**CHƯƠNG 5: TEXT**

****

**📘 Bài 1 – Tiền xử lý văn bản**

**❓ Đề bài:**

Trình bày các bước tiền xử lý cho đoạn văn:

“Sáng nay 24/5, tại TP.HCM, giá USD tự do đạt mốc 25.000 đồng/USD!”

**📌 Công thức áp dụng:**

Không có công thức toán học, nhưng các bước chuẩn bao gồm:

* Lowercasing
* Loại bỏ ký tự đặc biệt
* Tokenization
* Stopword removal
* Lemmatization / Stemming

**✅ Giải chi tiết:**

**Bước 1 – Lowercasing**:

“sáng nay 24/5, tại tp.hcm, giá usd tự do đạt mốc 25.000 đồng/usd!”

**Bước 2 – Loại ký tự đặc biệt và số**:

“sáng nay tại tphcm giá usd tự do đạt mốc đồng usd”

**Bước 3 – Tokenization (tách từ)**:

["sáng", "nay", "tại", "tphcm", "giá", "usd", "tự", "do", "đạt", "mốc", "đồng", "usd"]

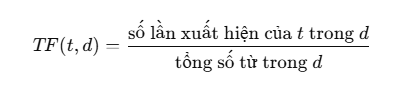
**Bước 4 – Loại từ dừng**: (VD: “tại”, “nay”...)

["sáng", "tphcm", "giá", "usd", "tự", "do", "đạt", "mốc", "đồng", "usd"]



**Bước 1: Tính TF (Term Frequency)**

TF là tần suất xuất hiện của một từ trong tài liệu, tính theo công thức:



Tài liệu:

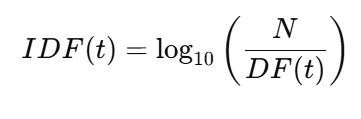
* **d1**: ["học", "máy", "và", "trí", "tuệ", "nhân", "tạo"] → 7 từ  
  → "máy" xuất hiện 1 lần ⇒ TF = 1/7
* **d2**: ["trí", "tuệ", "nhân", "tạo", "là", "tương", "lai"] → 7 từ  
  → "máy" xuất hiện 0 lần ⇒ TF = 0/7 = 0
* **d3**: ["máy", "học", "giúp", "xây", "dựng", "mô", "hình", "dự", "đoán"] → 9 từ  
  → "máy" xuất hiện 1 lần ⇒ TF = 1/9

✅ **Kết quả TF:**

* TF("máy", d1) = **1/7 ≈ 0.143**
* TF("máy", d2) = **0**
* TF("máy", d3) = **1/9 ≈ 0.111**

**Bước 2: Tính IDF (Inverse Document Frequency)**

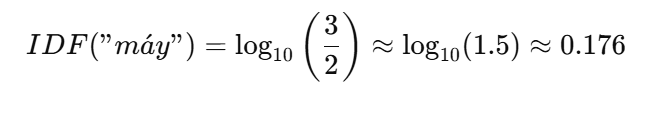
IDF đo lường tầm quan trọng của một từ trong toàn bộ tập tài liệu:



Trong đó:

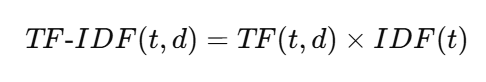
* **N** = Tổng số tài liệu = 3
* **DF(t)** = Số tài liệu chứa từ **t**

→ Từ "máy" xuất hiện trong **2** tài liệu: d1, d3

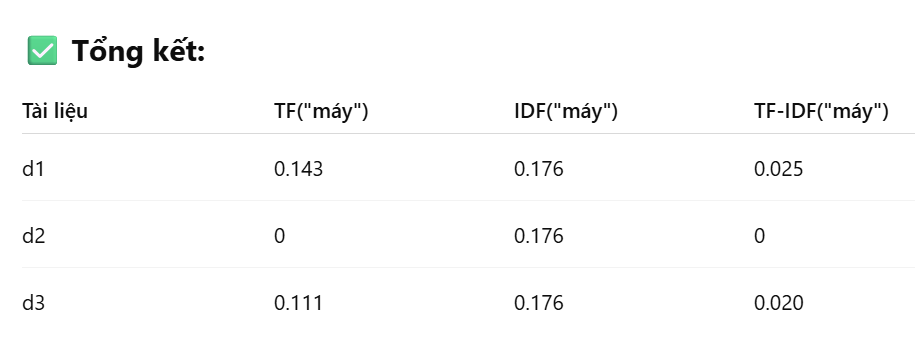


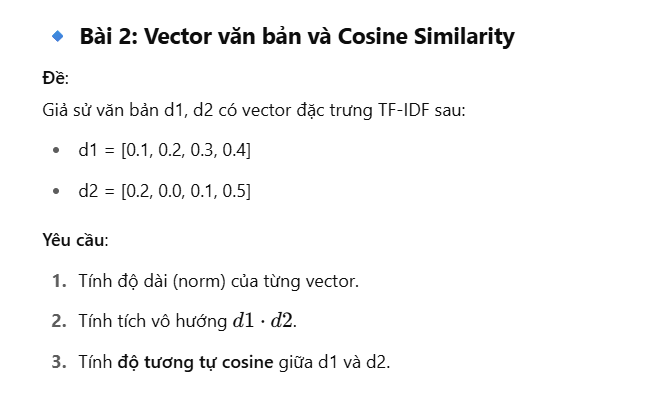
✅ **Kết quả IDF: 0.176**

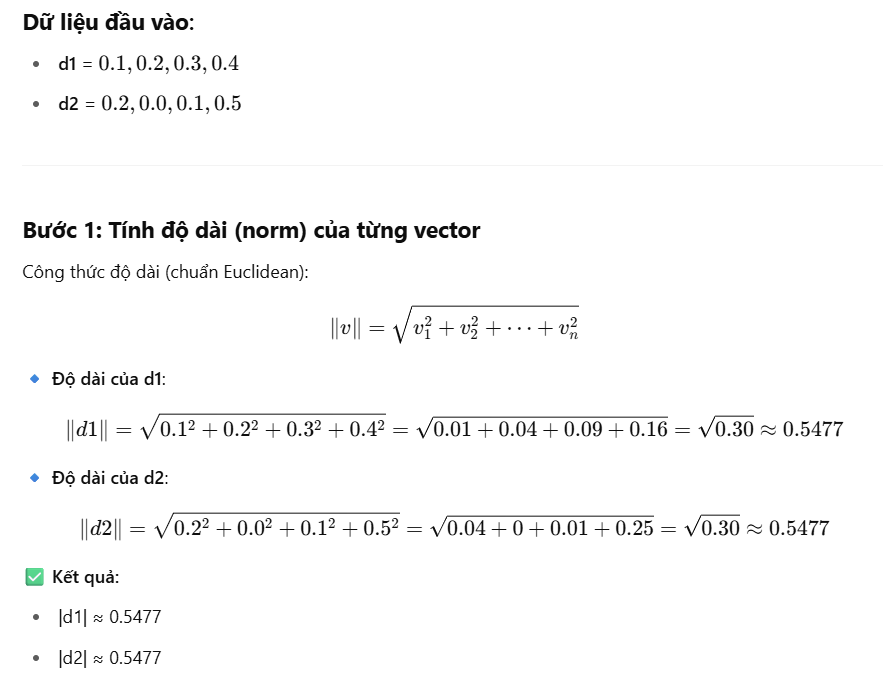
**Bước 3: Tính TF-IDF**

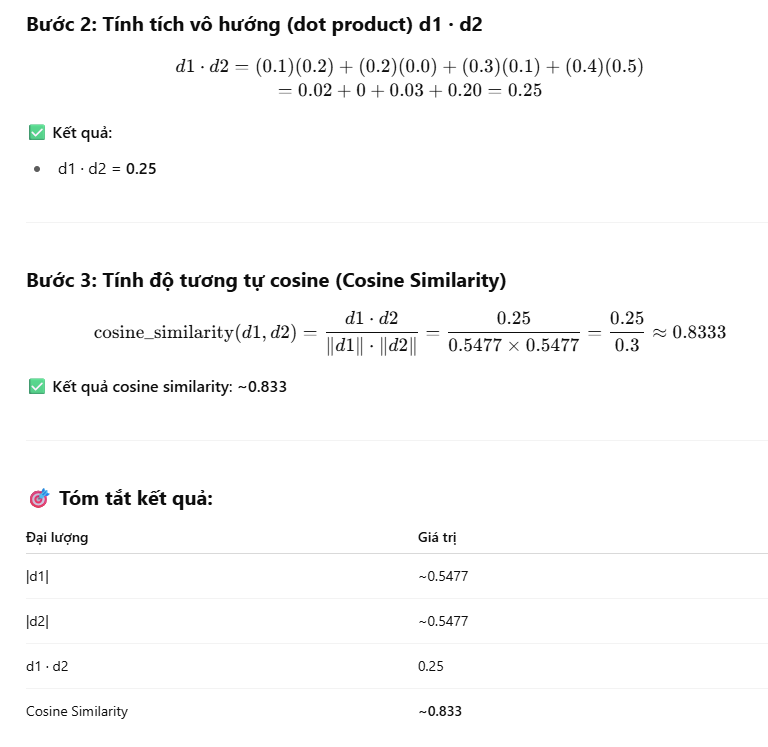


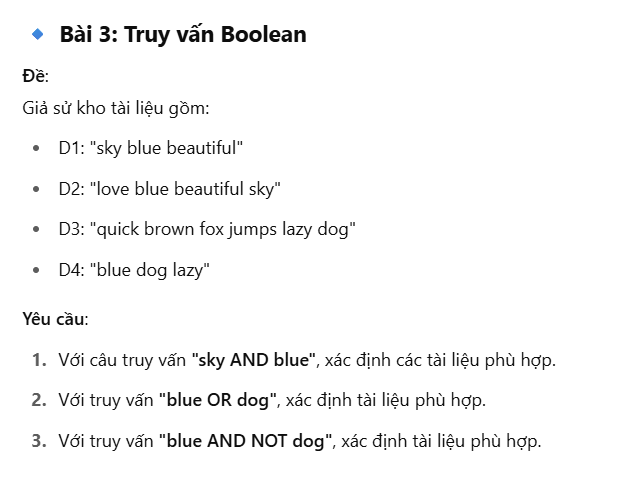
* TF-IDF("máy", d1) = 1/7 × 0.176 ≈ **0.025**
* TF-IDF("máy", d2) = 0 × 0.176 = **0**
* TF-IDF("máy", d3) = 1/9 × 0.176 ≈ **0.020**











**Dữ liệu tài liệu:**

* **D1**: "sky blue beautiful" → {sky, blue, beautiful}
* **D2**: "love blue beautiful sky" → {love, blue, beautiful, sky}
* **D3**: "quick brown fox jumps lazy dog" → {quick, brown, fox, jumps, lazy, dog}
* **D4**: "blue dog lazy" → {blue, dog, lazy}

**Yêu cầu 1: Truy vấn "sky AND blue"**

→ Tìm tài liệu chứa **cả hai từ** sky **và** blue

* **D1**: có "sky", "blue" ✅
* **D2**: có "sky", "blue" ✅
* **D3**: không có "sky" ❌
* **D4**: có "blue" nhưng không có "sky" ❌

✅ **Kết quả: D1, D2**

**Yêu cầu 2: Truy vấn "blue OR dog"**

→ Tìm tài liệu chứa **ít nhất một** trong hai từ blue **hoặc** dog

* **D1**: có "blue" ✅
* **D2**: có "blue" ✅
* **D3**: có "dog" ✅
* **D4**: có cả "blue" và "dog" ✅

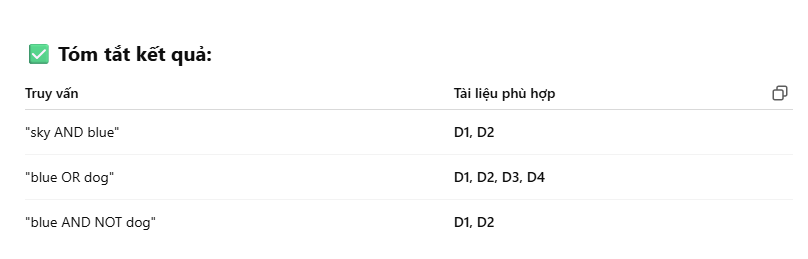
✅ **Kết quả: D1, D2, D3, D4**

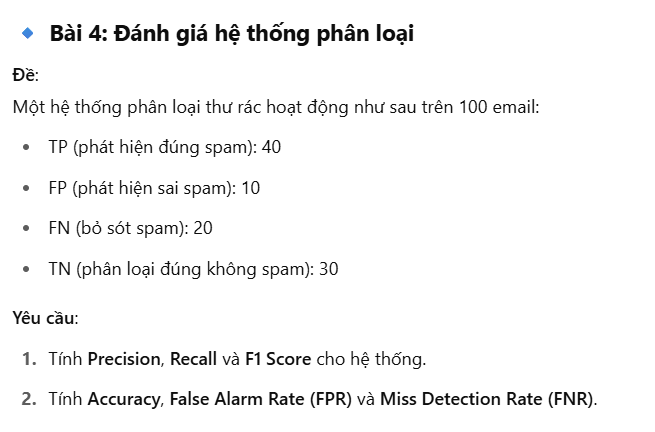
**Yêu cầu 3: Truy vấn "blue AND NOT dog"**

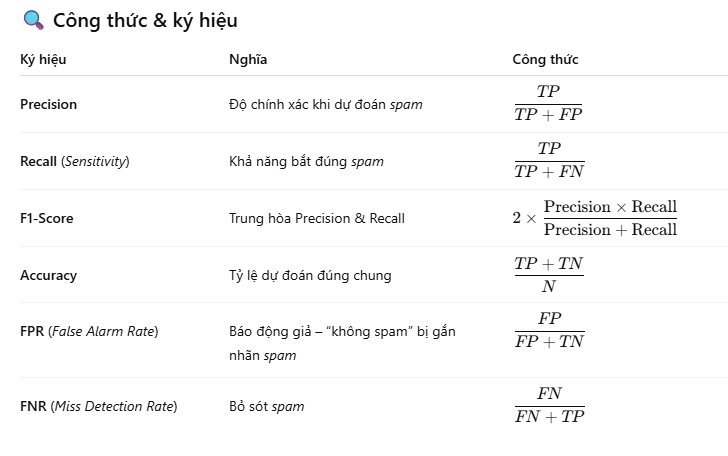
→ Tìm tài liệu chứa **"blue"** và **không chứa "dog"**

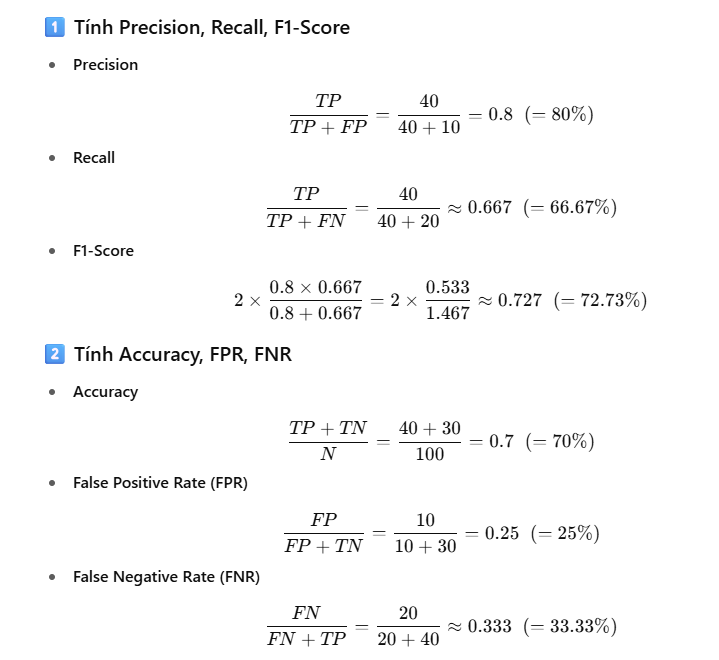
* **D1**: có "blue", không có "dog" ✅
* **D2**: có "blue", không có "dog" ✅
* **D3**: không có "blue" ❌
* **D4**: có "blue", có "dog" ❌

✅ **Kết quả: D1, D2**

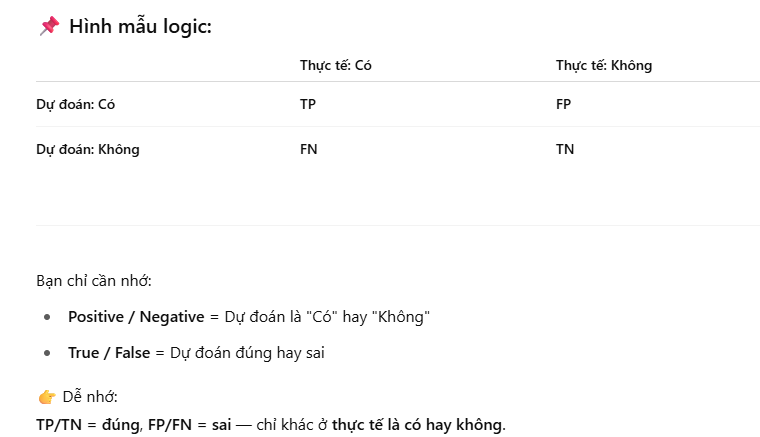


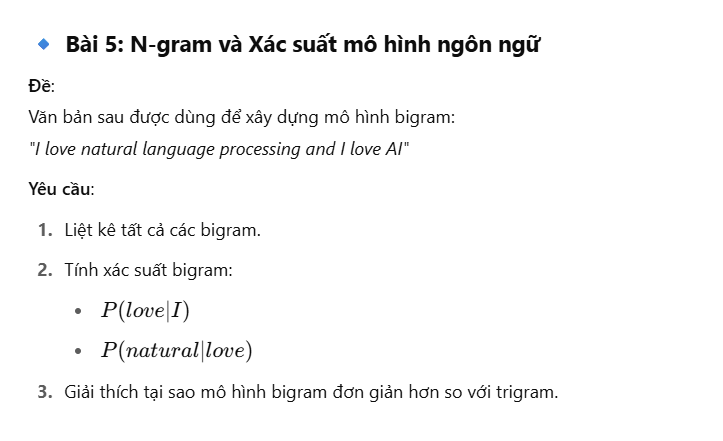












Văn bản:

"I love natural language processing and I love AI"

Ta xây dựng **mô hình bigram** (chuỗi gồm 2 từ liên tiếp).

**✅ 1. Liệt kê tất cả các bigrams**

Trước tiên, ta tách câu thành các từ:

[I, love, natural, language, processing, and, I, love, AI]

Sau đó, liệt kê tất cả các cặp 2 từ liên tiếp (**bigrams**):

1. (I, love)
2. (love, natural)
3. (natural, language)
4. (language, processing)
5. (processing, and)
6. (and, I)
7. (I, love) ✅ (xuất hiện lại)
8. (love, AI)

➡️ Danh sách các **bigrams**:

[(I, love), (love, natural), (natural, language),

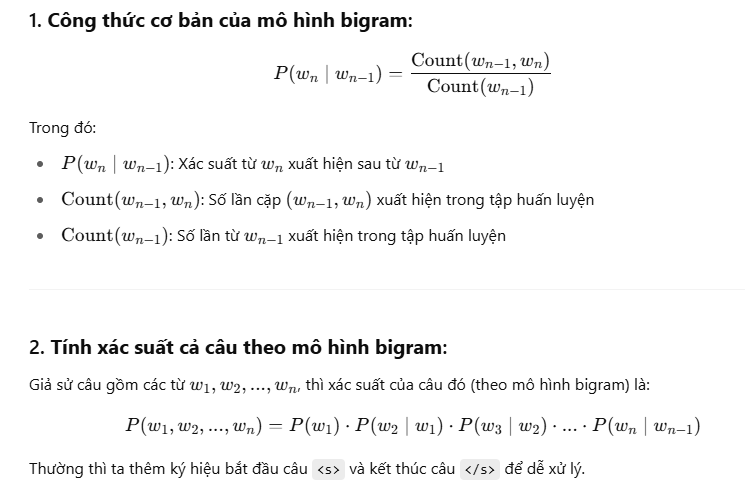
(language, processing), (processing, and),

(and, I), (I, love), (love, AI)]

**✅ 2. Tính xác suất bigram:**

**a.** ****

Công thức tính xác suất bigram:



Trong văn bản:

* Số lần xuất hiện của **(I, love)**: **2 lần**
* Số lần xuất hiện của **"I"**: **2 lần**

P(love∣I)=22=1.0P(love \mid I) = \frac{2}{2} = 1.0P(love∣I)=22​=1.0

**b. P(natural∣love)P(natural \mid love)P(natural∣love)**

* Số lần **(love, natural)**: **1 lần**
* Số lần **"love"** xuất hiện: **2 lần** (love natural và love AI)

P(natural∣love)=12=0.5P(natural \mid love) = \frac{1}{2} = 0.5P(natural∣love)=21​=0.5

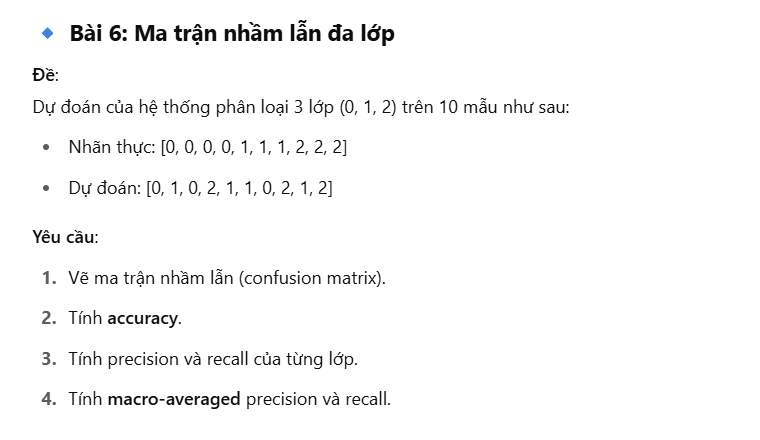
**✅ 3. Vì sao mô hình bigram đơn giản hơn trigram?**

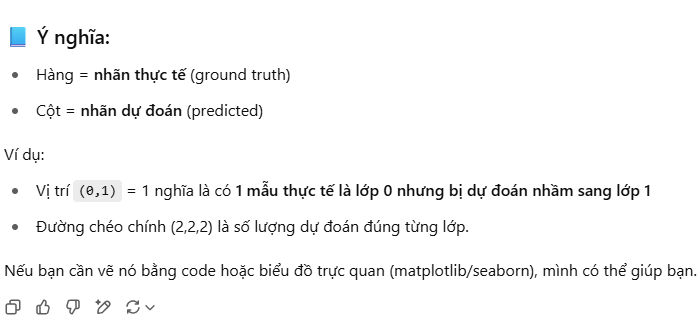
**So sánh đơn giản:**

| **Mô hình** | **Dựa vào mấy từ trước?** | **Ví dụ khi đoán từ sau "I love"** |
| --- | --- | --- |
| **Unigram** | Không dựa từ nào | Chọn từ tiếp theo theo xác suất riêng từng từ |
| **Bigram** | Dựa vào **1 từ trước** | Xác suất theo P(wi∣wi−1)P(w\_i \mid w\_{i-1})P(wi​∣wi−1​) → dựa vào "love" |
| **Trigram** | Dựa vào **2 từ trước** | P(wi∣wi−2,wi−1)P(w\_i \mid w\_{i-2}, w\_{i-1})P(wi​∣wi−2​,wi−1​) → dựa vào "I love" |

➡️ **Bigram đơn giản hơn** vì:

* Dễ tính toán hơn (ít dữ liệu hơn)
* Cần ít bộ nhớ và xử lý hơn
* Tránh được **"sparse data"** (dữ liệu hiếm) so với trigram







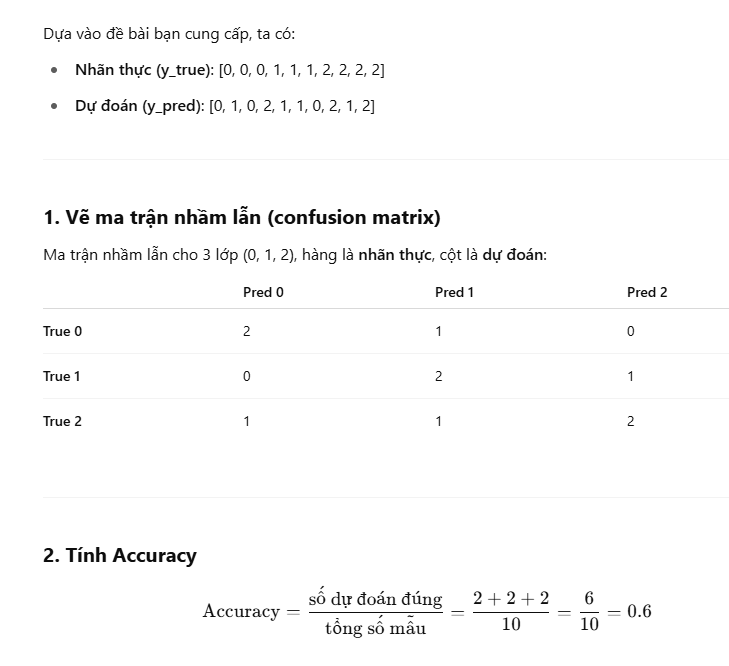
**📌 1. Khái niệm ma trận nhầm lẫn**

Ma trận nhầm lẫn là một bảng dùng để mô tả hiệu suất của một mô hình phân loại. Nó so sánh giữa:

* **Nhãn thực tế** (ground truth)
* **Nhãn dự đoán** (do mô hình đưa ra)

**✅ Cách đọc ma trận này**

* **(0,0) = 2**: Có 2 mẫu **thực sự là lớp 0** và được **dự đoán đúng là lớp 0**
* **(0,1) = 1**: Có 1 mẫu **thực sự là lớp 0**, nhưng bị **dự đoán nhầm thành lớp 1**
* **(2,1) = 1**: Có 1 mẫu **thực sự là lớp 2**, nhưng bị **dự đoán nhầm thành lớp 1**
* Các ô trên đường chéo (ô [0,0], [1,1], [2,2]) là các **dự đoán đúng**



**3. Tính Precision và Recall của từng lớp**

**Precision (P)** = TPTP + FP\frac{\text{TP}}{\text{TP + FP}}TP + FPTP​  
**Recall (R)** = TPTP + FN\frac{\text{TP}}{\text{TP + FN}}TP + FNTP​

**Lớp 0:**

* TP = 2 (dự đoán 0, đúng nhãn 0)
* FP = 1 (dự đoán 0 nhưng nhãn thật là 2)
* FN = 1 (nhãn thật là 0 nhưng dự đoán sai là 1)

→

* Precision₀ = 2 / (2+1) = 2/3 ≈ **0.667**
* Recall₀ = 2 / (2+1) = 2/3 ≈ **0.667**

**Lớp 1:**

* TP = 2 (dự đoán 1, đúng nhãn 1)
* FP = 2 (dự đoán 1 nhưng nhãn là 0 và 2)
* FN = 1 (nhãn là 1 nhưng dự đoán sai là 2)

→

* Precision₁ = 2 / (2+2) = 2/4 = **0.5**
* Recall₁ = 2 / (2+1) = 2/3 ≈ **0.667**

**Lớp 2:**

* TP = 2 (dự đoán 2, đúng nhãn 2)
* FP = 1 (dự đoán 2 nhưng nhãn là 1)
* FN = 1 (nhãn là 2 nhưng dự đoán sai là 1)

→

* Precision₂ = 2 / (2+1) = 2/3 ≈ **0.667**
* Recall₂ = 2 / (2+1) = 2/3 ≈ **0.667**

**4. Macro-averaged Precision và Recall**

Tính trung bình cộng của precision và recall các lớp:

* **Macro-averaged Precision** = (0.667 + 0.5 + 0.667) / 3 ≈ **0.611**
* **Macro-averaged Recall** = (0.667 + 0.667 + 0.667) / 3 = **0.667**

**📘 Bài 6 – Gán nhãn NER**

**❓ Đề bài:**

Gán nhãn NER cho câu:

“Nguyễn Văn A làm việc tại VinGroup ở Hà Nội.”

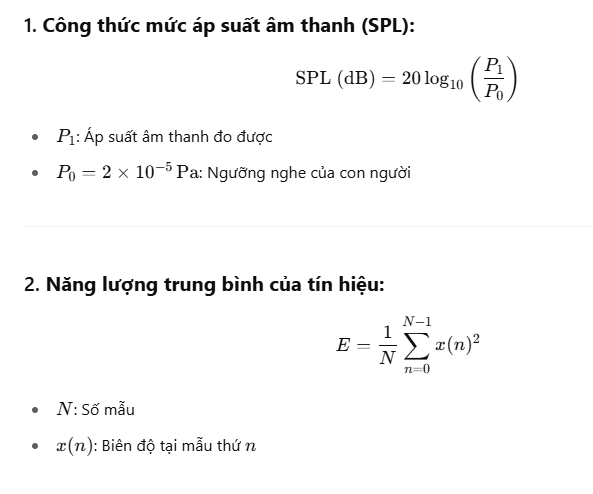
**📌 Các nhãn phổ biến:**

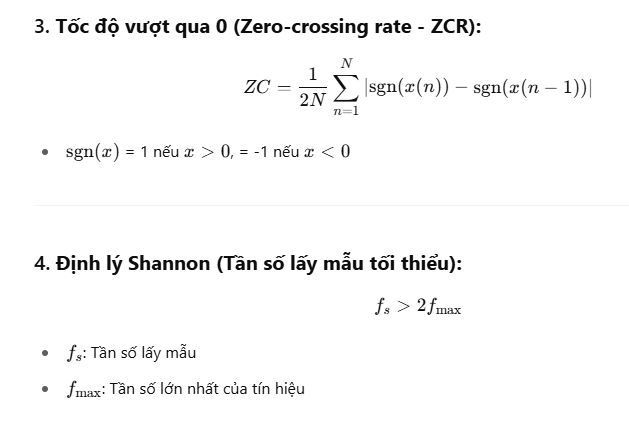
* PER: Tên người
* LOC: Địa điểm
* ORG: Tổ chức

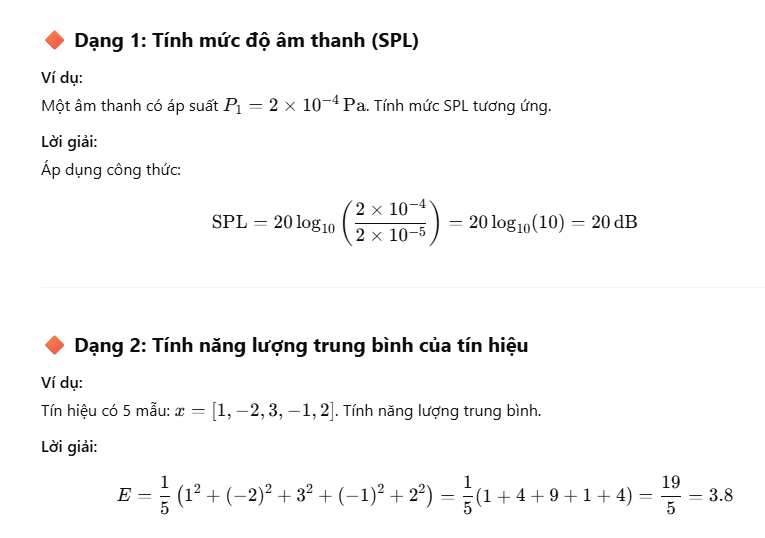
**✅ Giải chi tiết:**

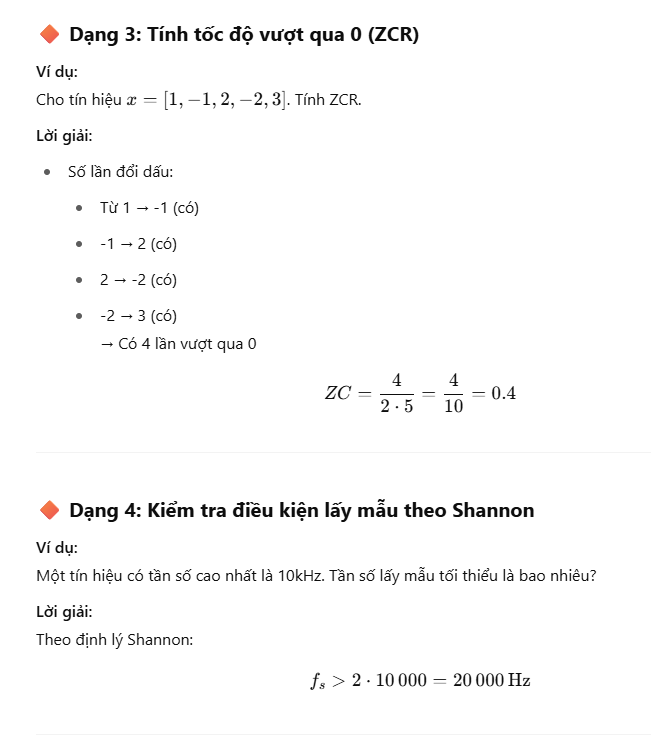
| **Từ** | **Nhãn** |
| --- | --- |
| Nguyễn Văn A | B-PER |
| làm việc | O |
| tại | O |
| VinGroup | B-ORG |
| ở | O |
| Hà Nội | B-LOC |

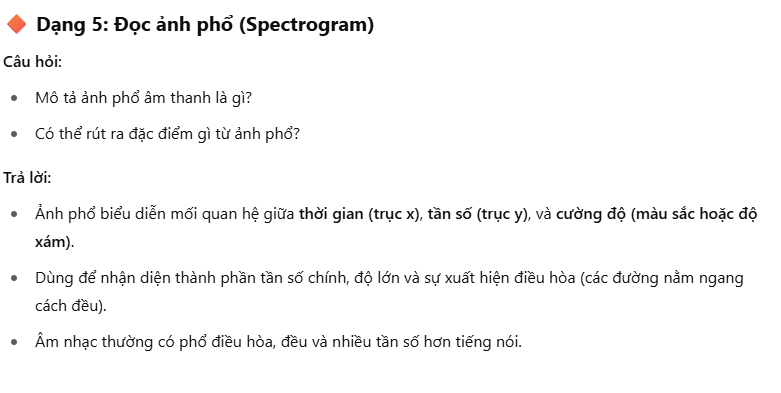
**CHƯƠNG 6: SOUND**

****

****

****

****

****

**6. So sánh tín hiệu âm nhạc và tiếng nói**

Bài tập: Nêu 2 điểm khác biệt giữa phổ âm nhạc và phổ tiếng nói.

Lời giải:

1. Băng thông: Âm nhạc có băng thông rộng hơn tiếng nói (nhiều tần số cao hơn)
2. Điều hòa: Âm nhạc thường có thành phần điều hòa rõ hơn và đều hơn tiếng nói

**7. Mô tả quy trình số hóa âm thanh**

Bài tập: Mô tả quy trình số hóa tín hiệu âm thanh từ analog sang digital.

Lời giải:

1. Lấy mẫu (Sampling): Biến tín hiệu liên tục thành rời rạc theo thời gian
2. Lượng tử hóa (Quantization): Biến giá trị biên độ thành số nguyên gần nhất
3. Mã hóa (Encoding): Biến giá trị đã lượng tử hóa thành dãy bit nhị phân

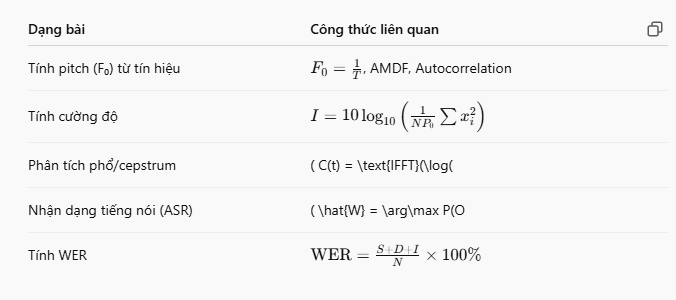
**8. Nhận diện điều hòa từ phổ tín hiệu**

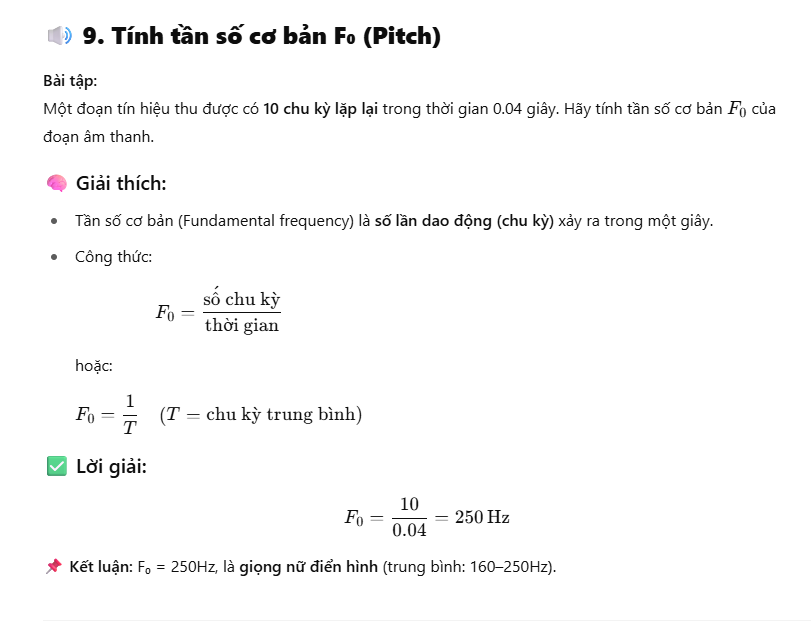
Bài tập: Một phổ tín hiệu có các đỉnh tại 200Hz, 400Hz, 600Hz, 800Hz. Tín hiệu này có phải điều hòa không? Vì sao?

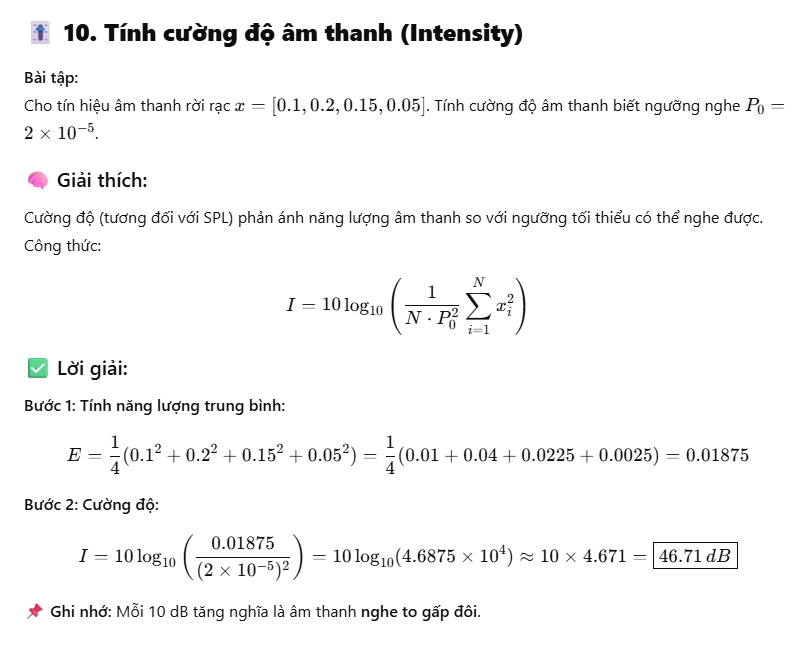
Lời giải:

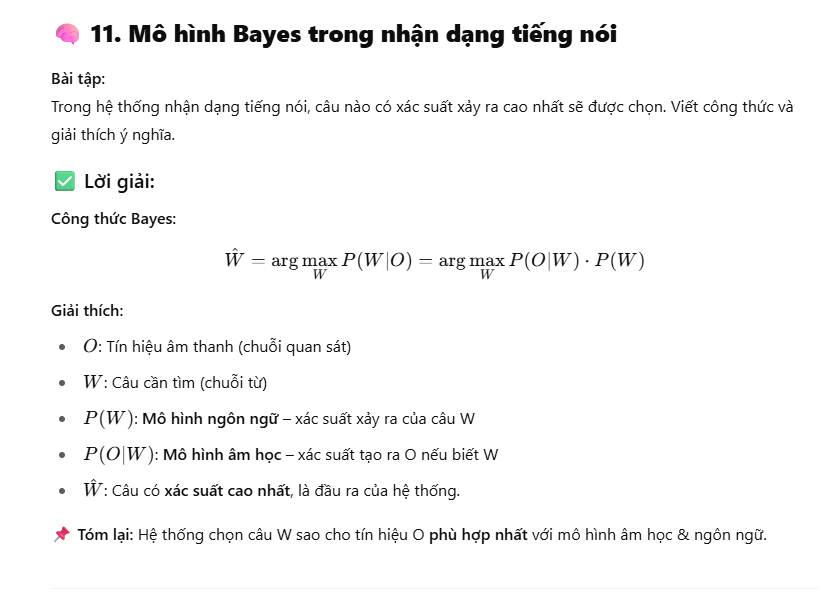
* Có tần số gốc là 200Hz
* Các đỉnh là bội số nguyên: 200, 2x200, 3x200, 4x200 → Điều hòa
* Đây là tín hiệu có cấu trúc điều hòa, đặc trưng của âm nhạc hoặc giọng hát

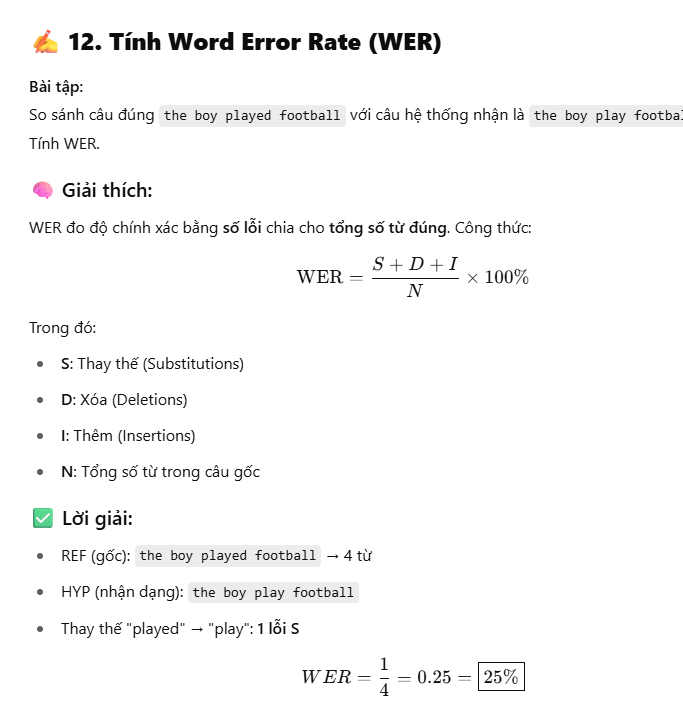
**CHƯƠNG 7: SPEECH**

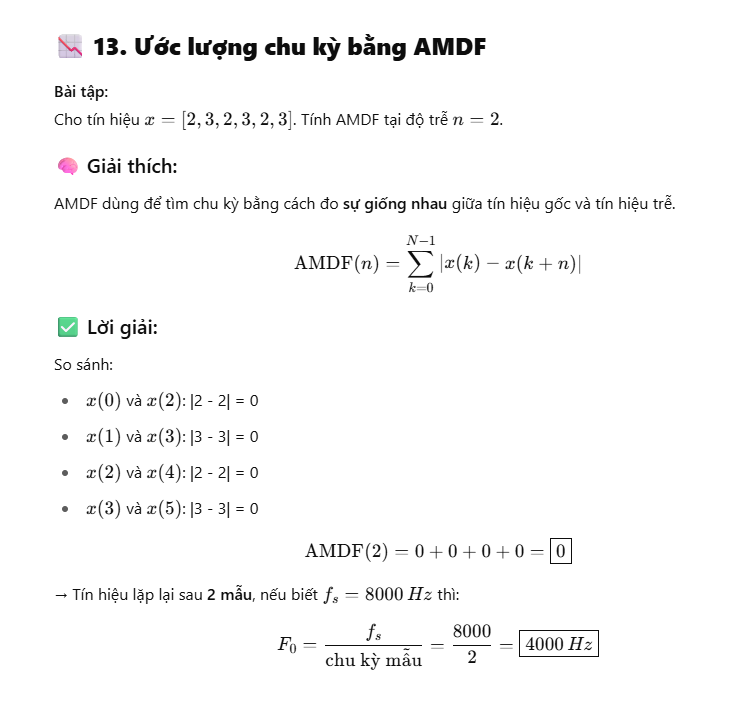
****

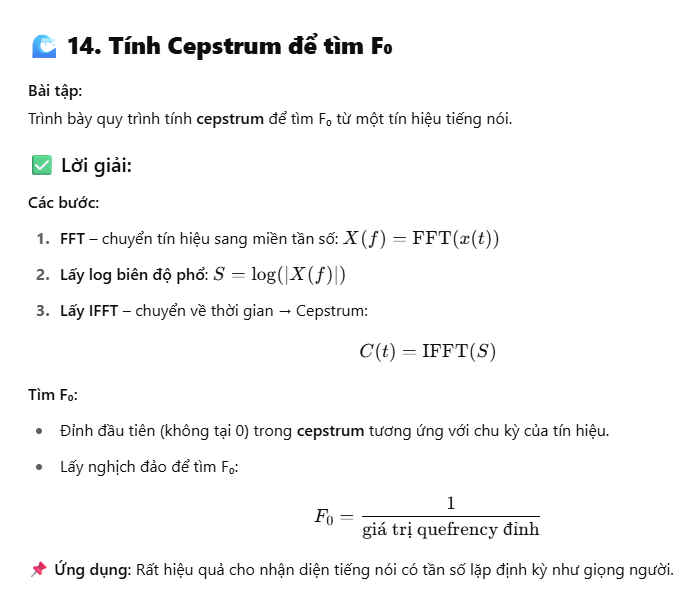
****

****

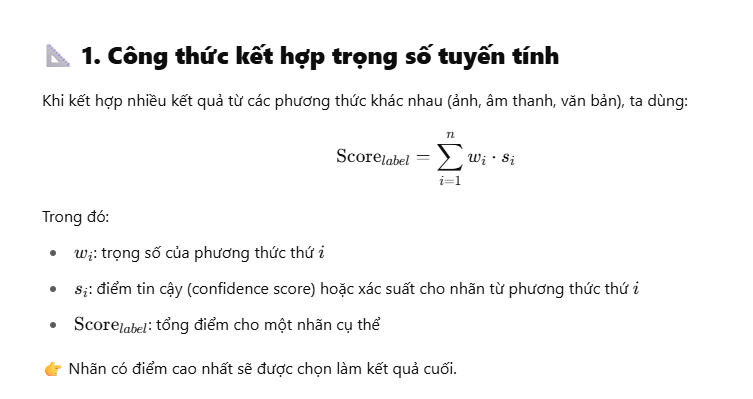
****

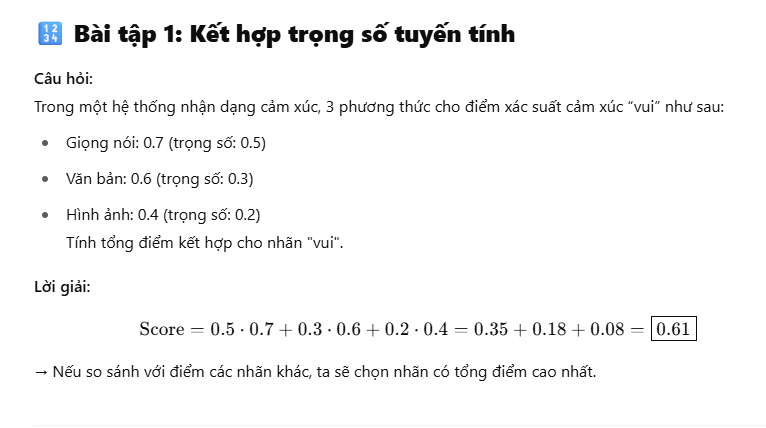
****

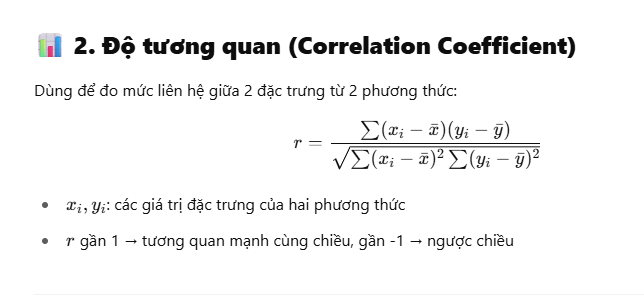
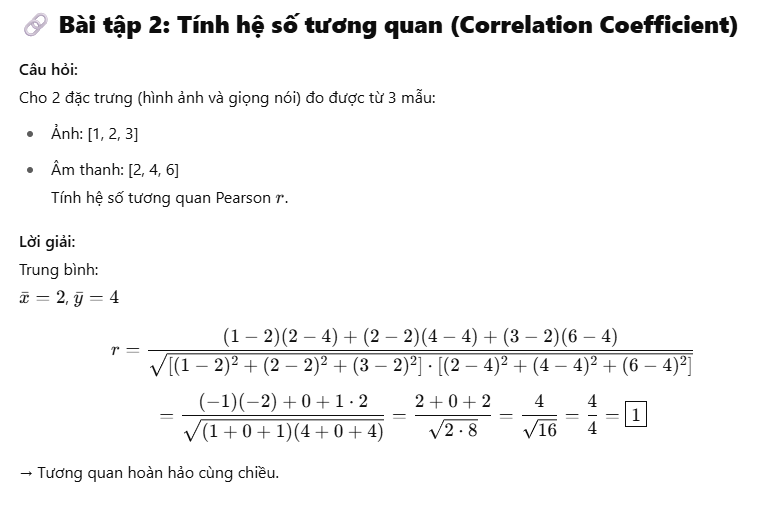
****

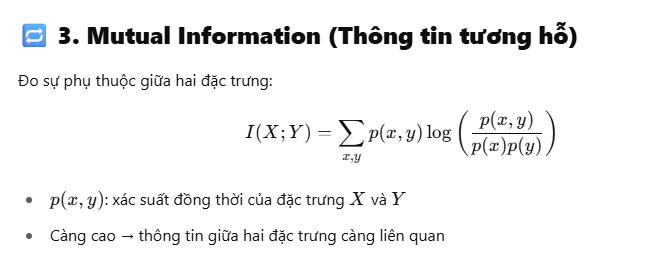
****

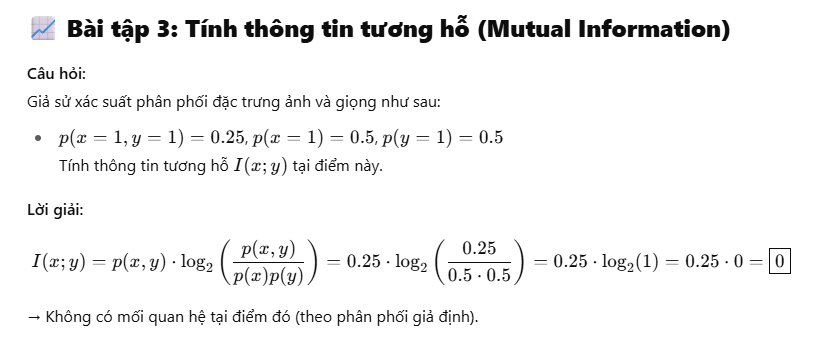
**CHƯƠNG 8**

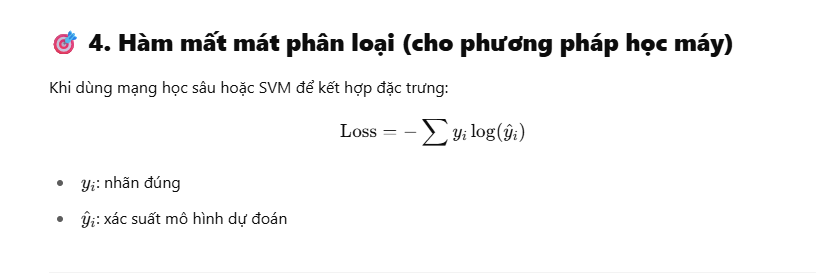
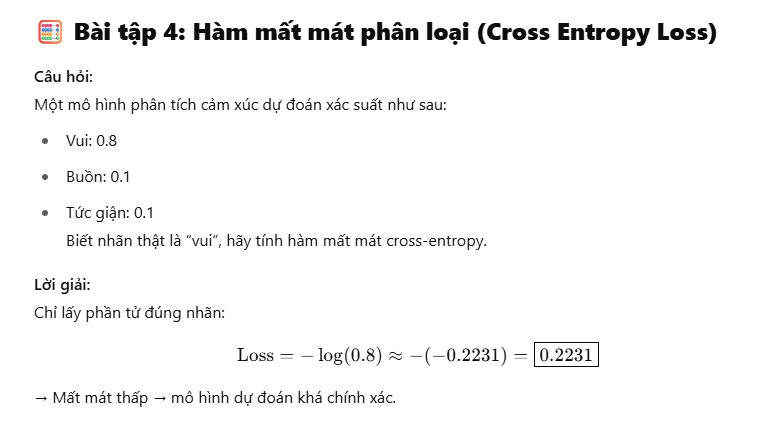
****

****

**** ****



****

**** ****