Laboratorio 5: Representaciones Vectoriales de Texto (PPMI, TF-IDF y Word2Vec)

Procesamiento de Lenguaje Natural

14 de agosto de 2025

En esta actividad explorará diferentes formas de representar texto: **PPMI** (Positive Pointwise Mutual Information), **TF-IDF** y **Word2Vec**. Se evaluará su impacto en tareas de clasificación y similitud de documentos.

Criterios de evaluación

1. **Preprocesamiento del corpus**: Seleccione un corpus (utilice el dataset `20newsgroups` de `scikit-learn` importando todas las categorías de politics y la de autos). Realice tokenización, pase a minúsculas y elimine caracteres no alfabéticos. Justifique sus decisiones de limpieza.

2. **Construcción de representación TF-IDF**: Utilice `TfidfVectorizer` de `scikit-learn` para construir representaciones de los documentos. Analice:

   - Palabras con mayor peso en algunos documentos.
   - Limitaciones de TF-IDF respecto a la semántica.

3. **Construcción de representación PPMI**:

   a) Construya una matriz de co-ocurrencia palabra-contexto con ventana deslizante.

   b) Calcule la matriz PPMI:

   $$PPMI(w, c) = \text{máx}\left(0, \log \frac{P(w, c)}{P(w) \cdot P(c)}\right)$$

   c) Discuta ventajas y desventajas de PPMI respecto a TF-IDF.

4. **Construcción de representación Word2Vec**: Entrene un modelo Word2Vec (puede usar `gensim`). Genere embeddings de documentos como el promedio de los embeddings de sus palabras. Analice:

   - Palabras más cercanas y lejanas en el espacio vectorial.

- Diferencias con TF-IDF y PPMI.

5. **Evaluación comparativa**: Con cada representación (TF-IDF, PPMI, Word2Vec), entrene un clasificador simple (ej. Regresión Logística). Compare las precisiones obtenidas y presente una tabla comparativa.

6. **Discusión final**: Reflexione sobre:

   - Cómo cada representación captura (o no) relaciones semánticas.
   - Escenarios donde cada técnica es más útil.
   - Limitaciones prácticas (memoria, tiempo de cómputo, interpretabilidad).

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