2]:
```
A, B=15, 6
print(A)
print(B)
```

```
15
6
```

In [3]:
```
A, B=B, A
print(A)
print(B)
```

```
6
15
```

3. Python的单行注释和多行注释分别怎么使用?

In [5]:
```
# 这是单行注释，一般用于注释代码的原理。<br>
'''
这是多行注释；<br>
多行注释一般用于代码的解释和说明文字。<br>
'''

```

Out[5]: '\n这是多行注释；
\n多行注释一般用于代码的解释和说明文字。
\n'

In [ ]:
```
4.  Python用什么标识分割模块(CELL)？
#%%用于分割CELL
```

5. 阐释时间列的强平稳性和弱平稳性的区别。
答（参考）：
平稳性定义时间序列的统计特征不随时间变化。
强平稳性是指时间序列的变量的联合概率分布不随时间变化，由于在实际使用中，联合概率分布很难直接测量，所以该定义在实践中很少使用。
弱平稳性为了弥补强平稳性在实践中测量和检验的实操性，检验时间序列的一阶和二阶动量，即平均值和方差不随时间变化，这样的定义不但方便测量，也便于检测。

6、 白噪音具备怎么样的特征，列举两项白噪音在科学、金融、经济、工程等领域的应用。 答（参考）：
白噪音是一个随机过程，其随机变量之间相互独立（不相干），其概率分布符合标准正态分布，即空间或时间上的平均值为0，标准差为1。 白噪音是理论上的标准随机过程，一般应用时间序列的噪音检测，也用于科学、金融等领域的随机过程模拟。

7. 描述土地或图片资料如何进行聚类分析？描述具体步骤.

答（参考）：

在地图图片中，水域、林地、土壤等元素的颜色各不相同，假设图片表示为$I(x, y, 3)$，可将图片的像素展开为R(x,y),G(x,y),B(x,y)的三个维度的强度信息。 将像素视为RGB三原色的向量。一般同类颜色的元素在三维RGB空间中距离较近，形成聚类。这样通过聚类算法，在三维空间内能够区分颜色相同的区域，从而标记不同特征 的区域信息，进行聚类分析。

```
8. 阐释距离函数和相似度的关联和差异？<br>

答（参考）：<br>
**距离函数**一般有定义数据点的高维度矢量之差的函数定义，代表矢量之间的距离，该距离可
**相似度**一般定义为距离的倒数，归一化在0和1之间，描述数据点之间的相似度。相似度越高

```

9. 为什么K-means一般需要重复多次?
答（参考）：
由于在计算K-means聚类的时候，聚类的质心的初始值随机产生。在计算K-means聚类时，虽然可能已经获得正确的聚类中心，但是不能排除一定收敛到全局最小值，所以为了排斥随机质心导致的影响，需要多次重复该过程，以确保收敛到全局最小值。

```
10. 为什么K-means算法用于异常值多的数据容易出现大的偏差？如何解决该问题？<br>
答（参考）：<br>
由于K-means定义的距离为欧式几何距离，异常值一般距聚类中心较远，对计算结果的影响很大，
解决方案：（1）人工删除异常值，再进行聚类分析；（2）改用对异常值不敏感的算法，比如数

```

11. 以下三维数据为什么可以降低维度，降维考虑的主要点是什么，请解释?
答（参考）：
该数据虽然表示为三维数据，但是通过探索性分析发现，多数数据点分布在一个二维平面上，所以数据的本质属于二维。可通过PCA或者SVD等算法进行降维，降维之后的二维数据，不但有利于发现数据特征，而且极大降低了计算量。

12. 解释监督性学习、无监督性学习，强化学习的意义和他们之间的区别?

监督学习（Supervised learning）：

监督学习即具有特征（feature）和标签（label）的，即使数据是没有标签的，也可以通过学习特征和标签之间的关系，判断出标签——分类。

简言之：提供数据，预测标签。比如对动物猫和狗的图片进行预测，预测label为cat或者dog。

通过已有的一部分输入数据与输出数据之间的对应关系，生成一个函数，将输入映射到合适的输出，例如分类和回归问题.

无监督学习（Unsupervised learning）：

无监督学习即只有特征，没有标签，只有特征，没有标签的训练数据集中，通过数据之间的内在联系和相似性将他们分成若干类——聚类。根据数据本身的特性，从数据中根据某种度量学习出一些特性。

eg.比如一个人没有见过恐龙和鲨鱼，如果给他看了大量的恐龙和鲨鱼，虽然他没有恐龙和鲨鱼的概念，但是他能够观察出每个物种的共性和两个物种间的区别的，并对这两种动物予以区分。

简言之：给出数据，寻找隐藏的关系。

强化学习（Reinforcement learning）：

强化学习与半监督学习类似，均使用未标记的数据，但是强化学习通过算法学习是否距离目标越来越近，我理解为激励与惩罚函数。类似生活中，女朋友不断调教直男友变成暖男。

简言之：通过不断激励与惩罚，达到最终目的。

区别：

（1）监督学习有反馈，无监督学习无反馈，强化学习是执行多步之后才反馈。

（2）强化学习的目标与监督学习的目标不一样，即强化学习看重的是行为序列下的长期收益，而监督学习往往关注的是和标签或已知输出的误差。

（3）强化学习的奖惩概念是没有正确或错误之分的，而监督学习标签就是正确的，并且强化学习是一个学习+决策的过程，有和环境交互的能力（交互的结果以惩罚的形式返回），而监督学习不具备。

```
13。*解释机器学习中，什么是过拟合和欠拟合?

**欠拟合**，根本的原因是特征维度过少，导致拟合的函数无法满足训练集，误差较大。<br>

欠拟合问题可以通过增加特征维度来解决。<br>

**过拟合** 根本的原因则是特征维度过多，导致拟合的函数完美的经过训练集，但是对新数据的

解决过拟合问题，则有2个途径：<br>

减少特征维度：可以人工选择保留的特征，或者模型选择算法 <br>
正则化：保留所有的特征，通过降低参数θ的值，来影响模型<br>
```