Práctica 2

Procesamiento de Lenguaje Natural
Facultad de Ingeniería, UNAM

A partir del corpus seleccionado en la tarea anterior realizar un modelo del lenguaje neuronal con base en la arquitectura propuesta por Bengio (2003). El corpus ya debe estar preprocesado. Síganse los siguientes pasos:

1. Trabajar con las palabras stemmizadas.
2. Insertar símbolos de inicio y final de cadena.
3. Obtener los bigramas que aparecen en el texto (indexar numéricamente).
4. Entrenar con los bigramas la red neuronal y obtener los valores para los hiperparámetros. Tomar de 100 unidades para la primera capa oculta (capa lineal) y 300 para la segunda capa oculta (capa con tanh).
5. Evaluar el modelo (con Entropía y/o Perplejidad).
6. Calcular la proabilidad de 5 oraciones no vistas en el entrenamiento.
7. Guardar los vectores de la capa de embedding asociados a las palabras (por ejemplo como un diccionario) (se usarán en la siguiente tarea).

Puntos a evluar

1. Entrega a tiempo del trabajo.
2. Código bien realizado y, principalmente, comentado adecuadamente.
3. Proprocesamiento y separación adecuada de los datos (entrenamiento 70 % y evaluación 30 %).
4. Haber manejado los casos problemáticos: 1) palabras desconocidas; 2) manejo de los stems y sus palabras.
5. Calculo de la probabilidad de las oraciones del punto 7.