



Chứng minh rằng tổng xác suất của tất cả các chuỗi có thể từ tập từ vựng \mathcal{V} trong hai trường hợp sau:

- **Không có từ kết thúc** $\langle /s \rangle$: Mô hình ngôn ngữ có thể sinh chuỗi dài vô hạn.

$$\sum_{n=1}^{\infty} \sum_{x_{1:n}} P(x_{1:n}) = \infty$$

- **Có từ kết thúc** $\langle /s \rangle$: Mô hình ngôn ngữ phải dừng lại việc sinh chuỗi bằng từ kết thúc câu $\langle /s \rangle$.

$$\sum_{n=1}^{\infty} \sum_{x_{1:n}} P(x_{1:n}, \langle /s \rangle) = 1$$



Cho tập dữ liệu gồm nhiều văn bản thuộc các lớp $C = \{c_1, c_2, \dots, c_k\}$ và mỗi văn bản chứa các từ từ tập từ vựng \mathcal{V} . Hãy sử dụng phương pháp MLE để tính:

- **Xác suất tiên nghiệm của lớp c_j :**

$$\hat{P}(c_j) = \frac{\text{count}(c_j)}{N_{doc}}$$

Trong đó:

- ◇ $\text{count}(c_j)$: số văn bản thuộc lớp c_j .
- ◇ N_{doc} : tổng số văn bản.

- **Xác suất có điều kiện của từ w_i trong lớp c_j :**

$$\hat{P}(w_i | c_j) = \frac{\text{count}(w_i, c_j)}{\sum_{w \in \mathcal{V}} \text{count}(w, c_j)}$$

Trong đó:

- ◇ $\text{count}(w_i, c_j)$: số lần từ w_i xuất hiện trong lớp c_j .
- ◇ $\sum_{w \in \mathcal{V}} \text{count}(w, c_j)$: tổng số lần xuất hiện của tất cả các từ trong lớp c_j .