UNIVERSIDAD
TORCUATO DI TELLA

Escuela de Negocios
Licenciatura en Tecnología Digital

Trabajo práctico 3
Lenguaje

Autores: Noguera Azul, Rocio Gonzalez Cingolani, Pancaldi Valentina y Jordan Paula.

Buenos Aires, 29 de Noviembre 2024

Este proyecto tiene como objetivo desarrollar y evaluar un modelo generador de texto basado en n-gramas, aplicando técnicas de procesamiento de lenguaje natural (NLP) para crear texto coherente y fluido a partir de un corpus de entrenamiento en castellano.

Para lograr este objetivo, el primer paso fue la construcción de un corpus de texto que refleje de manera auténtica el uso del español rioplatense y específicamente jerga argentina. Para ello, se realizaron varias etapas de recopilación y generación de datos:

1. Recopilación de textos reales mediante web scraping:
    - Se desarrolló un script de web scraping para extraer artículos de diarios argentinos populares, como Clarín y La Nación. Esto permitió capturar un lenguaje más contemporáneo y alineado con el contexto cultural y social actual del país.
    - Se aplicaron técnicas de limpieza y filtrado del texto extraído para eliminar elementos no deseados, como publicidad, encabezados repetitivos y contenido irrelevante, dejando solo el contenido relevante para el corpus.
2. Generación de datos sintéticos:
    - Dado que los datos obtenidos del scraping no eran suficientes para capturar toda la riqueza del lunfardo y las expresiones coloquiales argentinas, se complementó el corpus con texto generado artificialmente.
    - Este texto sintético fue cuidadosamente diseñado para incluir frases típicas, jergas y modismos que reflejan el habla cotidiana de los argentinos en contextos como la vida diaria, el fútbol, el asado y la familia.
3. Complementación con el Proyecto Gutenberg:
    - Se incorporaron textos literarios del Proyecto Gutenberg que contienen obras clásicas en español. Si bien estos textos no siempre reflejan el lunfardo argentino, fueron útiles para proporcionar diversidad lingüística y riqueza gramatical al corpus.
4. Integración y limpieza del corpus:
    - Se unificaron todas las fuentes de datos en un único archivo de texto estructurado. Se aplicaron técnicas de normalización, como la eliminación de caracteres no deseados y la conversión a minúsculas, para garantizar la uniformidad.
    - Además, se aseguraron de que los textos no tuvieran información redundante ni contenido que pudiera sesgar negativamente el modelo.
5. Construcción del modelo basado en n-gramas:
    - Una vez construido el corpus, se implementaron modelos generadores de texto utilizando bigramas, trigramas, y hasta modelos basados en 5-gramas. Estos modelos fueron entrenados para predecir la próxima palabra en una secuencia dada, generando texto que sigue patrones similares a los del corpus.
    - Se diseñaron funciones específicas para evaluar y ajustar la coherencia del texto generado, asegurando que mantuviera un estilo consistente y un lenguaje natural.
6. Evaluación y ajuste del modelo:
    - Los resultados del modelo fueron evaluados en función de la coherencia lingüística, la naturalidad y la capacidad de reproducir patrones característicos del español argentino.

- Se realizaron iteraciones para ajustar los parámetros del modelo y optimizar la generación de texto.

Ejercicio 3

¿Cómo es la calidad de los textos generados, a medida que aumentan n y/o la cantidad de datos de entrenamiento? ¿Qué tipos de errores se producen?

Calidad de los textos generados a medida que aumenta n:

A medida que aumentamos el valor de **n** , el modelo logra capturar un contexto lingüístico más amplio, mejorando la coherencia y fluidez de los textos generados. Esto se debe a que cada palabra o frase está condicionada por una secuencia más extensa de palabras anteriores, lo que permite reflejar con mayor precisión el estilo del corpus, que incluye jerga y expresiones propias del español argentino. Pero también incrementa la necesidad de un corpus más extenso y diverso para evitar problemas de repetición o falta de generalización.

Bigramas:

Con n=2 , el modelo genera texto basado únicamente en la palabra anterior. Esto resulta en frases que pueden sonar fragmentadas y carecer de una estructura gramatical adecuada. Además, tienden a repetirse patrones comunes del corpus, y las oraciones suelen estar desconectadas entre sí.

Ejemplo de texto generado: *"La república argentina y ánimo de valdivia donde quiera reemplazarlo se sonaba en efecto dos años más vibrante y gobernados por amor…"*

Aunque algunas combinaciones son válidas, las frases no presentan continuidad narrativa ni fluidez.

Trigramas:

Al pasar a n=3 , el modelo tiene acceso a un contexto mayor, lo que permite construir frases más coherentes y estructuradas. Esto mejora significativamente la gramática y da lugar a textos más cercanos al lenguaje natural, aunque aún pueden presentar incoherencias narrativas entre oraciones consecutivas.

Ejemplo de texto generado: *"La selección está jugando cada vez más añadió no sin echar una mirada de cariño duerme niño todavía no han pensado por eso los estafadores dignos de fama malogran un esfuerzo que le cuadre como lo dice muy contento."*

Aquí, las frases son más completas y gramaticalmente correctas, aunque el texto generado puede carecer de cohesión temática o narrativa global.

4-gramas y superiores:

Con n=4, o valores más altos, el modelo captura patrones más complejos y reproduce estructuras gramaticales completas, logrando textos mucho más naturales y contextualmente relevantes. Sin embargo, si el corpus no es lo suficientemente grande, el modelo puede memorizar frases completas, resultando en texto repetitivo o poco creativo.

Ejemplo de texto generado (n=5): *"En argentina llevamos el fútbol en la sangre y cada partido es como una nueva final del mundo"*

Este texto muestra una mayor coherencia contextual y narrativa, aunque puede volverse predecible si se limita a patrones frecuentes del corpus.

Cantidad de datos de entrenamiento:
Al aumentar la cantidad de datos, el modelo mejora en la predicción de secuencias de palabras, lo que ayuda a reducir errores gramaticales y repeticiones. Con un conjunto de datos más extenso, el modelo tiene más ejemplos de combinaciones de palabras, lo que ayuda a capturar patrones de lenguaje más variados. Sin embargo, el modelo aún puede producir frases incoherentes o inconsistencias en el flujo de ideas al no tener una comprensión semántica real.

Tipos de errores comunes:

- Repeticiones: En bigramas, se puede observar que el modelo genera secuencias repetitivas, como *"es muy, es muy, es muy…".*
- Incongruencias gramaticales: El modelo generalmente falla en construir una estructura gramatical sólida en frases largas.
- Frases sin sentido: El modelo, en muchas ocasiones, produce secuencias de palabras que, aunque gramaticalmente correctas, carecen de sentido lógico.

¿Cuánto se parecen los textos generados a los textos originales, a medida que aumentan n y/o la cantidad de datos de entrenamiento?

Similitud con el texto original:
A medida que aumentamos n, el texto generado se vuelve más similar al original en términos de estructura y gramática, ya que el modelo recuerda secuencias de palabras más largas. En el caso de trigramas o valores de n mayores, es posible que el modelo reproduzca frases completas del texto original, especialmente si el corpus de entrenamiento es pequeño.

Diversidad y variación:
Con un valor de n bajo, el texto generado suele tener una diversidad de palabras alta pero una coherencia baja. Con un n mayor y una cantidad de datos extensa, el modelo equilibra mejor la diversidad con la coherencia, acercándose a la estructura del texto original. Sin embargo, es importante notar que con valores muy altos de n, el modelo puede caer en la repetición exacta de fragmentos del texto de entrenamiento, especialmente si el corpus es pequeño, lo que limita la creatividad del modelo.

¿Qué grado de creatividad ven en estos modelos? ¿Y de inteligencia?
Creatividad:
Los modelos de n-gramas tienen un grado de creatividad limitada, ya que solo pueden generar combinaciones de palabras basadas en el texto de entrenamiento y en la probabilidad de aparición de cada secuencia de palabras. La creatividad del modelo depende en gran medida de la diversidad del corpus de entrenamiento; sin embargo, su creatividad es limitada, ya que no "comprende" el contexto ni el significado de las palabras. Los modelos de n-gramas pueden producir combinaciones novedosas, pero a menudo carecen de coherencia global, lo cual disminuye la percepción de creatividad.

Inteligencia:
Este modelo no posee inteligencia real, ya que simplemente cuenta y combina palabras en función de patrones de frecuencia en el corpus. No tiene una comprensión del significado, ni es capaz de razonar o mantener una narrativa coherente. Su “inteligencia” es puramente estadística y sintáctica, y en general, presenta un bajo grado de comprensión contextual. A medida que aumenta n, el modelo parece más “inteligente” debido a la mejora en la gramática y la fluidez del texto, pero este efecto es el resultado de patrones de probabilidad, no de comprensión real.

Fine-tuning GPT-2

Se decidió realizar una implementación adicional para comparar la performance de los n-gramas con un modelo más avanzado. Para ello, se optó por utilizar **GPT-2**, preentrenado en español, ajustándolo mediante fine-tuning para que aprenda a producir texto coherente y con jerga argentina.

El ajuste fino se realizó utilizando el mismo corpus que se utilizó para los N-gramas. Este enfoque permitió que el modelo adaptara su conocimiento general del español a un ámbito más específico y local.

Se notó una mejora significativa en la fluidez y coherencia contextual del texto generado en comparación con los modelos basados en n-gramas. Mientras que los n-gramas tienden a limitarse a patrones locales, GPT-2, gracias a su capacidad de capturar dependencias más largas, logra construir narrativas más completas y con mayor sentido. Además, la jerga y expresiones típicas argentinas aparecieron de manera más natural, reflejando mejor el corpus de entrenamiento.

Ejemplo de texto generado por GPT-2 Fine-tuned:
“La selección argentina ha sido una de las mejores del torneo. El partido fue muy bueno y el equipo ganó el partido. Pero el último partido es importante, ya que es el momento más importante para la selección, porque es la última oportunidad de hacer que el mundo salga de la pobreza. Y el final.”