Práctica 3

Procesamiento de Lenguaje Natural
Facultad de Ingeniería, UNAM

Elegir entre uno de los dos siguientes proyectos:

1. **Embedding documento-palabra:** Crear vectores continuos a partir del Corpus 1 ('CorpusDocs'). Tomar en consideración los siguientes puntos:
   - La entrada de la red neuronal son los tipos en todos los documentos. Debe hacerse un stemming.
   - La salida de la red son los documentos; específicamente, la probabilidad de un documento dada la palabra.
   - Utilizar una arquitectura similar a la de Word2Vec, únicamente con las variaciones indicadas ¿Cómo sería la función de error?
   - En este caso, ¿qué representa cada una de las amtrices de peso?
2. **Embeding de gráficas:** Crear vectores continuas a partir del Corpus 2 ('graph.txt'). Tomar en cuenta los siguientes puntos:
   - El documento representa una gráfica que relaciona palabras en una lengua conectadas a palabras en otra lengua. El formato general es:

     ```
     w1
         w2
         w3
         ...
     w4
         w5
         ...
     ```

     En este caso, la palabra w1 está conectada con las palabras w2 y w3, mientras que la palabra w4 está conectada con w5.
   - El método para aprender los vectores debe ser similar a Word2Vec.
   - La función de error corresponde a las conexiones ($1 - p$ si están conectadas, $0 - p$ si no lo están).
   - La entrada y la salida de la red son todas las palabras. Así, el modelo es más similar a Word2Vec, excepto por la función de error.