María Fernanda Pérez Ruiz A01742102

DÍA | MES | AÑO

Las técnicas de N-gramas buscan estimar la probabilidad de secuencias de palabras basadas en datos de entrenamiento.

>> Absolute discounting: Reduce una cantidad fija, D, de la cuenta de los N-gramas no observados. Esto resuelve el problema de asignar probabilidad a secuencias que no aparecen en el corpus.
Su formula (expresión matematica involucrada):

$$P_{AD}(w_i | w_{i-n+1}^{i-1}) = \frac{\max(C(w_{i-n+1}^{i}) - D, 0)}{C(w_{i-n+1}^{i-1})} + \lambda P_{backoff}(w_i | w_{i-n}^{i-1})$$

donde C es la frecuencia observada y
λ ajusta la redistribución de la probabilidad.

>> Smoothing Kneser-Ney: Extiende el descuento absoluto incorporando información sobre el número de contextos únicos en los que una palabra ha aparecido, resolviendo mejor la asignación de probabilidad a palabras raras o no vistas.
Su formula (expresión matematica involucrada)

$$P_{KN}(w_i | w_{i-n+1}^{i-1}) = \max(C(w_{i-n+1}^{i}) + D, 0) / C(w_{i-n+1}^{i-1}) + \lambda P_{KN}(w_i | w_{i-n+2}^{i-1})$$

Ambas técnicas resuelven la problemática de asignar probabilidades a secuencias no vistas en el corpus, evitando probabilidades nulas.

Un ejemplo de Absolute discounting:

Suponiendo que tenemos los siguientes trigrama y sus conteos en un corpus:

- "el gato come": 4
- "gato come pescado": 1
- "come pescado fresco": 0 (no visto)

Con absolute discounting (D = 0.75), la probabilidad del trigrama "gato come pescado" sería:

$$P_{AD}(\text{pescado} \mid \text{gato come}) = \frac{\max(1-0.75, 0)}{C(\text{gato come})} + \lambda P_{\text{backoff}}(\text{pescado} \mid \text{come})$$

Redistribuimos parte de la probabilidad de los trigramas vistos a los no vistos, como "come pescado fresco".

Y al hacerlo para Kneser, Ney, además del descuento, se consideraría cuántos contextos distintos preceden a "pescado", mejorando la probabilidad para palabras raras en contextos conocidos.

Este enfoque evita frases no vistas como "come pescado fresco" tengan probabilidad cero.