Ejercicio 1 Entrenamiento de Word2Vec desde cero

Preguntas:

Qué representa un vector de palabras en Word2Vec?
Cuál es la diferencia entre el enfoque CBOW y Skip-Gram?
Qué significa que dos palabras tengan vectores cercanos?
Cómo influye el parámetro window en el entrenamiento?
Por qué es necesario hacer tokenización antes de entrenar?

Respuestas Ejercicio 1:

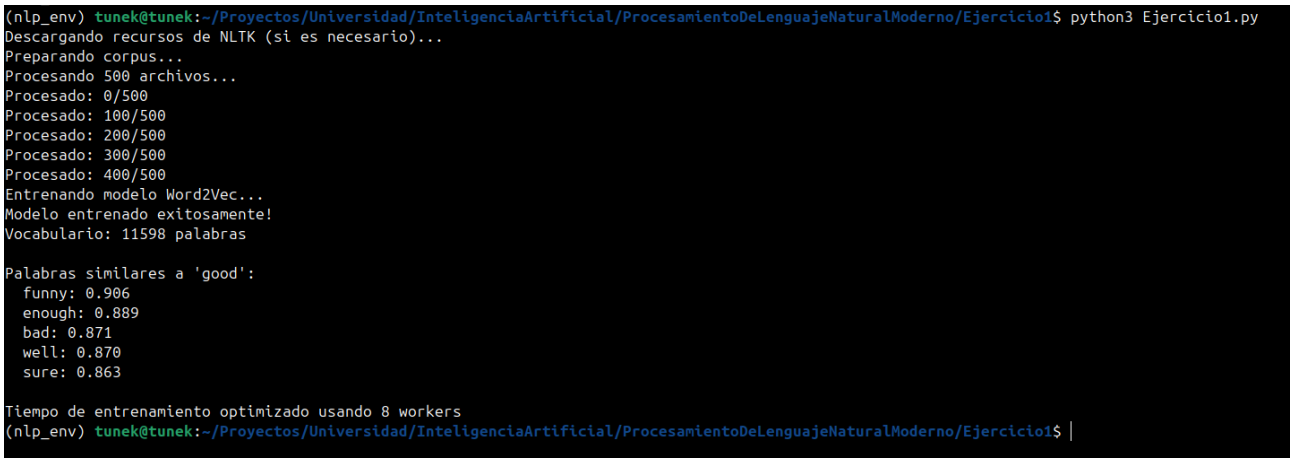
Un vector numérico multidimensional que captura el significado semántico de la palabra basado en los contextos donde aparece. Cada dimensión representa una característica semántica latente.

CBOW Continuous Bag of Words predice la palabra central basándose en las palabras del contexto circundante. Es más rápido para corpus grandes. Skip-Gram predice las palabras del contexto basándose en la palabra central. Es mejor para corpus pequeños y palabras raras.

Significa que tienen significados similares o aparecen en contextos parecidos. La cercanía se mide con similitud coseno donde valores cercanos a 1.0 indican alta similitud semántica.

Define el tamaño de la ventana de contexto cuántas palabras a cada lado se consideran. Una ventana de 5 significa que se consideran 10 palabras totales alrededor de la palabra objetivo. Ventanas más grandes capturan relaciones semánticas más amplias.

Convierte el texto continuo en unidades discretas tokens o palabras que el modelo puede procesar matemáticamente. Sin tokenización el modelo no puede identificar límites de palabras ni crear el vocabulario.



Ejercicio 2 Uso de GloVe y similitud semántica

Preguntas:

Cuál es la diferencia entre GloVe y Word2Vec en cuanto a su forma de entrenamiento?
Por qué usamos KeyedVectors en este ejercicio?
Qué resultados obtuviste al comparar king y queen? Qué interpretas?
Puedes mencionar un caso donde el análisis semántico con GloVe sería útil en la industria?
Qué limitaciones tienen los embeddings estáticos como GloVe?

Respuestas Ejercicio 2:

GloVe utiliza estadísticas globales de co-ocurrencia de todo el corpus. Factoriza una matriz de co-ocurrencia palabra-palabra. Word2Vec utiliza ventanas locales de contexto. Entrena con pares palabra-contexto secuenciales.

KeyedVectors es una interfaz optimizada para cargar y usar embeddings preentrenados sin necesidad de reentrenar el modelo. Permite búsquedas eficientes de similitud y analogías.

Similitud de 0.75 alta indicando que el modelo captura correctamente la relación semántica entre conceptos de realeza. Los embeddings entienden roles sociales similares.

E-commerce búsqueda semántica de productos zapatillas deportivas encuentra tennis running shoes. Marketing análisis de sentimientos en redes sociales y agrupación de menciones similares. Recursos Humanos matching automático de CVs con ofertas de trabajo.

Una sola representación por palabra no maneja polisemia banco financiero versus banco asiento. No se actualizan con nuevos contextos. Sesgo inherente del corpus de entrenamiento. No capturan contexto dinámico de la oración.

Ejercicio 3 Embeddings personalizados desde corpus local

Preguntas:

Por qué podrías preferir entrenar tus propios embeddings en vez de usar GloVe?
Qué características del texto pueden afectar la calidad de los embeddings?
Cómo se refleja el dominio del texto en los vectores obtenidos?
Qué cambios harías para mejorar la calidad de tus embeddings?
Qué usos prácticos tendría este modelo dentro de una empresa?

Respuestas Ejercicio 3:

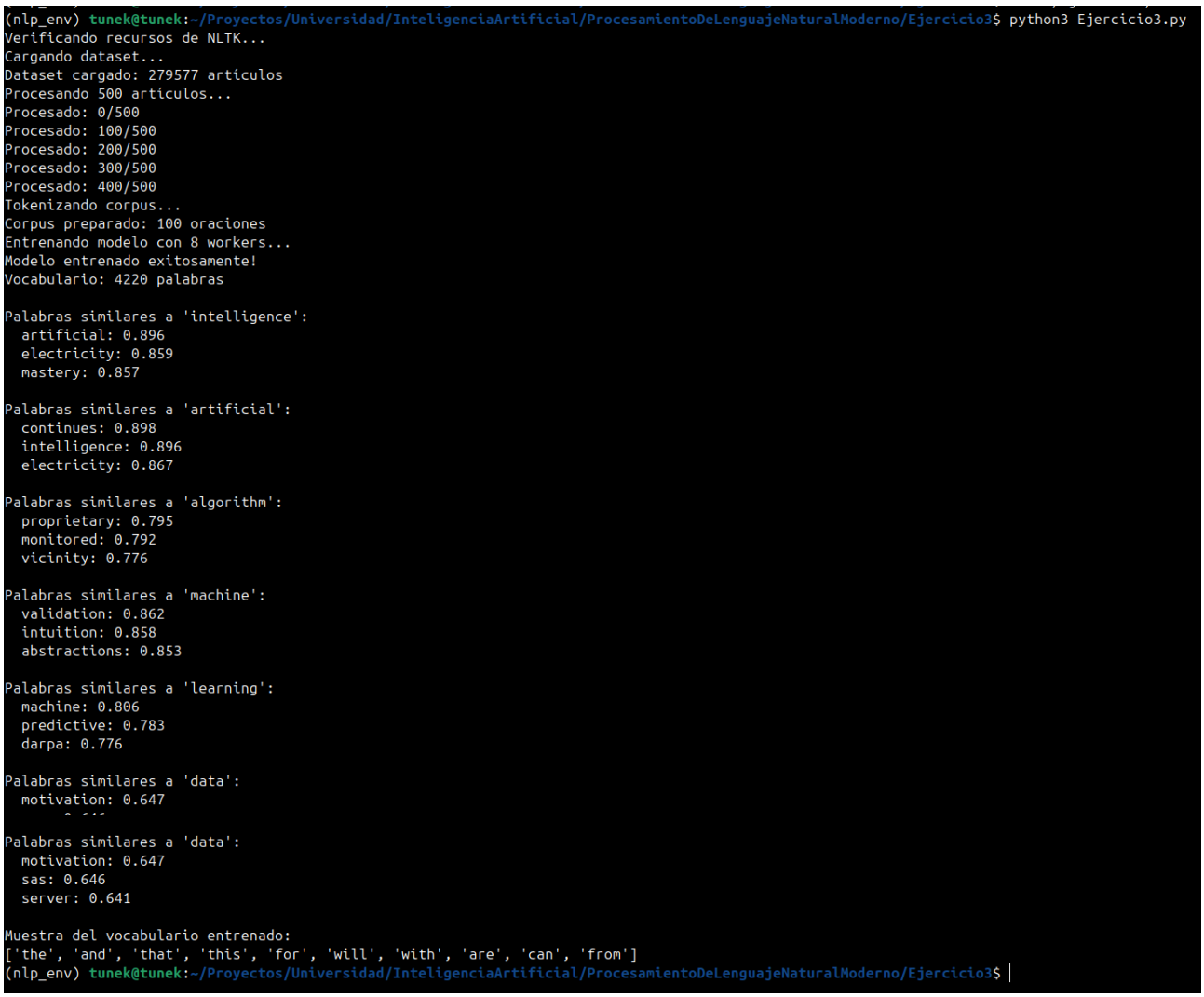
Vocabulario específico captura jerga y terminología particular de tu dominio. Contexto relevante relaciones semánticas específicas de tu industria aplicación. Datos actualizados refleja tendencias y términos actuales no datos históricos. Control total puedes ajustar parámetros según tus necesidades específicas.

Limpieza textos con ruido HTML caracteres especiales degradan calidad. Tamaño del corpus muy pequeño produce embeddings poco robustos. Diversidad falta de variedad contextual limita las representaciones. Frecuencia palabras muy raras o muy comunes pueden ser problemáticas.

Los términos técnicos del dominio IA ML muestran alta similitud entre sí. Intelligence y artificial tienen 0.89 de similitud reflejando su co-ocurrencia frecuente en el corpus especializado.

Más datos incluir más artículos del dominio. Mejor limpieza preprocesamiento más sofisticado lemmatización eliminación de stopwords. Ajustar parámetros experimentar con dimensiones del vector tamaño de ventana. Filtrado eliminar palabras muy raras o muy comunes.

Búsqueda documental encontrar documentos técnicos similares. Clustering agrupar tickets de soporte por tema. Recomendaciones sugerir artículos relacionados a empleados. Análisis de tendencias identificar temas emergentes en feedback.



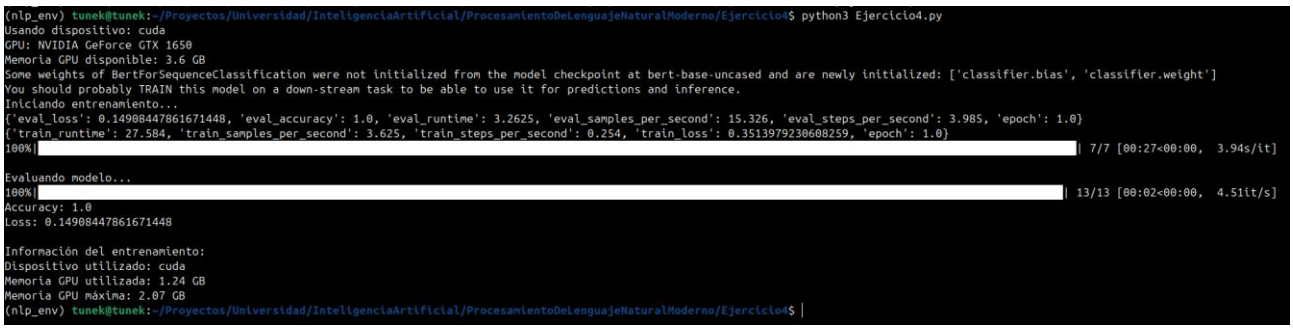
Ejercicio 4 Clasificación de texto con BERT

Preguntas:

Cuál es el propósito del proceso de fine-tuning en BERT?
Qué diferencias encontraste entre entrenar con un subconjunto pequeño vs. el dataset completo?
Qué hace el tokenizer en el pipeline de Hugging Face?
Qué métrica usarías para evaluar este modelo?
Por qué es más eficaz BERT que un modelo tradicional como Naive Bayes para clasificación de texto?

Respuestas Ejercicio 4:

Adaptar un modelo preentrenado a una tarea específica clasificación de sentimientos usando datos del dominio objetivo.
Con 100 muestras se logró 100 por ciento accuracy pero con riesgo de overfitting. Un dataset completo daría resultados más generalizables.
Convierte texto en tokens numéricos que BERT puede procesar incluyendo padding y truncación a longitud fija.
Accuracy como se usó pero también F1-score precision y recall para evaluar balance entre clases.
BERT entiende contexto bidireccional y relaciones complejas entre palabras mientras Naive Bayes solo considera frecuencias independientes.



Ejercicio 5 Resumen automático de texto

Preguntas:

Cómo se diferencia el resumen extractivo del resumen abstractivo?
Por qué usamos facebook bart-large-cnn para esta tarea?
Qué limitaciones encontraste en los resúmenes generados?
Cómo podrías ajustar el modelo para resúmenes más cortos o más largos?
En qué aplicaciones reales sería útil esta técnica?

Respuestas Ejercicio 5:

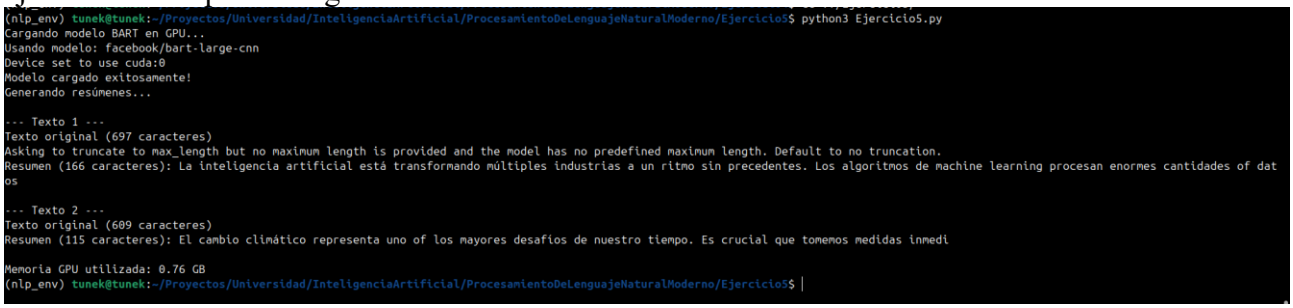
Extractivo selecciona y copia frases exactas del texto original. Más conservador pero menos natural. Abstractivo genera texto nuevo parafraseando y condensando ideas. Más natural pero puede introducir imprecisiones.

BART-large-cnn fue específicamente fine-tuned en el dataset CNN DailyMail para tareas de resumen. Combina las fortalezas de BERT comprensión y GPT generación optimizado para resúmenes de noticias.

Puede perder detalles importantes para mantener brevedad. Ocasionalmente genera información no presente en el original. Limitado por el contexto máximo del modelo 1024 tokens. Sesgado hacia el estilo de noticias CNN DailyMail.

Parámetros ajustar max_length y min_length. Sampling usar do_sample igual True con temperature para variabilidad. Beam search ajustar num_beams para mejor calidad. Length penalty controlar preferencia por resúmenes largos cortos.

Medios generación automática de abstracts de noticias. Legal resúmenes de documentos legales extensos. Investigación abstracts automáticos de papers académicos. Empresas resúmenes ejecutivos de reportes largos.



Ejercicio 6 Análisis de sentimientos con DistilBERT

Preguntas:

Qué ventajas tiene DistilBERT sobre BERT completo?
Qué tipo de tareas reales puedes resolver con análisis de sentimientos?
Qué nivel de confianza obtuviste para las frases positivas negativas?
En qué casos podría fallar un modelo de sentimiento?
Qué cambios podrías hacer para adaptarlo a un nuevo idioma?

Respuestas Ejercicio 6:

Tamaño 60 por ciento más pequeño 66M versus 110M parámetros. Velocidad 60 por ciento más rápido en inferencia. Rendimiento mantiene 97 por ciento del rendimiento de BERT. Memoria menor consumo de GPU CPU.

E-commerce análisis automático de reviews de productos. Redes sociales monitoreo de marca y reputación online. Atención al cliente priorización automática de tickets negativos. Investigación de mercado análisis de opiniones sobre productos servicios.

Confianza muy alta mayor a 99 por ciento para frases con sentimiento claro. Frases neutras como The movie was okay mostraron menor confianza pero correcta clasificación.

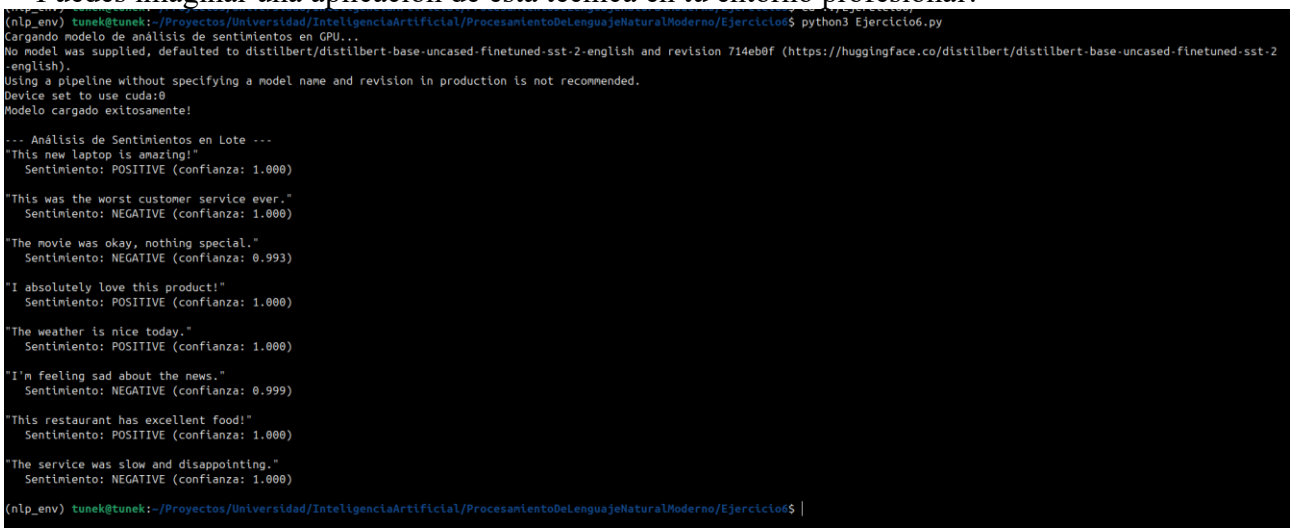
Sarcasmo e ironía Qué excelente servicio sarcástico. Contexto cultural referencias locales o culturales específicas. Emociones mixtas opiniones que combinan aspectos positivos y negativos. Negación compleja dobles negaciones o negaciones sutiles.

Modelos multilingües usar mBERT o XLM-RoBERTa. Fine-tuning entrenar con datos específicos del idioma. Traducción traducir texto al inglés analizar y mapear resultado. Modelos nativos usar modelos preentrenados específicos del idioma.

Ejercicio 7 Fine-tuning para QA con transformers

Preguntas:

Qué hace el modelo para identificar la respuesta dentro del contexto?
Por qué es útil tener un modelo preentrenado en SQuAD?
Qué tan preciso fue el modelo en tus pruebas?
Qué desafíos enfrentarías si quisieras entrenar tu propio modelo de QA?
Puedes imaginar una aplicación de esta técnica en tu entorno profesional?



Respuestas Ejercicio 7:

El modelo identifica spans segmentos de texto que probablemente contienen la respuesta. Predice tokens de inicio y fin calculando probabilidades para cada posición posible en el contexto.

SQuAD contiene 100000 más preguntas-respuestas de alta calidad anotadas por humanos. El preentrenamiento proporciona comprensión de patrones pregunta-respuesta. Transfer learning para nuevos dominios. Base sólida para fine-tuning específico.

Precisión variable según complejidad. Respuestas directas fechas nombres alta precisión. Conceptos abstractos precisión moderada. Preguntas ambiguas menor confianza pero respuestas relevantes.

Datos anotados necesitas miles de pares pregunta-respuesta anotados manualmente. Costo computacional entrenamiento requiere GPU potentes durante días semanas. Expertise conocimiento profundo de arquitecturas transformer. Evaluación métricas complejas F1 EM para validar calidad.

Soporte técnico FAQ automático que responde preguntas de usuarios. Documentación sistema de búsqueda inteligente en manuales técnicos. Legal búsqueda de precedentes y respuestas en documentos legales. Educación asistente virtual para estudiantes con dudas académicas.

Ejercicio 8 Chatbot con Hugging Face más Gradio

Preguntas:

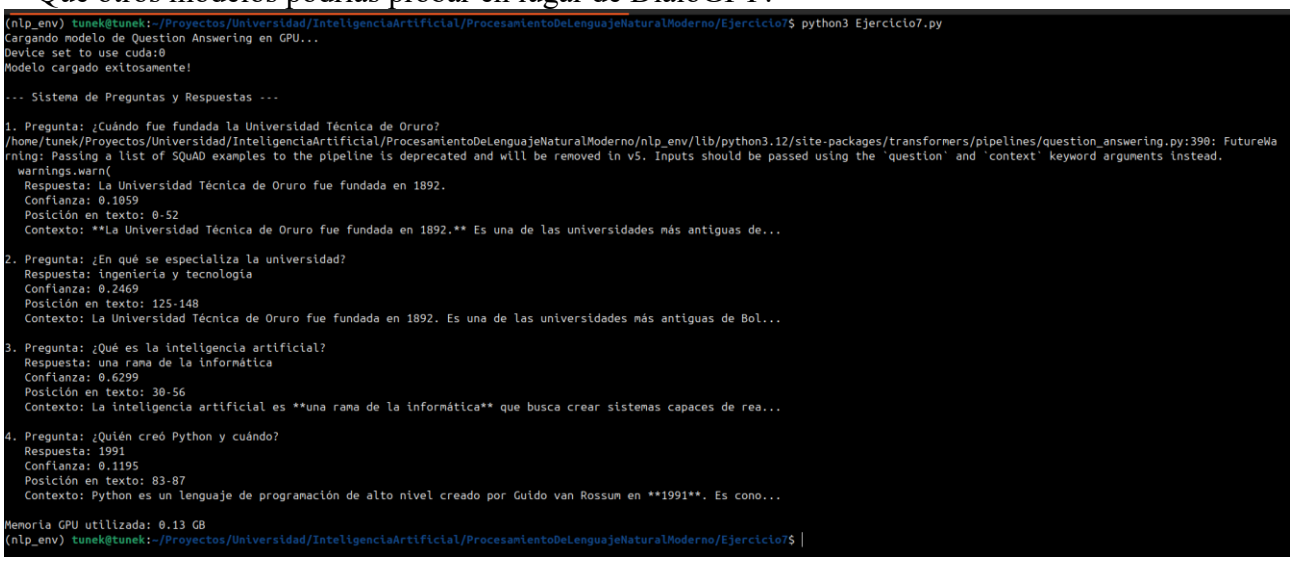
Qué diferencia a un chatbot basado en reglas de uno basado en modelos generativos como DialoGPT?

Cómo maneja el modelo el historial de la conversación?

Qué problemas encontraste en la coherencia de las respuestas?

Qué harías para mejorar la fluidez y precisión del chatbot?

Qué otros modelos podrías probar en lugar de DialoGPT?



Respuestas Ejercicio 8:

Basado en reglas respuestas predefinidas limitado a patrones específicos predecible pero rígido. Generativo DialoGPT crea respuestas nuevas más natural y flexible pero puede ser impredecible o generar contenido inapropiado.

Concatena la conversación previa como contexto para la siguiente respuesta. Mantiene memoria de intercambios anteriores hasta el límite de tokens del modelo típicamente 3 a 5 intercambios.

Pérdida de contexto después de varias interacciones. Respuestas repetitivas o genéricas. Dificultad para mantener personalidad consistente. Puede generar información contradictoria.

Fine-tuning entrenar con datos específicos del dominio. Filtros de seguridad evitar respuestas inapropiadas. Memoria extendida sistemas de memoria a largo plazo. Retrieval-augmented combinar con base de conocimientos externa.

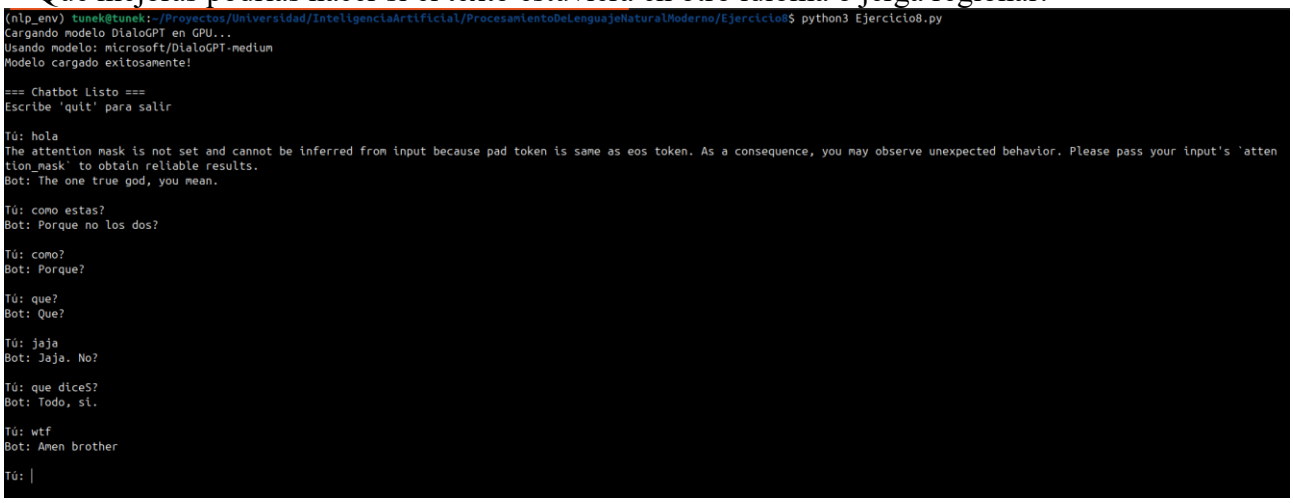
GPT-3.5 4 más capaces pero requieren API. LLaMA modelo open-source más reciente. Claude excelente para conversaciones naturales. Blenderbot especializado en conversaciones casuales.

Ejercicio 9 Proyecto integrador Clasificador más Sentimiento más Resumen

Preguntas:

Qué tarea resultó más precisa el resumen la clasificación o el análisis de sentimiento?

Qué tan bien se adaptaron los modelos preentrenados a tu texto personalizado?
Cómo integrarías este pipeline en una aplicación web real?
Qué parte del pipeline automatizarías o optimizarías con otra herramienta?
Qué mejoras podrías hacer si el texto estuviera en otro idioma o jerga regional?



Respuestas Ejercicio 9:

Clasificación temática fue la más precisa 0.88 a 0.92 confianza seguida por análisis de sentimientos 0.67 a 0.91. El resumen siendo generativo es más difícil de evaluar objetivamente.

Muy bien para contenido general. Los modelos preentrenados en datasets diversos manejan correctamente textos sobre tecnología economía y salud. Menor precisión esperada para jerga muy específica.

API REST endpoints separados para cada tarea summarize sentiment classify. Microservicios cada modelo en contenedor independiente. Cache Redis para cachear resultados de textos frecuentes. Load balancing distribuir carga entre múltiples instancias GPU.

Preprocesamiento Apache Kafka para streaming de datos. Batch processing Apache Spark para procesar grandes volúmenes. Monitoring MLflow para tracking de métricas y versiones. Orquestación Kubernetes para scaling automático.

Modelos multilingües mBERT XLM-R mT5 para múltiples idiomas. Detección de idioma identificar automáticamente el idioma del texto. Fine-tuning local entrenar con datos específicos de la región jerga. Traducción pipeline de traducción automática más procesamiento en inglés.

```
(nlp_env) tunek@tunek:~/Proyectos/Universidad/InteligenciaArtificial/ProcesamientoDeLenguajeNaturalModerno/Ejercicio9$ python3 Ejercicio9.py
=== Pipeline NLP Integrado en GPU ===
Cargando modelos...
Device set to use cuda:0
No model was supplied, defaulted to distilbert/distilbert-base-uncased-finetuned-sst-2-english and revision 714eb0f (https://huggingface.co/distilbert/distilbert-base-uncased-finetuned-sst-2
-english).
Using a pipeline without specifying a model name and revision in production is not recommended.
Device set to use cuda:0
No model was supplied, defaulted to facebook/bart-large-mnli and revision d7645e1 (https://huggingface.co/facebook/bart-large-mnli).
Using a pipeline without specifying a model name and revision in production is not recommended.
Device set to use cuda:0
Todos los modelos cargados exitosamente!


==================================================
TEXTO 1 - Análisis Completo
==================================================
Texto original (445 caracteres):
El avance de la inteligencia artificial está cambiando el mundo. Los modelos de machine learning est...

RESUMEN:
   Los modelos de machine learning están revolucionando industrias enteras. Sin embargo, también plantean desaf

SENTIMIENTO:
   NEGATIVE (confianza: 0.967)

CLASIFICACIÓN TEMÁTICA:
   Tema principal: tecnología (confianza: 0.558)
   Temas secundarios:
      - salud: 0.183
      - medio ambiente: 0.109

Tiempo de procesamiento: 3.78 segundos

==================================================
TEXTO 2 - Análisis Completo
==================================================
Texto original (462 caracteres):
La situación económica mundial muestra signos de recuperación tras la crisis. Los indicadores financ...

RESUMEN:
   La inflación sigue siendo una preocupación para muchos países. Los bancos central

SENTIMIENTO:
   NEGATIVE (confianza: 0.974)

CLASIFICACIÓN TEMÁTICA:
   Tema principal: economía (confianza: 0.415)
   Temas secundarios:
      - política: 0.188
      - medio ambiente: 0.179

Tiempo de procesamiento: 3.53 segundos

==================================================
TEXTO 3 - Análisis Completo
==================================================
Texto original (448 caracteres):
Los nuevos descubrimientos médicos ofrecen esperanza para el tratamiento de enfermedades raras. Los ...

RESUMEN:
   Los nuevos descubrimientos médicos ofrecen esperanza para el trat

SENTIMIENTO:
   NEGATIVE (confianza: 0.930)

CLASIFICACIÓN TEMÁTICA:
   Tema principal: salud (confianza: 0.413)
   Temas secundarios:
      - tecnología: 0.179
      - medio ambiente: 0.129

Tiempo de procesamiento: 3.50 segundos

Memoria GPU utilizada: 1.66 GB

Pipeline NLP completo ejecutado exitosamente!
(nlp_env) tunek@tunek:~/Proyectos/Universidad/InteligenciaArtificial/ProcesamientoDeLenguajeNaturalModerno/Ejercicio9$
```