T5: Modelos de Lenguaje Neuronales

CIMAT, Procesamiento de Lenguaje Natural, Ciencias de la Computación

Profesor: Dr. Adrián Pastor López Monroy

Entregar: Martes 19 de Marzo de 2024 antes de las 23:59:59

Instrucciones

Realiza los siguientes puntos en un notebook de Python *lo mejor organizado y claro posible.* Ponga su nombre al notebook (e.g., adrian_pastor_lopez_monroy.ipynb) y también en la primera celda del notebook junto con el número de Tarea. Sube al classroom el notebook como un archivo, que deberá haber sido ejecutado en tú máquina y mostrar el resultado en las celdas.

Se vale pedir ayuda y/o copiar con atribución entre los miembros de la clase y apegándose estrictamente a los siguientes puntos:

1. Del total de actividades que se solicitan hacer (3 con valor para esta tarea) solo puedes pedir ayuda/copiar en un total de **una**.
2. Para los puntos dónde se pide ayuda, brevemente escribe en qué pediste ayuda y a quién.
3. Si tuviste que reusar alguna parte de código que no es tuyo, deja claro dos cosas: 1) brevemente porque tuviste dificultad para hacerlo, 2) cómo lo solucionó tu compañero.

1 Modelos de Lenguaje Neuronales

1. (50pts) Con base en la implementación mostrada en las prácticas del NLM, construya un modelo de lenguaje neuronal a nivel de carácter. Tomé en cuenta secuencias de tamaño 6 o más para el modelo, es decir hasta 5 caracteres o más en el contexto. Ponga al modelo a generar texto 3 veces, con un máximo de 300 caracteres. Escriba 5 ejemplos de oraciones y mídales el likelihood. Escriba un ejemplo de estructura morfológica (permutaciones con caracteres) similar al de estructura sintáctica del profesor con 5 o más caracteres de su gusto (e.g., "ando "). Calcule la perplejidad del modelo sobre los datos val.
2. (50pts) Con base en la implementación mostrada en clase, construya un modelo de lenguaje neuronal a nivel de palabra, pero preinicializado con los embeddings proporcionados. Tomé en cuenta secuencias de tamaño 4 para el modelo, es decir hasta 3 palabras en el contexto. Después de haber entrenado el modelo, recupere las 10 palabras más similares a tres palabras de su gusto dadas. Ponga al modelo a generar texto a partir de tres secuencias de inicio de su gusto. Escriba 5 ejemplos de oraciones y mídales el likelihood. Proponga un ejemplo para ver estructuras sintácticas (permutaciones de palabras de alguna oración) buenas usando el likelihood a partir de una oración que usted proponga. Calcule la perplejidad del modelo sobre los datos val. Compárelo con la perplejidad del modelo de lenguaje sin embeddings preentrenados (el visto en clase). DISCUTA BREVEMENTE.
3. (OPCIONAL: 30pts extra en ESTA tarea, calificación máxima para promediar con otras tareas: 130) A partir del modelo anterior haga un modelo de lenguaje que integre una conexión directa de la capa de embeddings hacía la salida, justo como lo proponía Bengio. Discuta sobre las diferencias en el proceso de entrenamiento y la perplejidad respecto al modelo anterior y el visto en clase.