Laboratorio: Modelos de Lenguaje con N-gramas, Espacidad y Suavizado

Procesamiento de Lenguaje Natural

31 de julio de 2025

En esta actividad explorará la construcción de modelos de lenguaje basados en n-gramas, el problema de la **espacidad de datos** y el impacto de diferentes técnicas de **suavizado**. Se trabajará con el corpus *cess esp* de NLTK.

Criterios de evaluación

1. **Preprocesamiento del corpus**: Tokenice el texto, conviértalo a minúsculas y elimine signos de puntuación si es necesario. Justifique cualquier decisión de limpieza adicional.

2. **Construcción de modelos de n-gramas**: Genere modelos para unigramas, bigramas y trigramas. Calcule las probabilidades MLE:

$$P(w_n \mid w_{n-1}, \ldots, w_{n-k+1}) = \frac{C(w_{n-k+1}, \ldots, w_n)}{C(w_{n-k+1}, \ldots, w_{n-1})}$$

   para $k = 1, 2, 3$.

3. **Análisis de espacidad**: Calcule el porcentaje de n-gramas del conjunto de prueba que no aparecen en el conjunto de entrenamiento para cada valor de n. Discuta por qué ocurre este fenómeno.

4. **Implementación de suavizado**: Implemente y compare al menos tres métodos:

   a) Suavizado de Laplace (Add-1).

   b) Interpolación lineal

   c) Kneser-Ney (puede usar librerías como `nltk.lm` para facilitar la implementación).

5. **Cálculo de perplejidad**: Implemente una función para calcular la perplejidad de una secuencia dada. Evalúe el conjunto de prueba con cada técnica de suavizado y presente una tabla comparativa.

6. **Generación de texto**: Utilizando cada modelo entrenado, genere texto a partir de una palabra inicial. Compare la coherencia de los textos para distintos valores de n y para diferentes técnicas de suavizado.

7. **Discusión final**: Analice:

    - El impacto del suavizado en la perplejidad.
    - Cómo cambia la coherencia del texto generado según n y el suavizado.
    - Qué tanto influye la espacidad en el rendimiento del modelo.

Recomendación: Usar `NLTK` para la generación de n-gramas y suavizados.

Formato de entrega: La actividad debe realizarse en un notebook (`.ipynb`) y entregarse en formato PDF en la plataforma.