Given the following three review texts and their class labels:

ID	Input review text	Class label
1	Good, thanks	Positive
2	No impressive, thanks	Negative
3	Impressive good	Positive

Determine the class label of the 4-th review text “No, thanks” using the Naïve Bayesian and *k*-NN (*k*=1) classifiers, respectively.

In the pre-processing step, all lower-case words were extracted, and all punctuations were discarded from all texts, as follows:

ID	good	thanks	no	impressive	Class label
1	1	1	0	0	Positive
2	0	1	1	1	Negative
3	1	0	0	1	Positive
4	0	1	1	0	?

(1) 基于贝叶斯公式，给定 ID=4 的测试文本，类标签为 Positive 和 Negative 的条件概率分别如下：

P(Class label=“Positive” | ID=4)
= P(Class label=“Positive”) × P(ID=4 | Class label=“Positive”) / P(ID=4)

P(Class label=“Negative” | ID=4)
= P(Class label=“Negative”) × P(ID=4 | Class label=“Negative”) / P(ID=4)

贝叶斯分类器的决策规则如下：

如果 *P*(Class label="Positive" | ID=4) > *P*(Class label="Negative" | ID=4)，该测试文本的类标签预测为"Positive"；
如果 *P*(Class label="Positive" | ID=4) < *P*(Class label="Negative" | ID=4)，该测试文本的类标签预测为"Negative"；
如果 *P*(Class label="Positive" | ID=4) = *P*(Class label="Negative" | ID=4)，无法预测。

基于题给训练数据集（三篇包含类标签的评论文本，ID 为 1、2、3）可知：

P(Class label="Positive") = 2/3,
P(Class label="Negative") = 1/3.

因此，

P(Class label="Positive" | ID=4)
= (2/3) × *P*(ID=4 | Class label="Positive") / *P*(ID=4)

P(Class label="Negative" | ID=4)
= (1/3) × *P*(ID=4 | Class label="Negative") / *P*(ID=4)

在计算 *P*(ID=4 | Class label="Positive")和 *P*(ID=4 | Class label="Negative")时，有如下两种方式：

a) 向量表示形式：

P(ID=4 | Class label="Positive")
= *P*("good"=0, "thanks" =1, "no"=1, "impressive"=0 | Class label="Positive")
= *P*("good"=0 | Class label="Positive") × *P*("thanks" =1 | Class label="Positive")
× *P*("no"=1 | Class label="Positive") × *P*("impressive"=0 | Class label="Positive")
= 0 × (1/2) × 0 × (1/2) = 0.

P(ID=4 | Class label="Negative")
= *P*("good"=0, "thanks" =1, "no"=1, "impressive"=0 | Class label="Negative")
= *P*("good"=0 | Class label="Negative") × *P*("thanks" =1 | Class label="Negative")
× *P*("no"=1 | Class label="Negative") × *P*("impressive"=0 | Class label="Negative")
= 1 × 1 × 1 × 0 = 0.

代入前面的式子，

P(Class label="Positive" | ID=4) = 0.
P(Class label="Negative" | ID=4) = 0.

故：无法预测（Unknown 或 new）。

b) 词袋表示形式：

P(ID=4 | Class label=“Positive”)
= *P*(“thanks”, “no” | Class label=“Positive”)
= *P*(“thanks” | Class label=“Positive”) × *P*(“no” | Class label=“Positive”)
= (1/4) × 0 = 0.
说明：类标签为“Positive”的训练文本，总词袋为{“good”, “thanks”, “good”, “impressive”}
故，*P*(“thanks” | Class label=“Positive”) = (1/4).

P(ID=4 | Class label=“Negative”)
= *P*(“thanks”, “no” | Class label=“Negative”)
= *P*(“thanks” | Class label=“Negative”) × *P*(“no” | Class label=“Negative”)
= (1/3) × (1/3).
说明：类标签为“Negative”的训练文本，总词袋为{“thanks”, “no”, “impressive”}
故，*P*(“thanks” | Class label=“Negative”) = (1/3).

代入前面的式子，

P(Class label=“Positive” | ID=4) = (2/3) * (1/4) * 0 / *P*(ID=4) = 0 / *P*(ID=4).
P(Class label=“Negative” | ID=4) = (1/3) * (1/3) * (1/3) / *P*(ID=4) = (1/27) / *P*(ID=4).

因：*P*(ID=4) > 0,

故：we assign “Negative” to the review text with ID equal to 4.

(2) We can use the Euclidean distance to measure the dissimilarity between paired texts:

$$d(\text{ID=4, ID=1}) = \sqrt{(0-1)^2+(1-1)^2+(1-0)^2+(0-0)^2} = \sqrt{2}$$

$$d(\text{ID=4, ID=2}) = \sqrt{(0-0)^2+(1-1)^2+(1-1)^2+(0-1)^2} = 1$$

$$d(\text{ID=4, ID=3}) = \sqrt{(0-1)^2+(1-0)^2+(1-0)^2+(0-1)^2} = 2$$

For the review text with ID equal to 4, the review text with ID equal to 2 (whose class label is “Negative”) is the most similar text. Thus, we assign “Negative” to the review text with ID equal to 4 according to the *k*-NN (*k*=1) classifier.