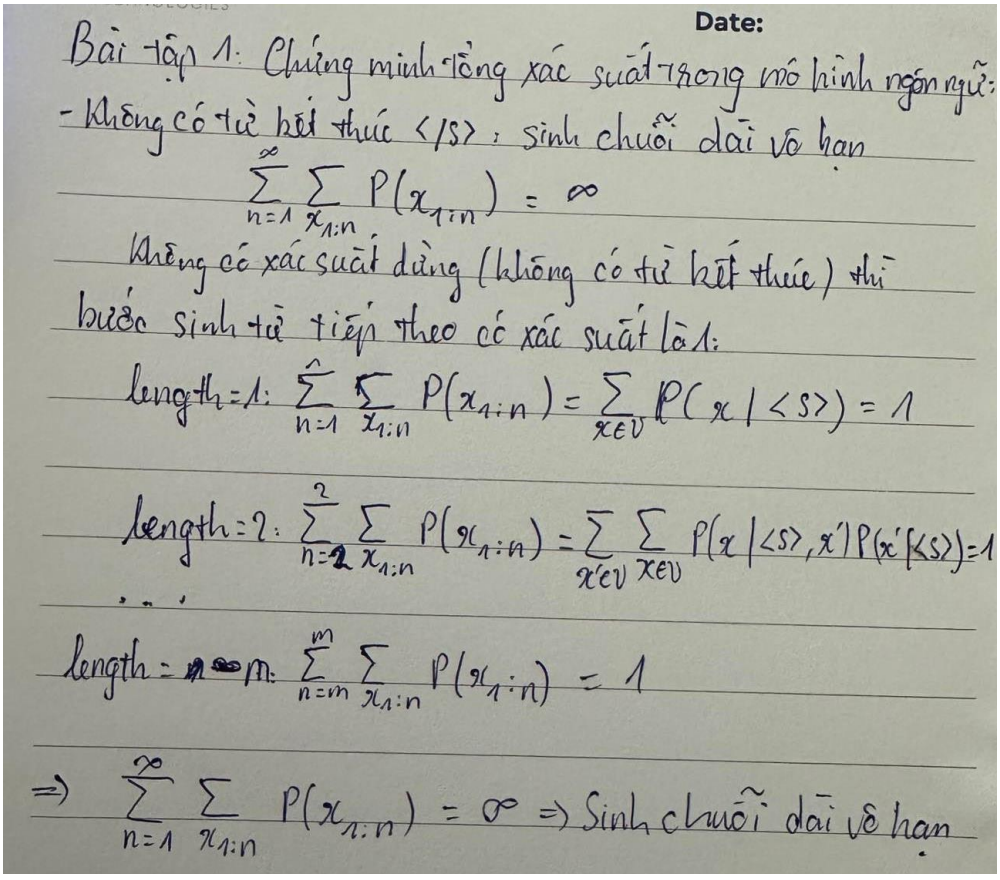
Xử Lý Ngôn Ngữ Tự Nhiên Trong Khoa Học Dữ Liệu – DS310.P11

Bài Tập Quá Trình 1

Bài 1: Chứng minh tổng xác suất trong mô hình ngôn ngữ



Date:

Bài tập 1: Chứng minh tổng xác suất trong mô hình ngôn ngữ:

- Không có từ kết thúc </s>: sinh chuỗi dài vô hạn

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

Không có xác suất dừng (không có từ kết thúc) thì bước sinh từ tiếp theo có xác suất là 1:

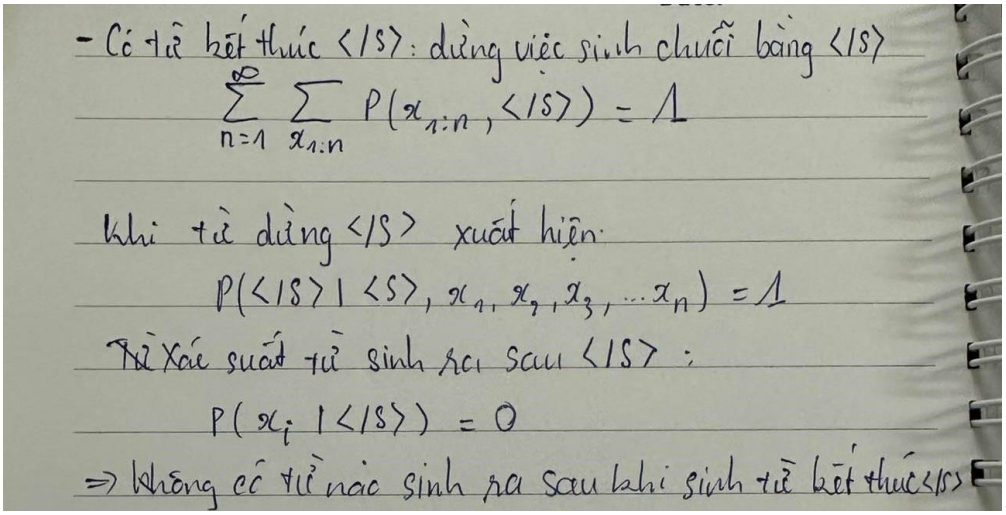
length = 1: $\sum_{n=1}^{1} \sum_{x_{1:n}} P(x_{1:n}) = \sum_{x \in V} P(x \mid <s>) = 1$

length = 2: $\sum_{n=2}^{2} \sum_{x_{1:n}} P(x_{1:n}) = \sum_{x' \in V} \sum_{x \in V} P(x \mid <s>, x') P(x' \mid <s>) = 1$

...

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

$$\Rightarrow \sum_{n=1}^{\infty} \sum_{x_{1:n}} P(x_{1:n}) = \infty \Rightarrow$$ Sinh chuỗi dài vô hạn



- Có từ kết thúc </s>: dừng việc sinh chuỗi bằng </s>

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

Khi từ dừng </s> xuất hiện:

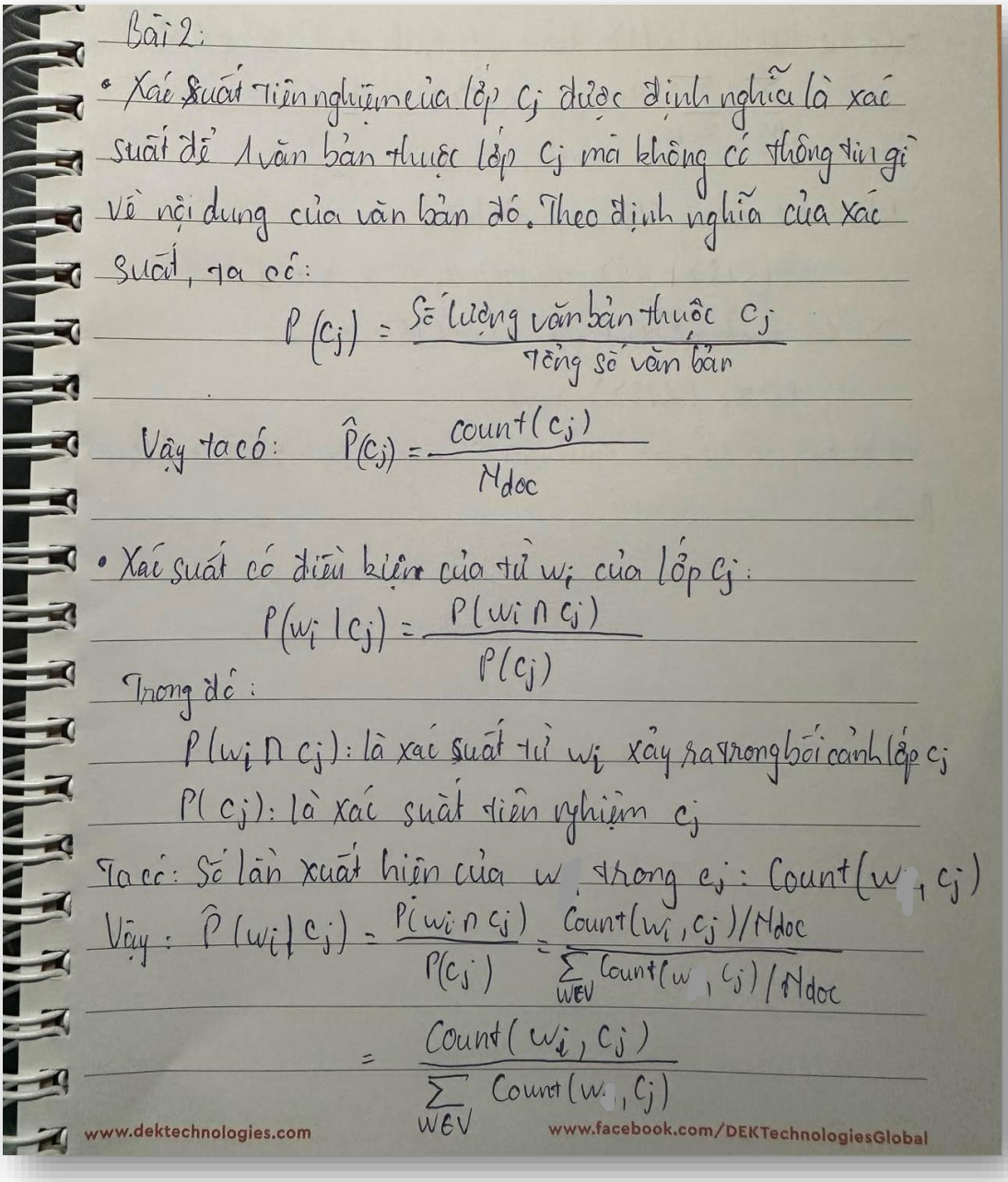
$$P(</s> \mid <s>, x_1, x_2, x_3, \ldots x_n) = 1$$

Xác suất từ sinh ra sau </s>:

$$P(x_i \mid </s>) = 0$$

⇒ Không có từ nào sinh ra sau khi sinh từ kết thúc </s>

Bài 2: Ước lượng xác suất bằng MLE



Bài 2:

- Xác suất tiên nghiệm của lớp c_j được định nghĩa là xác suất để 1 văn bản thuộc lớp c_j mà không có thông tin gì về nội dung của văn bản đó. Theo định nghĩa của xác suất, ta có:

$$P(c_j) = \frac{\text{Số lượng văn bản thuộc } c_j}{\text{Tổng số văn bản}}$$

Vậy ta có: $\hat{P}(c_j) = \frac{count(c_j)}{N_{doc}}$

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

$$P(w_i | c_j) = \frac{P(w_i \cap c_j)}{P(c_j)}$$

Trong đó:

$P(w_i \cap c_j)$: là xác suất từ w_i xảy ra trong bối cảnh lớp c_j

$P(c_j)$: là xác suất tiên nghiệm c_j

Ta có: Số lần xuất hiện của w_i trong c_j: $Count(w_i, c_j)$

Vậy: $\hat{P}(w_i | c_j) = \frac{P(w_i \cap c_j)}{P(c_j)} = \frac{Count(w_i, c_j)/N_{doc}}{\sum_{w \in V} Count(w, c_j)/N_{doc}}$

$$= \frac{Count(w_i, c_j)}{\sum_{w \in V} Count(w, c_j)}$$