Conferencia: Salamandra: una nueva colección de modelos de lenguaje multilingües del BSC

Tipo de evento
Congreso, jornada o conferencia
Temática
Científico/Tecnológico
Fecha de inicio
21/11/2024 - 09:30

En esta charla, se presenta Salamandra, una familia de modelos altamente multilingües que han sido preentrenados desde cero, y que ofrecen tres tamaños diferentes: 2 mil millones, 7 mil millones y 40 mil millones de parámetros, cada uno con versiones base e instruidas. Los modelos Salamandra se han desarrollado utilizando un amplio corpus de preentrenamiento de 7,8 billones de tokens, recopilados de datos altamente curados que abarcan textos en 35 idiomas europeos, así como lenguajes de programación. Todo el entrenamiento se realizó en el MareNostrum 5, un supercomputador pre-exaescala, alojado y operado por el Barcelona Supercomputing Center, utilizando el marco NeMo de NVIDIA. Este enfoque innovador en el entrenamiento de modelos posiciona a Salamandra como una herramienta robusta para la investigación, aplicaciones y productos multilingües, ya que los modelos están disponibles bajo la licencia Apache 2.0.

Aitor González-Agirre es el líder del equipo de Modelos de Lenguaje en la Unidad de Tecnologías del Lenguaje del Barcelona Supercomputing Center, responsable del desarrollo de Modelos de Lenguaje.
Recibió sus títulos de Máster y su Doctorado en Procesamiento del Lenguaje Natural (PLN) en la Universidad del País Vasco (UPV/EHU). Obtuvo el Premio a la Mejor Tesis en la Conferencia SEPLN de 2018. También cuenta con experiencia previa en diversas tareas clave de PLN, tales como Similitud Semántica Textual (STS), BARR2, MEDDOCAN, PharmaCoNER, MESINESP y CodiEsp, entre otras.
Su trabajo anterior incluye el desarrollo de modelos de lenguaje para español, catalán y dominios especializados como el biomédico y el legal. En 2022, ganó el Premio Archiletras a la Innovación por su trabajo en "MarIA: Spanish Language Models". Más recientemente, ha estado involucrado en el entrenamiento de la familia de modelos de lenguaje Salamandra, desarrollados desde cero utilizando 7.8 billones de tokens, con tamaños de modelos de 2 mil millones, 7 mil millones y 40 mil millones.

La charla en formato de vídeo puede ser visualizada desde aquí o desde el enlace adjunto a este evento:

Lugar

Salón de Grados A3

Locate

37.787688127577, -3.777852218935

Enlaces relacionados

- Enlace al vídeo