

HOMEWORK #2 (10分)

一、数据属性类型练习：(3分)

对于下面列出的每一个数据属性，从 {Nominal, Ordinal, Interval, Ratio} 四个选项选出最合适的类型，并说明理由。

- ✓ 教师的职称
- ✓ 手机号码
- ✓ 体重
- ✓ 出生日期
- ✓ 出生地
- ✓ 年龄

二、计算统计信息：(2分)

假定用于分析的数据为某食品中的脂肪含量。样本的值分别是：

脂肪(%)	9.0	26.5	7.8	17.8	31.4	26.5	27.4	26.5
脂肪(%)	34.6	43.0	28.8	33.4	27.4	34.1	32.9	41.2

注意这里所有样本属于同一组数据。

- 1) 计算这组数据的均值、中位数和众数；
- 2) 给出五数概括，并画出盒图。

三、文本数据的表示：(5分)

我们提供了纽约时报的部分新闻语料，在文件夹 `nyt corp0/`中，每一个文件表示一篇文档。使用任意一种编程语言，完成如下练习：

- 1) **文档的表示**：根据语料内容构造词典，然后将语料中的每篇文档都表示成词典上的 tf-idf 向量。
- 2) **词语的表示**：使用先前构造的词典，计算词语的共现矩阵，从而得到词语的共现向量。（共现矩阵的一个元素 $C(i, j)$ 表示词语 i 和词语 j 共同在文档中出现的次数）
- 3) **文档距离计算与分析**：任选一篇文档，使用 tf-idf 向量找出与它欧式距离最近/余弦相似度最高的各 5 篇文档，并简单分析这 10 篇文档是否与它内容相似。
- 4) **词语距离计算与分析**：任选一个词典中的词，使用共现向量找出与它欧式距离最近/余弦相似度最高的各 5 个词，并简单说明这 10 个词是否与它意思相近。

提交说明：需要提交源代码与报告。报告中简单说明 1)、2)的实现思路，如词典的构造细节，然后写清楚 3)、4)的结果与分析。

注意：请不要调用第三方的库来直接生成文档和词语的表示。