ĐẠI HỌC QUỐC GIA HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC BÁCH KHOA
KHOA KHOA HỌC VÀ KỸ THUẬT MÁY TÍNH

XỬ LÝ NGÔN NGỮ TỰ NHIÊN (BT) (CO3085)

KỲ 251: LAB 3

BÀI TẬP TOÁN

Giáo viên: Bùi Khánh Vĩnh
Sinh viên: Nguyễn Mạnh Dũng 2210583 (L01)

THÀNH PHỐ HỒ CHÍ MINH, 2025

Contents

1 Bài 1

Công thức trigram:

$$P(w_i \mid w_{i-2}, w_{i-1}) = \frac{\text{count}(w_{i-2}, w_{i-1}, w_i)}{\text{count}(w_{i-2}, w_{i-1})}$$

P(I | <s>, <s>) = $\frac{2}{3} = 0.67$
$P(am \mid \text{<s>}, I) = \frac{1}{2} = 0.5$
$P(Sam \mid I, am) = \frac{1}{2} = 0.5$
$P(\text{</s>} \mid am, Sam) = \frac{1}{1} = 1$
$P(Sam \mid \text{<s>}, \text{<s>}) = \frac{1}{3} = 0.33$
$P(I \mid \text{<s>}, Sam) = \frac{1}{1} = 1$
$P(am \mid Sam, I) = \frac{1}{1} = 1$
$P(\text{</s>} \mid I, am) = \frac{1}{2} = 0.5$
$P(do \mid \text{<s>}, I) = \frac{1}{2} = 0.5$
$P(not \mid do, I) = \frac{1}{1} = 1$
$P(like \mid do, not) = \frac{1}{1} = 1$
$P(green \mid like, not) = \frac{1}{1} = 1$
$P(eggs \mid green, like) = \frac{1}{1} = 1$
$P(and \mid green, eggs) = \frac{1}{1} = 1$
$P(Sam \mid eggs, and) = \frac{1}{1} = 1$
$P(< /s > \mid and, Sam) = \frac{1}{1} = 1$

2 Bài 2

Công thức bigram:

$$P(w_1 w_2 \dots w_n) \approx \prod_i P(w_i \mid w_{i-1})$$

Xác suất	Không biến đổi Laplace	Có biến đổi Laplace
P(i $\mid < s >$)	0.19	0.19
P(want $\mid i$)	0.33	0.21
P(chinese $\mid want$)	0.0065	0.0029
P(food $\mid chinese$)	0.52	0.52
P(</s>, food)	0.40	0.40
P(câu)	$8.477 \cdot 10^{-5}$	$2.4068 \cdot 10^{-5}$

Bảng 1: Xác suất sử dụng bigram

3 Bài 3

Xác suất unsmoothed cao hơn xác suất smoothed, vì smoothing làm phẳng phân phối, giảm xác suất ở những bigram quan sát được để phân bổ cho các bigram chưa quan sát. Nếu tất cả bigram đều tồn tại (p > 0), xác suất unsmoothed sẽ cao hơn xác suất smoothed

4 Bài 4

V = 11
Số lần xuất hiện từ "am": 3
Số lần xuất hiện cặp "am Sam": 2
Với bigram có dùng add-one smoothing, xác suất là:

$$\frac{2+1}{3+11} \approx 0.2143$$

5 Bài 5

Độ dài của corpus: 25
Số lần xuất hiện từ "Sam": 4
Số lần xuất hiện từ "am": 3
Số lần xuất hiện cặp "am Sam": 2
Công thức tính xác suất:

$$P(\text{Sam} \mid \text{am}) = \lambda_1 P_{\text{bi}} + \lambda_2 P_{\text{uni}} = \tfrac{1}{2} \cdot \tfrac{2}{3} + \tfrac{1}{2} \cdot \tfrac{4}{25} = \tfrac{31}{75} \approx 0.4133.$$

6 Bài 6

Xác suất trong mô hình unigram được tính theo công thức:

$$P(w_i) = \frac{\text{số lần xuất hiện của } w_i}{\text{độ dài của corpus}}$$

P(0) = $\frac{91}{100}$
P(3) = $\frac{1}{100}$
Xác suất: $P = (0.91)^8 \times (0.01)^1$.
Vậy $PP = P^{-\frac{1}{9}} \approx 1.81$