TF-IDF In-class exercise

You are provided with a small corpus of documents. Your task is to categorize a new document (Test Document) **using the K-Nearest Neighbors (KNN) algorithm with K = 1, based on the TF-IDF values** of the terms in each document.

Documents:

- **DOC1 (Category A):**
  - Words: apple, orange, fruit
  - Word counts: apple: 2, orange: 1, fruit: 3
- **DOC2 (Category B):**
  - Words: apple, banana, fruit
  - Word counts: apple: 1, banana: 2, fruit: 1
- **DOC3 (Category A):**
  - Words: banana, mango, fruit
  - Word counts: banana: 3, mango: 1, fruit: 2

Test Document (Unknown Category):

- Words: apple, mango, fruit
- Word counts: apple: 1, mango: 1, fruit: 2

	orange	apple	banana	fruit	mango
doc 1	1	2	0	3	0
doc 2	0	1	2	1	0
doc 3	0	0	3	2	1
new doc	0	1	0	2	1
	×idf orange	×idf apple	×idf banana	×idf fruit	× idf mango

	Doc1	Doc2	Doc 3	idf
orange	1 ×1.585	0	0	$\log_2 3 = 1.585$
apple	2 ×0.565	1 ×0.585	0	$\log_2 \frac{3}{2} = 0.585$
banana	0	2 ×0.585	3 ×0.585	$\log_2 \frac{3}{2} = 0.585$
fruit	3 ×0	1 ×0	2 ×0	$\log_2 1 = 0$
mango	0	0	1 ×1.585	$\log_2 3 = 1.585$