Задание 2. Опечатки

Дано:

- Встретилось слово hodtel
- od|o встретилась 9 раз — ошибка типа insertion
- d|s встретилась 7 раз — ошибка типа substitution
- $\frac{C(hotel)}{C(hostel)} = 5$
- $\frac{C('o')}{C('s')} = 1.2$

$$P(x|w) = \begin{cases} \frac{\text{ins}[w_{i-1}, x_i]}{\text{count}[w_{i-1}]}, & \text{if insertion} \\ \frac{\text{sub}[x_i, w_i]}{\text{count}[w_i]}, & \text{if substitution} \end{cases}$$

Формулы для расчета

Какое слово более вероятно: hotel или hostel?

Решение:

Общая формула:

$$P(w, x) = P(x|w)P(w)$$

Без сглаживания:

$$P_1(hostel, hodtel) = \frac{sub[d, s]}{count[s]}P(hostel) \qquad P_2(hotel, hodtel) = \frac{ins[od, o]}{count[o]}P(hotel)$$

Тогда отношение $\frac{P_2}{P_1}$:

$$\frac{P_2}{P_1} = \frac{ins[od, o] * count[s] * P(hotel)}{sub[d, s] * count[s] * P(hostel)} = \frac{9 * 5}{7 * 1.2} = \frac{75}{14} = 5.375$$

Со сглаживанием:

$$P_1 = \frac{sub[d, s] + 1}{count[s] + Alphabet}P(hostel) \qquad P_2 = \frac{ins[od, o] + 1}{count[o] + Alphabet}P(hotel)$$

Пусть $C('s') = x, C('o') = 1.2x$. Тогда отношение $\frac{P_2}{P_1}$:

$$\frac{P_2}{P_1} = \frac{(ins[od, o] + 1) * (count[s] + Alph) * P(hotel)}{(sub[d, s] + 1) * (count[o] + Alph) * P(hostel)} = \frac{(9 + 1) * 5 * (x + Alph)}{(7 + 1) * (1.2x + Alph)} = \frac{25}{4} * \frac{x + Alph}{1.2x + Alph}$$

Рассмотрим это отношение в пределе, когда частотность появления x стремится к бесконечно большому числу, а длина алфавита символов фиксированная:

$$\lim_{x \to +\infty} \frac{25}{4} * \frac{x(1 + \frac{Alphabet}{x})}{x(1.2 + \frac{Alphabet}{x})} = \frac{25}{4} \lim_{x \to +\infty} \frac{1}{1.2} = 5.208(3)$$