# Tecnicas N-grams

Adrian Pineda Sánchez
A00834710

## 0 Absolute Discounting

- **Problematica**

Aborda el problema de que muchos N-grams (secuencia de palabras) aparecen pocas o ninguna vez en los datos de entrenamiento. Sin suavizado, las probabilidades se vuelven 0, lo que afecta la capacidad del modelo para manejar textos nuevos. "Absolute Discounting" corrige esto restando una cantidad fija de cada N-grama observado y redistribuyendo esa masa de probabilidad a N-grams no observados.

- **Ejemplo:** Un bigrama "gato negro" que aparece 5 veces en el corpus, sin embargo el bigrama "gato blanco" no aparece. Sin suavizado su probabilidad seria 0; Absolute Discounting podria ajustar esto para el modelo.

- **Expresión Matemática:**

$$P_{abs}(w_i, w_{i-1}) = \frac{\max(c(w_{i-1}, w_i) - \mathcal{D}, 0)}{\sum_{w'} c(w_{i-1}, w')} + \alpha \cdot P_{abs}(w_i)$$

- $c(w_{i-1}, w_i)$ es el conteo del bigrama
- $\mathcal{D}$ es el valor fijo de descuento.
- $\alpha$ es una constante de normalización que garantiza que las probabilidades sumen 1.
- "$\max(c(w_{i-1}, w_i) - \mathcal{D}, 0)$ aseguro que no se generen probabilidades negativas.

# ② Knesser - Ney Smoothing

- **Problemática**

Soluciona un problema similar pero más específico que el anterior. En lugar de solo ajustar las probabilidades de los N-grams observados, considera el contexto en que una palabra aparece. Esto mejora los resultados al evitar que las palabras comunes en ciertos contextos tengan demasiada influencia en otros. Por ejemplo, en el corpus "San Francisco", "Francisco" aparecerá muchas veces, pero no siempre es una buena predicción para otros contextos donde se espera una palabra como "ciudad"

- **Ejemplo:**

"Necesito mis lentes". Un modelo de suavizado sin contexto podría asignar una alta probabilidad a "San Francisco" en lugar de "lentes", porque "Francisco" es una palabra frecuente después de "San". Knesser - Ney ajusta este comportamiento al considerar cuantos contextos únicos preceden a "lentes", otorgando una mejor estimación.

**Expresión matemática:**

$$P_{KN}(w_i / w_{i-1}) = \frac{\max(c(w_{i-1} w_i) - \mathcal{D}, 0)}{c(w_{i-1})} + \lambda \cdot P_{cont}(w_i)$$

$$P_{cont}(w_i) = \frac{|\{w_{i-1} : c(w_{i-1}, w_i) > 0\}|}{|\{w_i, w_{i+1}) : c(w_i, w_{i+1}) > 0\}|}$$

- $c(w_{i-1} w_i)$ es el conteo del bigrama.
- $\mathcal{D}$ es el valor del descuento.
- $P_{cont}(w_i)$ probabilidad de continuación, una palabra $w_i$ aparezca en contextos diferentes
- $\lambda$ constante de normalización que ajusta el peso del término de probabilidad de continuación.