Dennys Emmanuel Tezén Guerra comenzó la sesión dando la bienvenida y explicando que hoy se hablará sobre los conceptos de machine learning, incluyendo qué es una red neuronal, cómo se dividen, la matemática detrás de ellas y los tipos de datos, ya que esto será fundamental para las futuras clases más complejas. Asimismo, Dennys Emmanuel Tezén Guerra detalló que se abordarán los datos de entrenamiento, los algoritmos de machine learning y cómo estos aprenden, ya sea a través de inyección de información, corrección o sistemas de recompensas, además de la importancia de ajustar los modelos, un aspecto relevante en el desarrollo de aplicaciones con inteligencia artificial y una pregunta frecuente en exámenes.

Dennys Emmanuel Tezén Guerra explicó que en la clase se abordará el tema de la ingeniería de *prompts*, destacando su relevancia para el examen al enfocarse en los tipos de datos necesarios para que un *prompt* sea efectivo. Además, se discutirán los conceptos de *machine learning*, haciendo hincapié en el algoritmo GPT (*Generative Pre-training Transformer*), famoso por su capacidad de generar texto similar al humano, código, imágenes, audio y video. Finalmente, Dennys Emmanuel Tezén Guerra detalló el modelo BERT (*Bidirectional Encoder Representations from Transformers*), que se diferencia de GPT por su lectura bidireccional, lo que le permite predecir palabras considerando tanto el texto anterior como el posterior.

Dennys Emmanuel Tezén Guerra explicó que las redes neuronales recurrentes (RNN) están diseñadas para datos secuenciales y las redes residuales para reconocimiento de imágenes, siendo estas últimas más complejas. José Castro mencionó que ha estado trabajando en un modelo de visión por computadora llamado Yolo para un sistema de tráfico inteligente en Panajachel, el cual busca leer métricas de tráfico y próximamente incluirá reconocimiento de matrículas y un chatbot. José Castro también aclaró que la inteligencia artificial no es solo el modelo, sino la integración de diversas herramientas para crear una solución completa, destacando que los modelos de inteligencia artificial hacen lo que pueden con los datos que tienen.

Jose Castro explica que para una mayor precisión en el reconocimiento por visión por computadora, es crucial contar con buen hardware, como cámaras de calidad y un sistema eficiente para enviar el video. Dennys Emmanuel Tezén Guerra menciona que esto requeriría un equipo muy pesado y un alto consumo. Jose Castro añade que, aunque la tecnología está disponible, las principales limitaciones en Guatemala son la infraestructura y los altos costos de los recursos, ya que solo el sistema en la nube podría costar entre 20 y 25 mil dólares, sin incluir otros gastos. Dennys Emmanuel Tezén Guerra comenta que este proyecto es interesante y que ya habían investigado qué servicios de AWS podrían ejecutarlo, como EKS o Amazon S2, pero se encontraron con limitaciones con Lambda.

Dennys Emmanuel Tezén Guerra explicó el funcionamiento de los sistemas de audio basados en frecuencias, como Amazon Polly, que convierten texto a voz sin grabaciones humanas completas, utilizando sonidos pregrabados como base para generar nuevas ondas. Dennys Emmanuel Tezén Guerra también introdujo el concepto de red GAN, una red generativa antagónica utilizada para datos de contenido multimedia, y el GBOST, una implementación similar para modelos a gran escala. Finalmente, Dennys Emmanuel Tezén Guerra enfatizó la importancia de la calidad de los datos de entrenamiento en el aprendizaje automático, destacando que los resultados de un modelo dependen directamente de la precisión de los datos de entrada, utilizando ejemplos como

la proyección financiera y el etiquetado de imágenes para ilustrar cómo los datos erróneos pueden llevar a conclusiones incorrectas.

Dennys Emmanuel Tezén Guerra comenzó explicando los datos sin etiquetar, los cuales solo incluyen características de entrada sin las de salida, como colecciones de imágenes sin etiquetas asociadas. Luego, describió cómo un sistema identifica patrones en los datos de entrada para producir una salida, utilizando ejemplos como la identificación de manzanas verdes a través de atributos como el color y la forma. Finalmente, Dennys Emmanuel Tezén Guerra abordó los datos estructurados, como los tabulares organizados en filas y columnas, y los datos no estructurados, como el texto y el contenido multimedia, aclarando que, aunque un artículo pueda tener una estructura interna, no sigue el formato de filas y columnas de los datos estructurados.

Dennys Emmanuel Tezén Guerra definió los datos no estructurados como textos grandes que no siguen una estructura de filas y columnas, a diferencia de los datos estructurados donde la información está organizada en celdas. También explicó que los datos de imagen, que varían en formato y contenido, son otro tipo de datos no estructurados y se pueden analizar para diversas aplicaciones, como la gestión de contenido multimedia o la identificación de objetos. Luego introdujo el aprendizaje supervisado, que utiliza datos etiquetados para predecir resultados, y dio el ejemplo de la regresión, utilizada para predecir valores numéricos continuos.

Dennys Emmanuel Tezén Guerra explicó que en el aprendizaje automático, los datos deben estar etiquetados, como los ingresos de una empresa para predecir el próximo semestre o la clasificación de imágenes de gatos. También indicó que el proceso de aprendizaje automático comienza con la definición de un problema, seguido por el diseño de un algoritmo y la recopilación de datos, donde el conjunto de entrenamiento representa entre el 60% y el 80% del total de datos. Finalmente, subrayó la importancia de tener una gran cantidad de datos para entrenar el modelo, comparándolo con la necesidad de estudiar a fondo un tema antes de dar una charla.

Dennys Emmanuel Tezén Guerra explicó que el conjunto de datos de validación se utiliza para ajustar los parámetros y evaluar el rendimiento de un modelo, generalmente entre el 10% y el 20% del total de datos, buscando respuestas coherentes y rápidas. También mencionó el concepto de "feature engineering", un proceso que implica seleccionar y transformar datos sin procesar en características significativas para mejorar el rendimiento del modelo, como se ve en cómo Chat GPT mantiene información relevante sin cambios mientras varía otros detalles. Finalmente, describió el aprendizaje no supervisado, que busca descubrir patrones y relaciones en los datos sin etiquetas predefinidas, permitiendo que la máquina agrupe y detecte anomalías por sí misma, adaptándose a diferentes contextos de negocio.

Dennys Emmanuel Tezén Guerra explicó el agrupamiento de datos, el cual clasifica datos similares en grupos según sus características, como en los firewalls que agrupan direcciones IP con mala reputación. Dennys Emmanuel Tezén Guerra también describió la segmentación de clientes, dando el ejemplo de segmentar clientes de supermercado según sus hábitos de gasto, y cómo esto se aplica en servicios como AWS Personalize para ofrecer recomendaciones basadas en compras anteriores. Finalmente, Dennys Emmanuel Tezén Guerra habló sobre la detección de anomalías, ilustrando con la detección de fraudes bancarios, donde el sistema marca transacciones inusuales que podrían ser fraudulentas.

Dennys Emmanuel Tezén Guerra explicó que la anomalía de una compra depende del contexto individual, señalando que una compra de 3.000 $ en Europa, aunque inusual para algunos, podría ser normal para un ejecutivo. Luego, detalló los tipos de aprendizaje en la inteligencia artificial, comenzando con el aprendizaje semisupervisado, donde se utiliza una pequeña cantidad de datos etiquetados y una gran cantidad de datos no etiquetados para entrenar un sistema, y posteriormente describió el aprendizaje autosupervisado, en el cual el modelo genera sus propias etiquetas y se aplica en modelos de IA generativa como Bert y GPT. Finalmente, Dennys Emmanuel Tezén Guerra abordó el aprendizaje por refuerzo, donde un agente aprende a tomar decisiones para maximizar recompensas en un entorno controlado, utilizando como ejemplo los robots de ajedrez que buscan ganar el torneo.

Dennys Emmanuel Tezén Guerra explicó que el aprendizaje por refuerzo implica un agente, un entorno, una acción, una recompensa, un estado y una política, donde el agente observa el entorno, selecciona una acción basada en su política, y la recompensa obtenida actualiza su política para futuras decisiones, de manera similar a cómo los humanos aprenden de sus experiencias. También abordó el aprendizaje por refuerzo a partir de la retroalimentación humana, donde el modelo de inteligencia artificial aprende de las correcciones hechas por un humano, incorporando una función de recompensa para alinear las respuestas a los objetivos deseados. Dennys Emmanuel Tezén Guerra ilustró esto con ejemplos de cómo un modelo aprende de las correcciones para evitar errores y ajustar sus respuestas a necesidades específicas, especialmente útil en modelos personalizados.

Dennys Emmanuel Tezén Guerra explicó que se puede actualizar el algoritmo de una organización para que brinde respuestas sobre la migración a Windows 11, considerando aplicaciones que requieren Windows 10 y ofreciendo soluciones. Julio Cesar Chali preguntó sobre la aplicación teórica o práctica de los modelos en un examen, a lo que Dennys Emmanuel Tezén Guerra respondió que son más teóricos pero útiles para evaluar y ajustar modelos como los de Amazon Bedrock en la práctica, lo que se profundizará en la clase ocho. Julio Cesar Chali también consultó sobre la integración de estos modelos con otras plataformas fuera de AWS, y Dennys Emmanuel Tezén Guerra afirmó que es posible a través de APIs, como con OpenAI, aunque esto generaría costos por transferencia de datos al no ser una red interna de AWS. Dennys Emmanuel Tezén Guerra introdujo el tema de la ingeniería de prompts, definiéndola como la técnica para dar indicaciones a un modelo de IA con el fin de optimizar esas instrucciones.

Dennys Emmanuel Tezén Guerra enfatizó la importancia de contextualizar las instrucciones al solicitar un ensayo sobre la independencia de Guatemala, indicando que el ensayo debía ser de tipo académico universitario y abordar temas como los próceres y el grito de independencia. También destacó la necesidad de especificar la estructura del ensayo (introducción, desarrollo y conclusión) para obtener un resultado más preciso. Finalmente, Dennys Emmanuel Tezén Guerra explicó el concepto de "negative prompting" o instrucciones negativas, que permite indicar a la inteligencia artificial qué información se debe omitir en la respuesta, como el período de la Inquisición en un ensayo sobre la historia de la Iglesia Católica.

Dennys Emmanuel Tezén Guerra explicó el uso de las instrucciones de exclusión para evitar contenido no deseado, mantener la concentración y mejorar la claridad en las respuestas de los modelos de IA, poniendo como ejemplo cómo pedir un ensayo sobre IA en AWS sin incluir

servicios de infraestructura o seguridad. Continuó mostrando cómo, al usar esta técnica, el modelo se enfoca en servicios de IA específicos como Amazon Recognition, Compent, Transcribe, Poly, Translate, Lex, Bedrock y SageMaker, y cómo se puede aplicar para especificar qué información incluir o no, como al solicitar un resumen de la independencia sin detallar su historia. Finalmente, Dennys Emmanuel Tezén Guerra mencionó la optimización del rendimiento como un aspecto más avanzado que no está disponible en la versión gratuita de ChatGPT.

Dennys Emmanuel Tezén Guerra comenzó explicando cómo encontrar y abrir el servicio de Amazon Bedrock, señalando que la función GPT era una adición reciente. También indicó que, aunque no todo el contenido del examen es predecible, los parámetros que está a punto de explicar sí serán evaluados. Finalmente, Dennys Emmanuel Tezén Guerra instruyó a la audiencia sobre cómo seleccionar un modelo específico, como Cloud 3.5 en Anthropic, para experimentar con diferentes configuraciones.

Dennys Emmanuel Tezén Guerra explicó que para obtener respuestas más creativas en un algoritmo, se debe modificar el valor de "temperature", aumentando este valor, aunque un riesgo es que la respuesta puede perder coherencia y exactitud. Además, mencionó otro valor llamado "top p", que cuando está bajo, el porcentaje de palabras más probables generará una respuesta coherente, pero cuando está alto, se obtendrán más palabras, resultando en un resultado más creativo y diverso. Finalmente, Dennys Emmanuel Tezén Guerra afirmó que el valor de "top p" se recomienda mantener alto para que el algoritmo tenga una mayor variedad de palabras a elegir, sin afectar el idioma.

Dennys Emmanuel Tezén Guerra comenzó explicando que el valor "top" debía ser de 30 y que al probarlo en una aplicación web, un error inesperado apareció. Dennys Emmanuel Tezén Guerra explicó que la cuota era un problema y que tendría que crear un ticket, y después de varios segundos, el sistema dio una respuesta sobre las maravillas de Guatemala. Dennys Emmanuel Tezén Guerra también explicó que en la clase número ocho, cuando vean Amazon Bedrock y creen una aplicación usando esta herramienta, no debería haber problemas, y añadió que otros dos parámetros que se pueden variar son la longitud máxima de la respuesta y el acceso premium para respuestas más largas.

Actualización más reciente

Dennys Emmanuel Tezén Guerra explicó el "stop sequence" en Amazon Bedrock, que son tokens que le indican al modelo cuándo dejar de generar resultados, y mencionó que este servicio centraliza varias IA, incluyendo las de Open AI. Julio Cesar Chali preguntó si la información para estos modelos proviene solo de internet o de alguna base de conocimientos, y si se pueden configurar para acceder a información privada de centros de investigación o universidades. Dennys Emmanuel Tezén Guerra aclaró que los modelos tienen una base de datos interna para el entrenamiento y acceso a internet para buscar información, y que sí se pueden conectar a fuentes de información privada, aunque no a través de un parámetro directo, sino mediante un servicio adicional que permite el acceso a documentos privados sin entrenar el modelo principal.