N-Grams: Bigramas y Trigramas

Comprender el poder de los N-Gramas

Table of Contents

1

Introducción a los N-Gramas

Descripción general

- Los N-gramas son secuencias de 'n' palabras adyacentes en un texto, utilizadas en el procesamiento del lenguaje natural. Los bigramas tienen 2 palabras; Los trigramas tienen 3.
- Los N-gramas ayudan a predecir la siguiente palabra de una oración y son cruciales en las tareas de modelado del lenguaje y generación de texto.
- Los bigramas capturan relaciones cercanas entre palabras, mientras que los trigramas brindan más contexto y capturan relaciones entre tripletes de palabras.
- Comprender las diferencias entre bigramas y trigramas es esencial para aprovechar los n-gramas de manera efectiva para el análisis y la predicción de texto.

2

Bigramas en profundidad

Análisis

- Los bigramas analizan la coexistencia de dos palabras y ayudan a estudiar la proximidad y las asociaciones de palabras dentro de un corpus de texto.
- Las aplicaciones comunes de los bigramas incluyen predicción de texto, revisión ortográfica y etiquetado de partes del discurso en procesos de procesamiento del lenguaje natural.
- Los bigramas son más simples que los trigramas, pero son eficaces para capturar el contexto inmediato y las dependencias que a menudo ocurren en el uso del lenguaje.
- Los bigramas se utilizan ampliamente en diversas tareas de PLN debido a su simplicidad y eficiencia a la hora de capturar relaciones entre palabras dentro de datos de texto.

3

Trigramas revelados

Exploración

- Los trigramas amplían aún más el concepto al analizar la coexistencia de tres palabras, proporcionando un contexto más profundo y capturando dependencias de mayor alcance.
- Los trigramas son beneficiosos en tareas que requieren un contexto más extenso, como la generación de lenguaje, el análisis de sentimientos y el reconocimiento de entidades nombradas.
- Si bien los trigramas ofrecen un contexto más rico, pueden ser computacionalmente más costosos que los bigramas debido a la mayor combinación de palabras que capturan.
- En tareas complejas de modelado del lenguaje, los trigramas desempeñan un papel vital a la hora de mejorar la precisión predictiva y capturar patrones lingüísticos matizados.

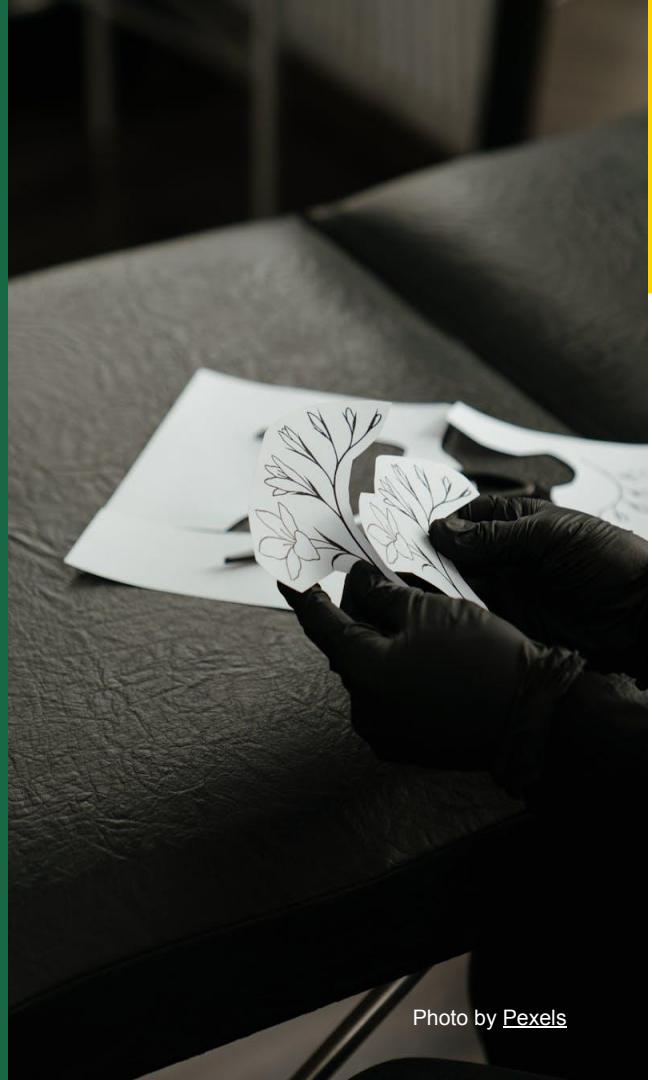
Photo by Pexels

4

Escenarios de uso

- Los bigramas son adecuados para tareas en las que las relaciones inmediatas entre palabras son importantes, como la predicción básica de texto o la identificación de frases comunes en un corpus de texto.
- Los trigramas brillan cuando se requiere un contexto lingüístico más profundo, como en el análisis de sentimientos, la traducción automática o la generación de secuencias de texto coherentes.
- La elección entre bigramas y trigramas depende de los requisitos específicos de la tarea, el conjunto de datos disponible y el nivel de complejidad lingüística deseado.
- Comprender los escenarios de aplicación ayuda a seleccionar el modelo de n-gramas apropiado para obtener resultados óptimos en diversas tareas de procesamiento del lenguaje natural.

5

Desafíos y consideraciones

- Elegir la 'n' correcta en n-gramas es crucial ya que los valores de 'n' más grandes dan como resultado datos más dispersos y aumentan la sobrecarga computacional, lo que afecta el rendimiento.
- La elección entre bigramas y trigramas también debe considerar el equilibrio entre la complejidad del modelo, los recursos computacionales y el nivel de matices lingüísticos requeridos.
- El preprocesamiento de datos, incluida la tokenización y la limpieza, es esencial para garantizar que los modelos de n-gramas capturen asociaciones y patrones de palabras significativos de manera efectiva.
- Equilibrar la granularidad del análisis con bigramas y la profundidad contextual de los trigramas es una consideración clave para optimizar el uso de n-gramas para aplicaciones específicas de PNL.

Beneficios y limitaciones

- Los bigramas ofrecen simplicidad, eficiencia y cálculo rápido, lo que los hace ideales para obtener información rápida a partir de datos de texto, pero pueden carecer de captura de matices contextuales extensos.
- Los trigramas destacan por capturar contextos más ricos y patrones lingüísticos matizados, pero pueden requerir más recursos computacionales y un cuidadoso ajuste de parámetros.
- La combinación de bigramas y trigramas en modelos híbridos puede aprovechar las fortalezas de ambos enfoques, equilibrando la eficiencia con la profundidad del contexto en las tareas de PLN.
- En la práctica, comprender los beneficios y las limitaciones de los bigramas y trigramas permite a los profesionales tomar decisiones informadas al diseñar modelos de lenguaje efectivos.

7

Introducción a los N-Gramas

Computando *P(w|h)*

Un modelo de N-grama bien elaborado puede predecir eficazmente la siguiente palabra de una frase, lo que consiste esencialmente en determinar el valor de p(w | h), donde h es la historia o el contexto y w es la palabra a predecir.

Veamos cómo predecir la siguiente palabra de una frase. Tenemos que calcular P(w|h), donde w es la palabra candidata a ser la siguiente. Consideremos la frase «Este artículo es sobre...». Si queremos calcular la probabilidad de que la siguiente palabra sea «PLN», la probabilidad puede expresarse como:

P(«PLN»|«Este», «artículo», «es», «sobre»)

7

Introducción a los N-Gramas

Computando *P(w|h)*

Para generalizar, la probabilidad condicional de la quinta palabra dadas las cuatro primeras puede escribirse como:

$p(w_5 \mid w_1, w_2, w_3, w_4)$ ó $p(w_n \mid w_1, w_2, \ldots, w_{n-1})$ $P(A|B) = \frac{P(A \cap B)}{P(B)}$

Cualquier probabilidad de distribución conjunta puede ser descompuesta en distribuciones de probabilidad sobre una variable. Esto se llama “regla de la cadena” o regla del producto, de la probabilidad.

7

Introducción a los N-Gramas

Computando *P(w|h)*

Para la probabilidad conjunta de que cada palabra de una secuencia tenga un valor determinado $P(X_1 = w_1, X_2 = w_2, X_3 = w_3, \ldots, X_n = w_n)$ utilizaremos $P(w_1, w_2, \ldots, w_n)$.
Ahora, ¿cómo podemos calcular probabilidades de secuencias enteras como $P(w_1, w_2, \ldots, w_n)$?
Una cosa que podemos hacer es descomponer esta probabilidad utilizando la regla de la cadena de probabilidad:

$$\begin{aligned} P(X_1 \ldots X_n) &= P(X_1)P(X_2|X_1)P(X_3|X_{1:2}) \ldots P(X_n|X_{1:n-1}) \\ &= \prod_{k=1}^{n} P(X_k|X_{1:k-1}) \end{aligned}$$

7

Introducción a los N-Gramas

Computando *P(w|h)*

La regla de la cadena muestra el vínculo entre el cálculo de la probabilidad conjunta de una secuencia y el cálculo de la probabilidad condicional de una palabra dadas las palabras anteriores. La ecuación de la lámina anterior sugiere que podemos calcular la probabilidad conjunta de toda una secuencia de palabras multiplicando varias probabilidades condicionales.

La intuición del modelo de n-gramas es que en lugar de calcular la probabilidad de una palabra dada toda su historia, podemos **aproximar** la historia sólo con las últimas palabras

7

Introducción a los N-Gramas

Computando *P(w|h)*

El modelo de bigramas, por ejemplo, aproxima la probabilidad de una palabra dadas todas las palabras anteriores $P(w_n|w_{1:n-1})$ utilizando sólo la probabilidad condicional de la palabra precedente $P(w_n|w_{n-1})$. En otras palabras, en lugar de calcular la probabilidad.

P(«PLN»|«Este», «artículo», «es», «sobre»)

lo aproximamos con la probabilidad

P(«PLN»|«sobre»)

7

Introducción a los N-Gramas

Computando *P(w|h)*

La suposición de que la probabilidad de una palabra depende únicamente de la palabra anterior se denomina suposición de Markov. Los modelos de Markov son la clase de modelos probabilísticos que suponen que podemos predecir la probabilidad de alguna unidad futura sin mirar demasiado al pasado. Podemos generalizar el bigrama (que mira una palabra en el pasado) al trigrama (que mira dos palabras en el pasado) y así al n-grama (que mira n-1 palabras en el pasado).