Máster en Inteligencia Artificial

PROCESAMIENTO DE LENGUAJE NATURAL

Refuerzo Actividad 2

Yeray Mezquita

Objetivos de la actividad

- Interiorizar le uso de word embeddings y transformers.
- Analizar, de forma empírica, el desempeño de diferentes modelos.
- Transmitir resultados.

Pregunta 1 (1 punto)

Utilizando el tokenizador de spacy, que ya conoces, calcula el número promedio de tokens de una muestra de 15 ficheros de la categoría «com.graphics». Indica el código utilizado y el resultado obtenido.

- Cuando se habla de muestras, se deben tomar de forma aleatoria.
- Importante: directorio com.graphics.
- Total de tokens entre el total de ficheros.

Pregunta 2 (1 punto)

El código proporcionado lee los ficheros uno a uno y, antes de generar el catálogo de datos de entrenamiento y validación, descarta las diez primeras líneas de cada fichero. ¿Cuál es el trozo de código en el que se realiza dicho descarte?, ¿por qué crees que se descartan dichas líneas?, ¿por qué diez y no otro número?

- lines = content.split("\n") lines = lines[10:]
- Se descartan porque son metadatos.
- Análisis de dónde estaba el límite de los metadatos para saber con exactitud porqué es diez y no otro número.

Pregunta 3 (1 punto)

¿Qué se controla con el parámetro validation_split?, ¿por qué se ha elegido ese valor?, ¿qué ocurre si lo modificas?

- Es el porcentaje que indica cómo se va a dividir el conjunto de datos entre entrenamiento y validación.
- Normalmente, es una convención para asegurar una buena evaluación sin sobreentrenamiento.
- Mayor validation_split:
  - Menos datos de entrenamiento: Puede reducir la capacidad del modelo.
  - Más datos de validación: Evaluación más robusta.
- Menor validation_split:
  - Más datos de entrenamiento: Mejor capacidad del modelo.
  - Menos datos de validación: Evaluación menos confiable.

Pregunta 4 (1 punto)

Imprime por pantalla un ejemplo (es decir, un elemento del array) de train_samples, val_samples, train_labels y val_labels. A tenor de las etiquetas que se utilizan, ¿qué tarea crees que se está intentando entrenar?

- Uso de una observación aleatoria de cada tipo.
- Clasificación del texto a un contexto.

Pregunta 5 (1 punto)

Con output_sequence_length se establece un tamaño fijo para la salida de Vectorizer. ¿Por qué se necesita un tamaño fijo y por qué se ha elegido el valor 200?

- Las RRNN necesitan una cantidad de datos fija para la estructura de la red.
- Uniformidad de entrada a los modelos.
- 200 tokens suelen ser suficientes para capturar el contexto significativo en muchos textos.
- Menos tokens pueden perder información crucial; más tokens pueden introducir ruido y aumentar el costo computacional.

Pregunta 6 (1.5 puntos)

Indica cuál es la precisión del modelo en el conjunto de datos de entrenamiento y en el conjunto de datos de validación. ¿Qué interpretación puedes dar? Haz, en este punto, un análisis comparativo de los dos modelos ejecutados

- Mostrar la precisión (accuracy) tanto en evaluación como en entrenamiento.
  - Acc
  - Val_acc
- Ser críticos con la interpretación:
  - Siempre decir un porqué.

Pregunta 7 (1 punto)

En la parte final del código se hace un análisis cualitativo de la salida. Explica el funcionamiento de este análisis e interpreta los resultados. Haz también, en este punto, un análisis comparativo de los dos modelos ejecutados.

- Comparativa de las predicciones.
- Comparativa del tiempo de ejecución.
- Evaluar si cuando fallan, es por algún tipo de ambigüedad en el texto o limitación del modelo.

Pregunta 8 (1.5 puntos)

Explica algunas de las limitaciones que puedes encontrar al modelo entrenado.

- Dependencia del dataset.
- Generalización.
- Requisitos computacionales.
- Idioma.
- Contexto y longitud del texto.
- Ambigüedad y subjetividad.
- Interpretabilidad.

Pregunta 9 (1 punto)

¿Qué sería necesario para que este modelo pueda interpretar textos en español?

- Utilizar datasets y embeddings específicos para el español.
- Ajustar el preprocesamiento y la tokenización al idioma español.
- Emplear modelos de lenguaje preentrenados en español y realizar fine-tuning adecuado.

Criterios de evaluación

Laboratorio: word embedings y transformers para clasificación de texto	Descripción	Puntuación máxima (puntos)	Peso %
Pregunta 1	La respuesta es válida y está bien argumentada.	1	10%
Pregunta 2	La respuesta es válida y está bien argumentada.	1	10%
Pregunta 3	La respuesta es válida y está bien argumentada.	1	10%
Pregunta 4	La respuesta es válida y está bien argumentada.	1	10%
Pregunta 5	La respuesta es válida y está bien argumentada.	1	10%
Pregunta 6	La respuesta es válida y está bien argumentada.	1.5	15%
Pregunta 7	La respuesta es válida y está bien argumentada.	1	10%
Pregunta 8	La respuesta es válida y está bien argumentada.	1.5	15%
Pregunta 9	La respuesta es válida y está bien argumentada.	1	10%
		10	**100 %**

UNIVERSIDAD INTERNACIONAL DE LA RIOJA

unir

www.unir.net

www.unir.net