
TEXT
SUMMARIZATION
NLP

EL RESUMEN DE TEXTO CONDENSA UNO O MÁS TEXTOS EN RESÚMENES MÁS CORTOS PARA MEJORAR LA EXTRACCIÓN DE INFORMACIÓN.

RESUMEN DE TEXTO

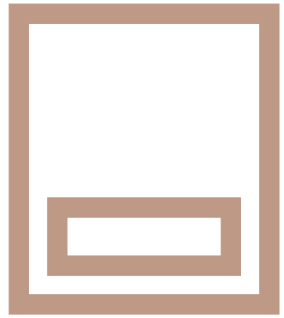
El resumen automático de texto (o resumen de documentos) es un método de procesamiento del lenguaje natural (NLP) que condensa la información de uno o más documentos de texto de entrada en un texto de salida original. Se debate la cantidad de texto de entrada que aparece en la salida: algunas definiciones indican solo el 10 %, otras el 50 %. Los algoritmos de resumen de texto a menudo utilizan arquitecturas de aprendizaje profundo, específicamente, *transformers*, para analizar documentos y generar resúmenes de texto.

TIPOS DE RESUMEN AUTOMÁTICO DE TEXTO

Hay dos tipos principales de resumen: extractivo y abstractivo.

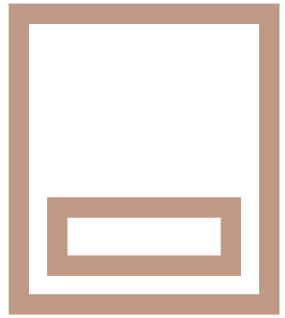
El resumen extractivo extrae oraciones no modificadas de los documentos de texto originales. Una diferencia clave entre los algoritmos extractivos es cómo puntúan la importancia de las oraciones y reducen la redundancia tópica. Las diferencias en la puntuación de las oraciones determinan qué oraciones extraer y cuáles conservar.

El resumen abstractivo genera resúmenes originales utilizando oraciones que no se encuentran en los documentos de texto originales. Dicha generación requiere redes neuronales y grandes modelos de lenguaje (LLM) para producir secuencias de texto semánticamente significativas.

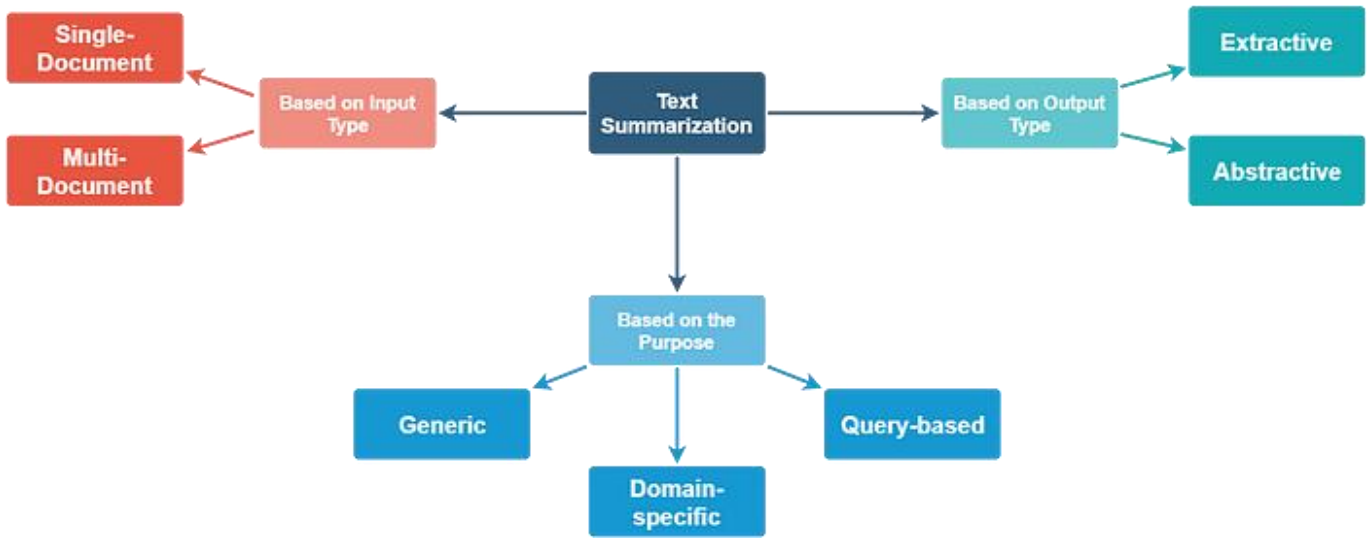


TIPOS DE RESUMEN AUTOMÁTICO DE TEXTO

Como se puede suponer, el resumen abstractivo de textos es más costoso computacionalmente que el extractivo, y requiere una comprensión más especializada de la inteligencia artificial y los sistemas generativos. Por supuesto, el resumen de texto extractivo también puede utilizar de redes neuronales *transformers*, como GPT, BERT y BART, para crear resúmenes. Sin embargo, los enfoques extractivos no requieren redes neuronales.



TIPOS DE RESUMEN AUTOMÁTICO DE TEXTO

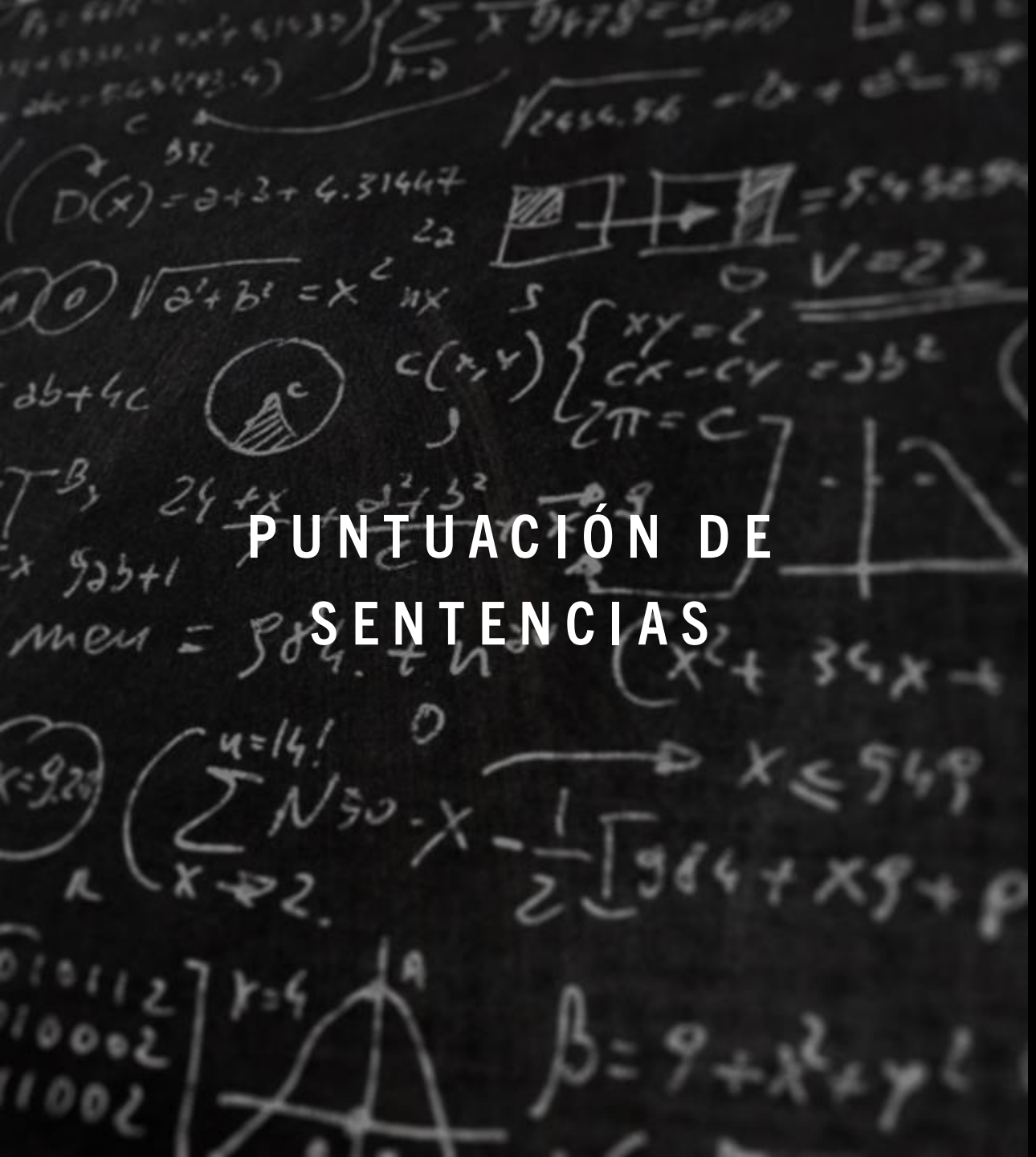


CÓMO FUNCIONA EL RESUMEN DE TEXTO EXTRACTIVO

Al igual que con otras tareas de NLP, el resumen de texto requiere que los datos de texto se sometan primero a un preprocesamiento. Esto incluye la tokenización, la eliminación de palabras vacías, stemming o lematización para que el conjunto de datos sea legible por un modelo de aprendizaje automático. Después del preprocesamiento, todos los métodos de resumen de textos extractivos siguen tres pasos generales e independientes: representación, puntuación de oraciones y selección de oraciones.

REPRESENTACIÓN

En la etapa de representación, un algoritmo segmenta y representa datos de texto preprocesados para su comparación. Muchas de estas representaciones se construyen a partir de modelos de *bolsa de palabras*, que representan segmentos de texto, como palabras u oraciones, como puntos de datos en un espacio vectorial. Los grandes conjuntos de datos de varios documentos pueden utilizar TF-IDF, una variante de bolsa de palabras que pondera cada término para reflejar su importancia dentro de un conjunto de texto.



PUNTUACIÓN DE SENTENCIAS

La puntuación de oraciones, como su nombre lo indica, puntúa cada oración de un texto de acuerdo con su importancia para ese texto. Diferentes representaciones implementan diferentes métodos de puntuación. Enfoques basados en grafos, calculan la centralidad de las oraciones. Estos algoritmos determinan la centralidad utilizando TF-IDF para calcular qué tan lejos puede estar un nodo de oración con respecto al centroide de un documento en el espacio vectorial.

SELECCIÓN DE ORACIONES

El último paso general en los algoritmos extractivos es la selección de oraciones. Al haber ponderado las oraciones por importancia, los algoritmos seleccionan las *n* oraciones más importantes para un documento o colección de este. Estas oraciones comprenden el resumen generado.

FUENTE

- IBM