UNIVERSIDADE DA CORUÑA

Grado en Inteligencia Artificial – Fundamentos de Procesamiento de Lenguaje Natural

Curso 2024-2025

Práctica PLN – Modelos neuronales para representación vectorial de palabras

En esta práctica, vamos a desarrollar dos modelos de redes neuronales diseñados para crear representaciones vectoriales de palabras, comúnmente referidas como *word embeddings*, utilizando grandes corpus de texto en lenguaje natural. Estas representaciones vectoriales son cruciales para diversas aplicaciones en el procesamiento del lenguaje natural (PLN) y representan la base sobre la cual se construyen los modelos de lenguaje modernos, como GPT. Estos modelos fueron pioneros en la capacidad de capturar relaciones semánticas entre palabras mediante vectores, un principio que sigue siendo fundamental en los modelos de lenguaje de la actualidad, aunque operando a una escala mucho mayor. Específicamente, implementaremos dos metodologías diferentes para la generación de estos vectores: la primera, centrada en predecir una palabra basándose en el contexto en el que aparece, y la segunda, que utiliza una palabra para predecir el contexto circundante.

Para más detalles sobre los detalles de la práctica, tenéis también disponible el documento guia_pln.pdf

Conjunto de datos: Para entrenar los modelos, se suministran varios conjuntos de datos junto con los materiales de esta práctica, disponibles en la carpeta denominada "datasets". Estos conjuntos de datos están en formato de texto plano y ya han sido preprocesados. Será necesario generar las muestras de entrenamiento utilizando estos datos. En particular, se requiere seleccionar al menos uno de los tres libros clásicos incluidos en los materiales de la práctica.

Modelos:

1. **Modelo basado en la predicción de palabras.** Predice una palabra objetivo a partir de un contexto circundante de palabras. La generación de muestras de entrenamiento implica el uso de una ventana deslizante que recorre todo texto. Específicamente, la ventana incluye un número determinado de palabras antes y después de la palabra objetivo, excluyendo la propia palabra objetivo. Para crear las muestras de entrenamiento, la ventana se desplaza a lo largo del texto, palabra por palabra. En cada posición, las palabras dentro de la ventana, excepto la palabra central (la salida del modelo), se utilizan como entrada al modelo, y la tarea del modelo es predecir la palabra central basándose en este contexto. Por ejemplo, si se elige una ventana de 5 palabras (2 palabras antes y 2 palabras después de la palabra objetivo), para cada palabra del texto, se toman las 2 palabras anteriores y las 2 posteriores como el contexto de entrada, y se intenta predecir la palabra en el centro de la ventana.

2. **Modelo basado en la predicción de contextos.** Predice el contexto de palabras circundantes a partir de una palabra objetivo. A diferencia del modelo previo, este enfoque acepta una única palabra como entrada y su objetivo es predecir las palabras que usualmente aparecen a su alrededor, es decir, su contexto. Para generar las muestras de entrenamiento en este modelo, se selecciona una palabra del texto como la palabra objetivo. Luego, para cada palabra objetivo, se crean pares de muestras utilizando la palabra objetivo y cada una de las

palabras de su contexto cercano, dentro de un rango predefinido. Por ejemplo, si se elige un rango de contexto de 2 palabras, para la palabra objetivo "gato" en la frase "el gato negro duerme", se generarían las siguientes muestras de entrenamiento: ("gato", "el"), ("gato", "negro"), ("gato", "duerme"). Cada muestra consiste en la palabra objetivo y una palabra de contexto, y el modelo se entrena para predecir la probabilidad de que estas palabras de contexto aparezcan cerca de la palabra objetivo.

Arquitectura: Para ambos enfoques, se debe implementar una arquitectura de red neuronal que incluya (más detalles se incluyen en el documento guia_pln.pdf):

1. Una capa de entrada que acepte la(s) palabra(s) como entrada.
2. Una capa de Embeddings que asocia cada ID de palabra a un vector, y que se irá actualizando durante el entrenamiento. Los valores de esta capa serán los *word embeddings* de cada término.
3. Una o varias capas ocultas que procesen estas entradas.
4. Una capa de salida que genere las predicciones para cada modelo

Se pide experimentar con al menos un tamaño de ventana de 5 (es decir, una palabra central, sus dos palabras anteriores y sus dos palabras posteriores). Según el libro seleccionado, se incluye con los materiales de la práctica una lista mínima de palabras que deberán ser visualizadas y comparadas, según lo indicado a continuación. También se valorará la inclusión de experimentos adicionales en la memoria, siempre que aporten valor y queden a criterio del alumno. Algunas opciones podrían ser considerar distintos tamaños de ventana, diferentes tamaños de embeddings, la visualización de palabras adicionales o el uso de datasets adicionales. La motivación del experimento deberá estar correctamente justificada.

Funcionalidades Requeridas:

1. Entrenar los modelos utilizando un conjunto de entrenamiento y validar los modelos con un conjunto de desarrollo para ajuste de hiperparámetros.
2. Análisis cualitativo de los resultados:
   a. Similitud Semántica: Para las palabras objetivo, se pide calcular la similitud semántica de cada una de esas palabras con respecto al resto de palabras del vocabulario. Para ello, se utilizarán métricas específicas, en concreto la similitud del coseno, que compara los vectores generados para cada palabra y determinan cómo de cercanas semánticamente están dos palabras. Para cada una de las palabras objetivo, han de mostrarse las 10 palabras más cercana semánticamente, y deberéis discutir en la memoria los resultados más relevantes que observéis.
   b. Visualización mediante t-SNE: Para comprender visualmente cómo se agrupan las palabras según su semántica, se debe emplear la técnica de reducción de dimensionalidad t-SNE (t-Distributed Stochastic Neighbor Embedding). Esta herramienta permite representar los vectores de palabras en un espacio bidimensional, facilitando la identificación de grupos de palabras con significados similares o relacionados. La visualización resultante será un gráfico que muestre estos agrupamientos. Se proporciona un código de ejemplo junto con los materiales de esta práctica. En concreto, se pide visualizar las palabras objetivo utilizando para ello las funciones visualize_tsne_embeddings y visualize_all_tsne_embeddings. Para el caso de visualize_tsne_embeddings, debéis mostrar dos gráficas: (i) la visualización de las palabras objetivo antes de entrenar la red y (ii) la visualización de las palabras objetivo después de

haber entrenado la misma. Ello permitirá descubrir visualmente las relaciones semánticas que ha aprendido la red.

Para la obtención de una nota superior a un 9, como parte de la práctica se pide también entrenar un modelo sobre el dataset text8 que corresponde a los primeros 100MB de la Wikipedia en inglés. Se valorará la inclusión y análisis de resultados sobre este corpus. Debido a que se trata de un conjunto de datos más grande, el entrenamiento puede tardar algunas horas en ejecutarse. Será suficiente con entrenarlo una única iteración y llegará con entrenar únicamente el primer modelo propuesto, el basado en predicción de palabras.

Entrega

Incluir un manual breve que detalle cómo ejecutar el código para entrenar y evaluar los modelos, un análisis de los resultados obtenidos, y una discusión sobre las diferencias observadas entre las dos implementaciones y sus ventajas e inconvenientes. La discusión no debe exceder las 3 páginas, pero se pueden usar páginas adicionales para incluir imágenes de gráficas y visualizaciones, etc.

La práctica se realizará en grupos de tres personas, respetando los formados al inicio de la asignatura.

La práctica deberá subirse al campus virtual, en el apartado habilitado para ello, como muy tarde el 4 de abril a las 23:59 horas. Únicamente un miembro del grupo debe subir la práctica.

La defensa (obligatoria) tendrá lugar durante la última semana de prácticas (del 7 al 10 de abril). Los estudiantes que no acudan a dicha defensa suspenderán la práctica con una nota de 0 puntos.