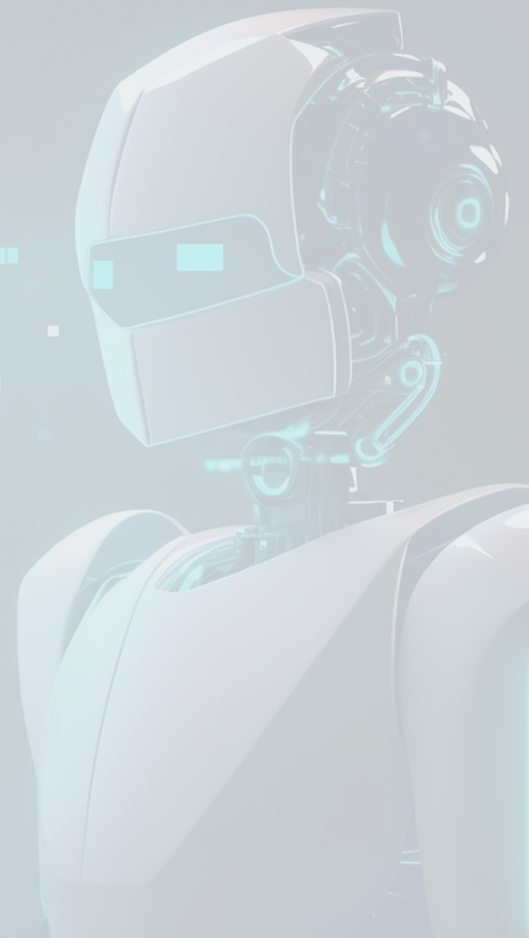




Introduction to Large Language Models and Agents

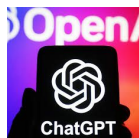
Ronald Cardenas Acosta, Ph.D.
Research Engineer in NLP
Huawei London



El avance agigantado de la Inteligencia Artificial



Moshi (2024)



ChatGPT (2023)



DALL-E 3

DALL-E (2021)



AlphaZero (2017)

LLaMa
Qwen
DeepSeek
Mistral
Claude
Gemini
O1
Nova

...



OpenAI's Operator
Anthropic's Computer Use
Huggingface's SmolAgents

Agentes

Formato del curso



Clases:

- 4 sesiones de teoría + demos (L-J)
- 1 sesión de tutoría dedicada a responder preguntas adicionales sobre material o sobre los trabajos (V)

Evaluación:

- 2 trabajos - implementación de conceptos básicos

Syllabus



Unidad 1	Conceptos Básicos	17/03
Unidad 2	Modelado de Lenguaje	18/03
Unidad 3	Pre-entrenamiento y Fine-tuning	19/03
Unidad 4	Post-entrenamiento, aplicaciones de LLMs	20/03
	Sesion de Tutoria & QA	21/03

Syllabus



Unidad 1	Conceptos Básicos
	<p>Teoría de Probabilidad</p> <p>Machine Learning</p> <p>Redes Neuronales y Backpropagation</p> <p>Teoria de la Informacion</p> <p>Tareas NLP</p> <p>Introduccion a Huggingface y Google Colab</p>

Syllabus



Unidad 2	Modelado de Lenguaje
	<p>Modelos de Lenguaje</p> <p>Word Embeddings</p> <p>Redes Recurrentes</p> <p>Transformers</p> <p>Trabajo de Laboratorio 1</p>

Syllabus



Unidad 3	Pre-entrenamiento y Fine-tuning
	<p>Pretraining</p> <p>Prompting</p> <p>Leyes de Escalamiento</p> <p>Métodos de Generación de Texto</p> <p>Generación de Texto: Demo</p>

Syllabus



Unidad 4	Post-entrenamiento, aplicaciones de LLMs
	<p>Post-training & Preference Optimizacion</p> <p>Entrenamiento Eficiente de LLMs</p> <p>Recursos e interfaces</p> <p>Aplicaciones de LLM en NLP</p> <p>Agentes</p>

Syllabus



Unidad 1	Conceptos Básicos
Unidad 2	Modelado de Lenguaje
Unidad 3	Pre-entrenamiento y Fine-tuning
Unidad 4	Post-entrenamiento, aplicaciones de LLMs

Unidad 1: Conceptos Basicos

Teoría de Probabilidad

Machine Learning

Redes Neuronales y Backpropagation

Teoría de la Información

Tareas NLP

Introduccion a Huggingface y Google Colab



Probabilidad

Lenguaje matemático usado para
cuantificar **incertidumbre**

χ : **Espacio de muestras**

Conjunto de todos los posibles
resultados de un experimento

X: variable aleatoria

Puede tomar cualquier valor de χ

Ejemplo: Espacio y variable discreta

Experimento: predecir el clima de hoy

$\chi = \{\text{soleado, lluvioso}\}$

$X = \text{el clima de hoy}$

$P(X=\text{soleado})$: *certidumbre* de que hoy sea
un día soleado

Probabilidad



Dados dos eventos A y B

Probabilidad de la unión

$$P(A \vee B) = P(A \cup B) = P(A) + P(B) - P(A \wedge B)$$

Probabilidad conjunta

$$P(A, B) = P(A \wedge B) = P(A \mid B) P(B)$$

Regla del producto

Probabilidad marginal

$$P(A) = \sum_b P(A, B=b) = \sum_b P(A \mid B=b) P(B=b)$$

*Regla de la suma o de
probabilidad total*

Probabilidad



Regla de la cadena

$$P(A, B, C, D) = P(A) P(B | A) P(C | A, B) P(D | A, B, C)$$

Regla de Bayes o Teorema de Bayes

$$p(X = x | Y = y) = \frac{p(X = x, Y = y)}{p(Y = y)} = \frac{p(X = x)p(Y = y | X = x)}{\sum_{x'} p(X = x')p(Y = y | X = x')}$$

$P(X=x | Y=y)$: probabilidad *a posteriori*

$P(X=x)$: probabilidad *a priori*

$P(Y=y | X=x)$: verosimilitud (*likelihood*)

$P(Y=y)$: evidencia

Probabilidad



Ejemplo: Teorema de Bayes

Te sometes a un test de cancer. Si el test sale *positivo*, cual es la probabilidad de que realmente tengas cáncer?

Probabilidad



Ejemplo: Teorema de Bayes

- Variable $X = \{0,1\}$: test sale positivo (1) o negativo (0)
- Variable $Y = \{0,1\}$: la persona tiene cancer (1) o no (0)
- Sensibilidad del test : $P(X=1 \mid Y=1) = 0.8$
- Probabilidad a priori de tener cáncer (estadísticas mundiales):

$$P(Y=1) = 0.004$$

- Prob. de que sea una falsa alarma (limitaciones de la tecnología de testeo):

$$P(X=1 \mid Y=1) = 0.1$$

Probabilidad



Ejemplo: Teorema de Bayes

Probabilidad de tener cáncer dado que el test es positivo es

$$P(Y = 1|X = 1) = \frac{P(X = 1|Y = 1)P(Y = 1)}{P(X = 1|Y = 1)P(Y = 1) + P(X = 1|Y = 0)P(Y = 0)}$$

$$P(Y = 1|X = 1) = \frac{0.8 \times 0.004}{0.8 \times 0.004 + 0.1 \times 0.996} = 0.031$$

Respuesta: **3.1%**

Distribución Bernoulli



- Distribución discreta de Probabilidad que toma valor de **1** si el experimento es exitoso con probabilidad **p**, y **zero (0)** si es fallido con probabilidad **q=1-p**
- Parametros: input $k=\{0,1\}$, probabilidad de éxito **p**
- Función de densidad

$$f(k; p) = p^k q^{1-k} = p^k (1 - p)^{1-k}$$

Distribución Binomial



- Distribución de probabilidad de éxito luego n experimentos Bernoulli sucesivos
- Probabilidad de que j experimentos fueron exitosos luego de n intentos, es

$$f(j, n; p) = \binom{n}{j} p^j q^{n-j} = \binom{n}{j} p^j (1 - p)^{n-j}$$

P: probabilidad de éxito de un solo intento

Distribución Categórica



- Extiende la definición de una distribución Bernoulli
 - Cada intento ahora puede tener n posibles resultados (o categorías)

$$f(x_i; p_1, \dots, p_n) = p_i.$$

Donde p_i es la probabilidad de obtener resultado i

Distribución Multi-nomial



- Generalización de la distribución Binomial y la Categórica
 - Resultado de **n** experimentos
 - Cada experimento puede tomar **k** valores

$$f(x_1, \dots, x_k; n; p_1, \dots, p_k) = \frac{n!}{x_1! \dots x_k!} p_1^{x_1} \cdot \dots \cdot p_k^{x_k}$$

Dond

x_i: número de veces que obtuvimos valor **i**

p_i: probabilidad de obtener valor **i** en cada intento

Estimando Distribuciones de Probabilidad

- Hasta ahora hemos asumido que los parámetros (θ) de una distribución conocidos y son usados para calcular $p(x | \theta)$
 - Existen casos en que θ es fácil de inferir (ej. Gaussian, Bernoulli, Binomial,...)
- En la mayoría de casos, $P(*)$ es complejo y tiene que aproximarse con **data**
- Ejemplo
 - X: N experimentos de tirar una moneda, k nos dieron cara
 - $\theta = p = k / N \rightarrow$ estimando θ con X
- Formalmente, queremos estimar $p(\theta | x)$: **Likelihood**
 - Mide cuán bien el parametro θ modela la data x

$$\mathcal{L}(\theta | x_1, \dots, x_n) = \prod_{i=1}^n f(x_i | \theta)$$

Estimando Distribuciones de Probabilidad

- Hasta ahora hemos asumido que los parámetros (θ) de una distribución conocidos y son usados para calcular $p(x | \theta)$
 - Existen casos en que θ es fácil de inferir (ej. Gaussian, Bernoulli, Binomial,...)
- En la mayoría de casos, $P(*)$ es complejo y tiene que aproximarse con **data**
- Ejemplo
 - X : N experimentos de tirar una moneda, k nos dieron cara
 - $\theta = p = k / N \rightarrow$ estimando θ con X
- Formalmente, queremos estimar $p(\theta | x)$: **Likelihood**
 - Mide cuán bien el parámetro θ modela la data x

$$\mathcal{L}(\theta | x_1, \dots, x_n) = \prod_{i=1}^n f(x_i | \theta)$$

Maximum Likelihood Estimation (MLE)



- Dado un conjunto de muestras \mathbf{x} , el parámetro θ que mejor modela \mathbf{x} está definido por

$$\hat{\theta}_{mle} = \operatorname{argmax}_{\theta} \mathcal{L}(\theta | \mathbf{x})$$



Machine Learning

Aprendizaje de Máquina (Machine Learning)



“Se dice que un programa de computador aprende de una *experiencia* E con respecto a una *clase de tareas* T y la *medida de desempeño* P , si su desempeño en las tareas T , medido por P , mejora con la experiencia E .”

~Thomas M. Mitchell

Aprendizaje de Máquina (Machine Learning)



- Conjunto de métodos para detectar patrones en data automáticamente, y usarlos para realizar predicciones

Supervisado	X: entrada Y: categoria, label de cada entrada X
No Supervisado	X: entrada
Por refuerzo	X,M,(Y,R) M: modelo o ambiente Y,R : output + recompensa

Aprendizaje Supervisado



- Dado una entrada x , asignar una categoría y
 - Ejemplo: clasificación de documento
 - Sea $P(y | x)$ el modelo de clasificación

$$\hat{y} = \arg \max_{y \in \mathcal{Y}} P(y|x)$$

Aprendizaje No Supervisado

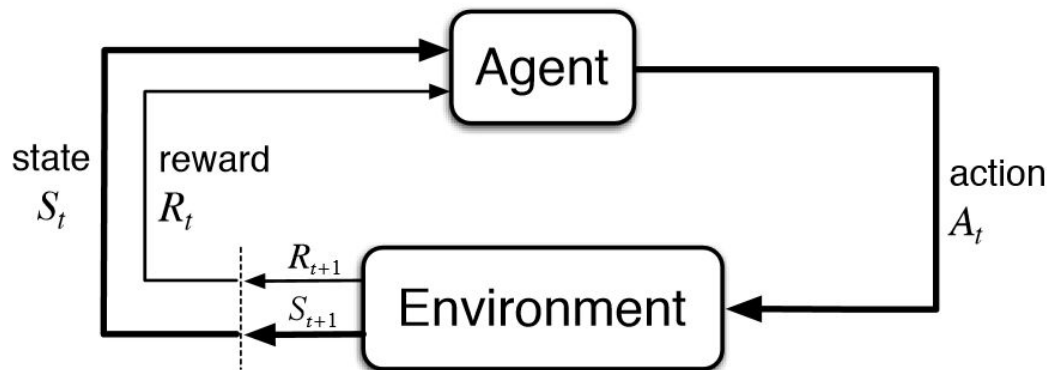



- Solo se dispone de X
- Objetivo: modelar características o patrones de X
- Ejemplos:
 - Ajustar una distribución de probabilidad a X
 - Agrupar muestras x de acuerdo a similaridad (clustering)
 - Obtener el espacio vectorial que describe x (manifold learning)
 - Obtener las direcciones más influyentes en un espacio vectorial (component analysis / decomposition)

https://scikit-learn.org/stable/unsupervised_learning.html

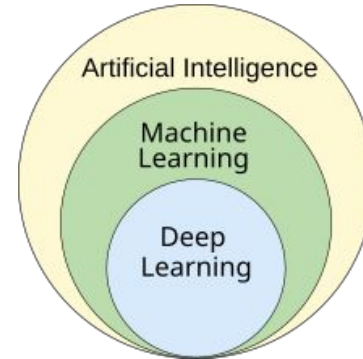
Aprendizaje por Refuerzo: Reinforcement Learning

- Paradigma distinto a supervisado / no supervisado
- Un modelo (**agente**) aprende mediante interacción con el **ambiente**
 - Cada **acción** está asociada a una “**recompensa**”
 - Recompensa alta -> mejor la probabilidad de escoger esa acción



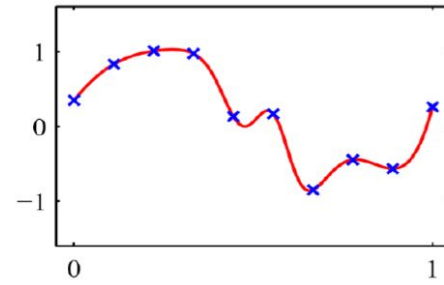
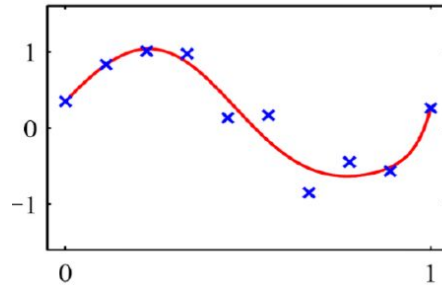
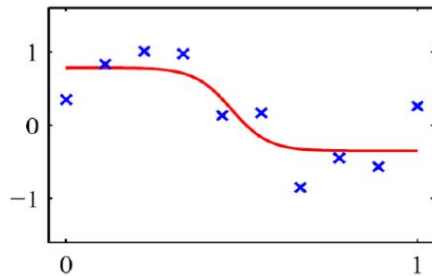


Aprendizaje Profundo: Redes Neuronales



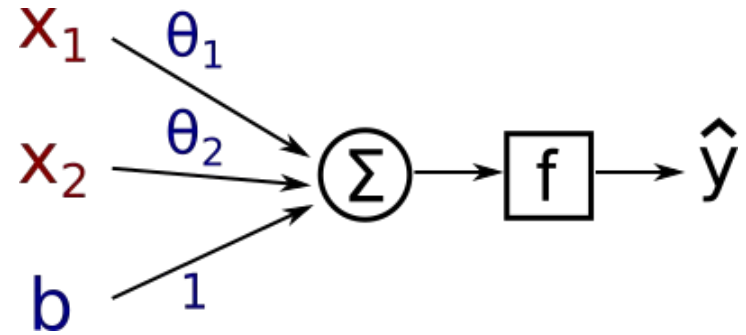
Redes Neuronales

- Diseñadas para realizar approximation de funciones
 - Ej. $p(Y=y | x)$ es una función, $f(x,y)$
- Expresividad depende de la arquitectura (ej. numero de capas, funcion de activacion, etc)



Perceptron

- Dado el datum $\langle x, y \rangle$
 - $X = [x_1, x_2]$
 - $Y = \{0, 1\}$
- Modelo con parametros
 - $w = [\theta_1, \theta_2]$: weights
 - b : bias
 - f : funcion de activacion

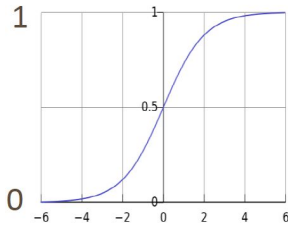


$$\hat{y} = f(w^T x + b)$$

- Sin funcion de activation
 - **Simplemente una transformación lineal!**

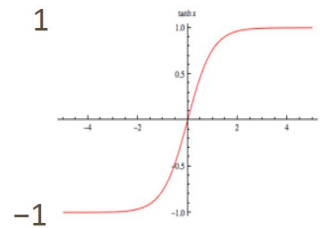
Funcion de Activacion

$$f(z) = \frac{1}{1 + \exp(-z)}$$



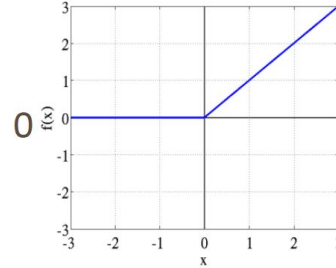
Logistic o sigmoid

$$\tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$$



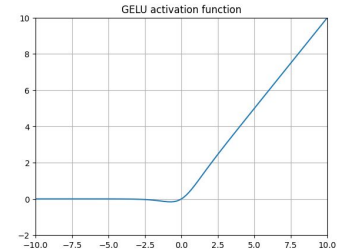
Tangente Hiperbolica
tanh

$$\text{ReLU}(z) = \max(z, 0)$$



Rectified Linear Unit

$$\begin{aligned} \text{GELU}(x) &= x \cdot P(X \leq x), X \sim N(0,1) \\ &\approx x \cdot \text{logistic}(1.702x) \end{aligned}$$

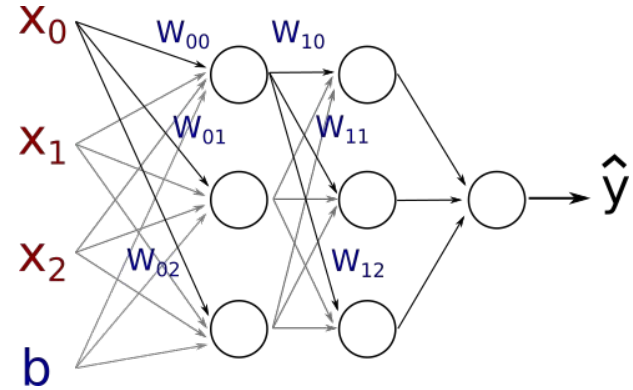


Cuando num. Clases > 2:
softmax

Y muchas mas... campo activo de investigación

Multi-layer Perceptron

- Dado el datum $\langle x, y \rangle$
 - $X = [x_0, x_1, x_2]$
 - $Y = \{0, 1\}$
- Modelo con parametros
 - W^0 = matriz de $[n_{\text{input}}] \times [n_{\text{neuron}}]$
 - W^1 = matrix $[n_{h0} - n_{h1}]$
 - W^2 = matrix $[n_{h1} - n_{\text{classes}}]$
 - b : bias
 - f : funcion de activacion



$$h_0 = f(W^0 x + b^0)$$

$$h_1 = f(W^1 h_0 + b^1)$$

$$\hat{y} = f(W^2 h_1 + b^2)$$

Entrenando una Red Neuronal



- Sea la función objetivo, o costo, $J(\theta) = Cost(\hat{y}, y; \theta)$
- Buscamos los parámetros θ que *minimicen* $J(\theta)$
 - **Recap:** previamente buscábamos maximizar likelihood $L(\theta)$
- Como aproximar θ ?
 - Métodos de gradiente

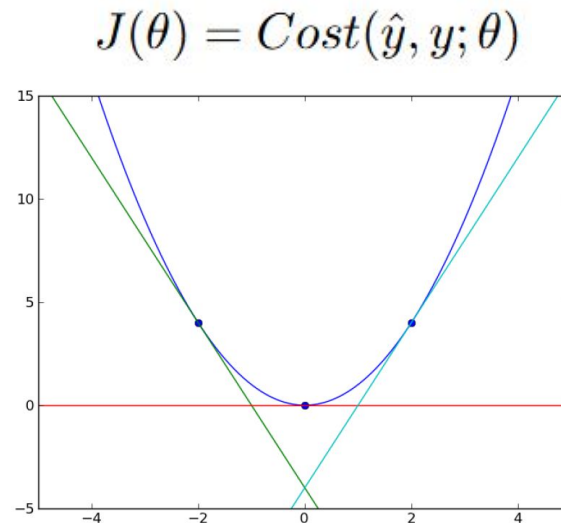
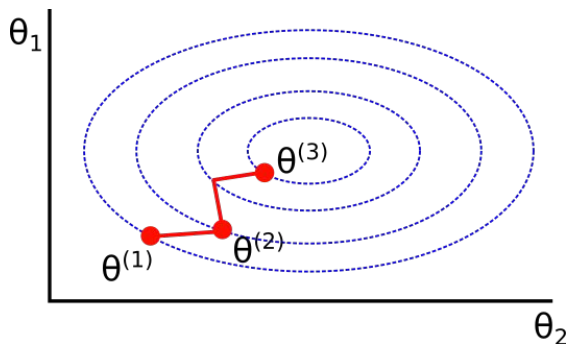
Gradient Descent:

$$\theta^{(i+1)} \leftarrow \theta^{(i)} - \alpha \nabla_{\theta} J$$

α : learning rate (tasa de aprendizaje)

Entrenando una Red Neuronal

- $J(^*)$ converge a un **valor mínimo** asociado a un grupo de parámetros θ
 - Si $J(^*)$ es convexo: mínimo **global**
 - Si $J(^*)$ no es convexo: mínimo **local**



Entrenando una Red Neuronal



- Dada la expresion $\theta^{(i+1)} \leftarrow \theta^{(i)} - \alpha \nabla_{\theta} J$
como calculamos la gradiente $\nabla J(\theta)$?
- Aplicamos *backpropagation*, simplemente

$$\nabla J(\theta) = 0$$

Idea principal: propagar el error (cuando $\hat{Y} \neq Y$) desde la última capa hasta la entrada

Backpropagation



Ejemplo: Clasificación binaria, function de activ=***sigmoid***
Logistic Regression

$$J(\theta) = Cost(\hat{y}, y; \theta)$$

$$J(\theta) = \frac{1}{N} \sum_i^N y_i \log \hat{y}_i - (1 - y_i) \log(1 - \hat{y}_i)$$

$$\begin{aligned} \nabla J(\theta) &= \frac{\partial J}{\partial \theta} = \frac{\partial J}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial \theta} \\ &= \frac{1}{N} \sum_{i=1}^N \hat{y}_i (y_i - \hat{y}_i) \cdot x_i \end{aligned}$$

Regla de la cadena en cálculo

Backpropagation



<https://xnought.github.io/backprop-explainer/>



Information Theory, *the essentials*

Nocion de Entropia



- Entropia ~ “caos”, lo opuesto al orden, confusion
- Importado de fisica
 - La entropía no disminuye a menos que se aplique energía
- Medida de incertidumbre
 - Baja entropía -> baja incertidumbre
 - alta entropía -> alta incertidumbre
 - Pero mas alta la “sorpresa” (informacion) que podemos obtener de un experimento

Entropia



- Sea:
 - $p(X)$: distribución de probabilidad sobre la variable aleatoria X
 - χ : conjunto de posibles resultados (*espacio muestral*)

$$H(X) = - \sum_{x \in \chi} p(x) \log_2 p(x)$$

- Unidades: bits (log10: nats)
- Notacion: $H(X) = H(p)$

Entropia: ejemplo

- Moneda sin adulterar: $\mathcal{X} = \{\text{cara, sello}\}$
 - $p(x=\text{cara}) = 0.5, p(x=\text{sello}) = 0.5$
 - $H(p) = -0.5 \log(0.5) + (-0.5 \log(0.5))$
 - $= 2 * (-0.5) * (-1) = 1$

- Dado sin adulterar de 32 lados
 - $p(x) = 1/32$, para cada lado x

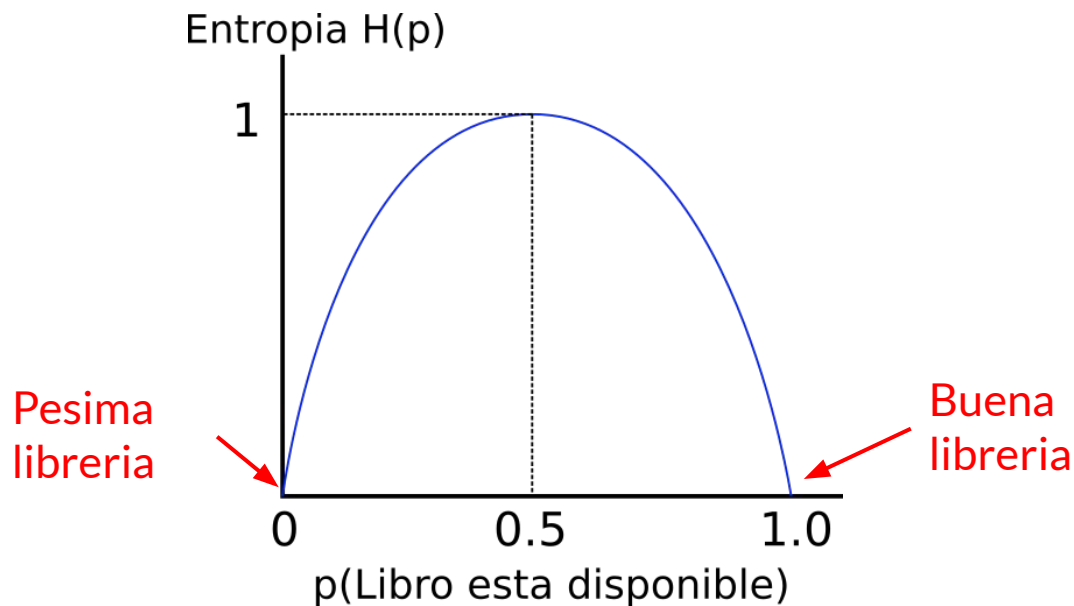
$$\begin{aligned} H(p) &= - \sum_{i=1 \dots 32} p(x_i) \log p(x_i) = -32 p(x_1) \log p(x_1) \\ &= -32 * (1/32) * (-5) = 5 \end{aligned}$$

Alguna idea de porque se mide en **bits**?

- Moneda adulterada
 - $p(\text{cara}) = 0.2 \rightarrow H(p) = 0.722$
 - $p(\text{cara}) = 0.01 \rightarrow H(p) = 0.081$

Entropia: ejemplo

- X = “disponibilidad de libros en una librería”



Entropia: casos limite

- Es posible que $H(p) = 0$?
 - Si *se sabe de antemano* el resultado de un experimento

$$\exists x \in \chi; p(x) = 1 \ \& \ \forall y \in \chi; y \neq x \Rightarrow p(y) = 0$$

- Limite superior
 - Ninguno en general
 - Para

$$|\chi| = n : H(p) \leq \log n$$

Nada es más incierto que la distribución uniforme

Perplexity (Perplejidad): Motivacion



- Recordando:
 - 2 resultados equiprobables: $H(p) = 1$ bit
 - 32 resultados equiprobables: $H(p) = 5$ bits
 - 4.3 billones de resultados equiprobables: $H(p) \sim 32$ bits
- Y si los resultados no son igualmente probables?
 - 32 resultados, 2 equiprobables en 0.5, el resto imposible
 - $H(p) = 1$ bit
- Necesitamos una métrica que compare la entropía (incertidumbre, dificultad de predicción) de variables con *diferente número de posibles resultados*

Perplexity



$$PPL(p) = 2^{H(p)}$$

- Una métrica más intuitiva que entropía
 - PPL=32 cuando hay 32 resultados equiprobables
 - PPL=2 para una moneda sin adulterar
 - ...
- Ampliamente usado en NLP, ejemplo
 - PPL = tamaño del vocabulario cuando todas las palabras son equiprobables
- Mientras más sesgada (biased) la distribución, mejor
 - Menos entropía, menos perplejidad

Entropia Conjunta y Condicional



- Dadas dos variables aleatorias X, Y con espacios muestrales χ, ψ
- Entropia Conjunta (Joint Entropy)

$$H(X, Y) = - \sum_{x \in \chi} \sum_{y \in \psi} p(x, y) \log p(x, y)$$

- Entropia Condicional (Conditional Entropy)

$$H(Y|X) = - \sum_{x \in \chi} \sum_{y \in \psi} p(x, y) \log p(y|x)$$

Entropia Conjunta y Condicional



- Dadas dos variables aleatorias X, Y con espacios muestrales χ, ψ
- Entropia Conjunta (Joint Entropy)

$$H(X, Y) = - \sum_{x \in \chi} \sum_{y \in \psi} p(x, y) \log p(x, y)$$

- Entropia Condicional (Conditional Entropy)

$$H(Y|X) = - \sum_{x \in \chi} \sum_{y \in \psi} p(x, y) \log p(y|x)$$

Entropía: Propiedades



- Regla de la cadena

$$H(X, Y) = H(Y|X) + H(X)$$

$$H(X, Y) = H(X|Y) + H(Y)$$

- $H(Y|X) \leq H(Y)$
- $H(X, Y) \leq H(X) + H(Y)$

Entropía como “*codificación*”



- “ $H(p)$ es el número mínimo (en promedio) de bits necesarios para codificar un mensaje (texto, secuencia, senhal, ...)”
- Algoritmos de compresión
 - Más efectivos en data con patrones repetidos (fácilmente predecibles = baja entropía)

Entropía de un Lenguaje

- Digamos que generamos la siguiente letra usando

$$p(l_{n+1}|l_1, \dots, l_n)$$

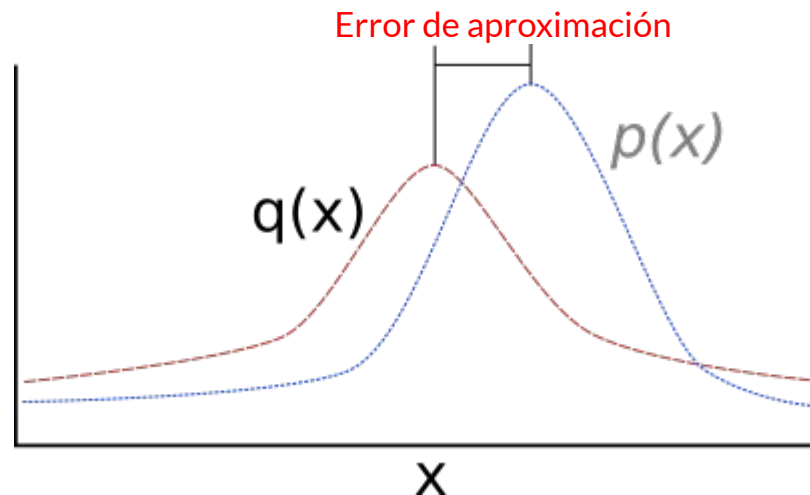
Donde l_1, \dots, l_n es la secuencia de **todas las letras** escritas hasta ahora

- Sea l_1, \dots, l_n el *historial*, y H el conjunto de todos los historiales
- Entonces, la entropía se define como

$$-\sum_{h \in H} \sum_{l \in \Lambda} p(l, h) \log p(l|h) \quad \text{impractico!}$$

Divergencia de Kullback-Leibler: Intuición

- $p(x)$: distribución de probabilidad que queremos aproximar
 - Descripción completa desconocida
 - Solo tenemos muestras
- Si utilizamos distribución $q(x)$ como substituta de $p(x)$
 - Cómo cuantificamos el error cometido?



Divergencia de Kullback-Leibler

- Medida de disimilaridad entre dos distribuciones de probabilidad

$$\mathbb{KL}(p||q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)}$$

$$\begin{aligned}\mathbb{KL}(p||q) &= \sum_{x \in \mathcal{X}} p(x) \log p(x) - \sum_{x \in \mathcal{X}} p(x) \log q(x) \\ &= -H(p) + H(p, q)\end{aligned}$$

- $H(p, q)$: cross-entropia
 - “Cantidad de bits necesarios para codificar p si usamos q ”
- “Distancia KL”, “Entropía Relativa”
 - Estrictamente no es medida de distancia: $\mathbb{KL}(p||q) \neq \mathbb{KL}(q||p)$



Tareas NLP

Qué *tareas* resolvemos en *NLP*?



- NLP = Natural Language Processing
 - “Natural” == humano (*antropocentrico*)
 - En contraste con otros lenguajes en el reino animal, o lenguajes sintéticos (ej. lenguajes de programación)
- Tarea = Task = habilidad general de procesamiento de lenguaje, evaluada con un grupo específico de datasets / benchmarks
- En esta sección : tareas modernas que el campo de NLP trabaja

NLP tasks



- Clasificación
 - De texto en diferentes granularidades:
 - palabras, frases, oraciones, documentos
 - De relaciones entre partes del texto
- Clustering
- Predicción estructurada
 - Syntactic parsing: predicción del árbol gramático
- Secuencia - Secuencia
 - Traducción
 - Pregunta - respuesta
 - Resumen
 - ...

Clasificación de Texto



Objetivo: Dado un texto, asignarle una categoría de entre un grupo predeterminado de categorías

*Indecopi niega adquisición empresarial por primera vez, ¿Grupo Gloria puede apelar?
Este es el primer caso en que Indecopi rechaza una operación de concentración desde que se implementaron estos controles hace tres años...*



Finanzas

Categorías disponibles: Deportes, Finanzas, Política

Clasificación de Texto



Formalmente

$$\hat{y} = \arg \max_{y \in \mathcal{Y}} P(y|x)$$

- x : secuencia de palabras
- $\mathcal{Y} = \{\text{Deportes, Finanzas, Politica}\}$, set de categorías
- $P(y|x)$: probabilidad de elegir categoría y dado x

Clasificación de Texto



Podemos definir

$$P(y|x) \approx W \cdot f(x, y)$$

Entonces podemos encontrar un **W** que minimize

$$\mathcal{L} = - \sum y \cdot P(y|x) - (1 - y) \cdot (1 - P(y|x))$$

(Cross-Entropy Loss)

Usando Gradient Descent

$$W^{t+1} \leftarrow W^t - \alpha \cdot \nabla \mathcal{L}$$



Tareas NLP

Sentiment Analysis

Análisis de Opinión (Sentiment Analysis)



Objetivo: Dada una reseña de un producto, asignarle una categoría que indique la satisfacción del usuario o consumidor.

<i>Me encanta el nuevo diseño de su página web.</i>	Positivo
<i>No estoy seguro si me gusta el nuevo diseño.</i>	Neutral
<i>El nuevo diseño es horrible.</i>	Negativo

- x : reseña
- $Y = \{\text{Positivo, Neutral, Negativo}\}$

Análisis de Opinión (Sentiment Analysis)

Objetivo: Dada una reseña de un producto, asignarle una categoría que indique la satisfacción del usuario o consumidor.

<i>Si te gusta ejercitarte con música, estos earbuds son una buena opción...</i>	5 stars
<i>Para mi, fue una perdida de dinero...</i>	1 star
<i>En general estoy satisfecho con este producto...el único defecto...</i>	3 stars

- x : reseña
- $Y = \{1, 2, 3, 4, 5\}$



Tareas NLP

Inferencia Retorica:
Natural Language Inference

Implicacion retorica



Objetivo: Dadas dos oraciones, la *hipótesis* y la *premisa*, predecir si la hipótesis puede ser *inferida logicamente* de la premisa

Premisa

“Hugo es una gato siames” ->

Hipotesis

“Hugo es gato”

Tambien conodocoda como:

- *Natural Language Inference*
- *Rhetorical Text Entailment*

Implicacion retorica



Objetivo: Dadas dos oraciones, la *hipótesis* y la *premisa*, predecir si la hipótesis puede ser *inferida logicamente* de la premisa

Data: {X=(H,P), Y={Entailment, Contradiction, Neutral}}

Entailment	H solo es verdadero si P es verdadero
Contradiction	H puede ser verdadero sin importar P
Neutral	H y P son independientes

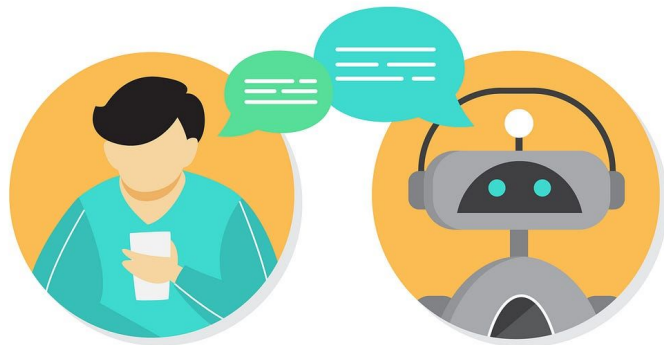


Tareas NLP

Sistemas de
Pregunta-Respuesta:
Question Answering

Question Answering

- Dada una pregunta o query **Q** y un contexto **C**,
 - Retornar un texto **A** con la información requerida



Question Answering



IBM

Watson, Jeopardy! champion

In a televised Jeopardy! contest viewed by millions in February 2011, IBM's Watson DeepQA computer made history by defeating the TV quiz show's two foremost all...

9 Dec 2023



Yahoo Movies UK

IBM Supercomputer 'Watson' Sweeps Jeopardy Practice Round

Meet Watson, aspiring Jeopardy champ. He's a black rectangle with a globe avatar, and he absorbed 200 million pages of text to prepare for...

1 week ago



Formulacion QA

Entrada

Query / Pregunta
+
Contexto
(articulo o documento)

Cuál es el mejor stock para comprar ahora?

Los mejores stocks para comprar ahora, de acuerdo a analistas

Si estás buscando una opinión experta, aquí están los 7 mejores stocks en el S&P 500 rankeados por analistas.

Compania, Recomendacion

Alexandria Real Estate Equities Inc., 1.17

Zoetis Inc., 1.18

Amazon.com Inc., 1.22

...

NerdWallet.com

Salida

Respuesta

El mejor stock para comprar ahora es stock de la compania Alexandria Real Estate Equities Inc

Tipos de Pregunta



Tipo	Descripción	Ejemplos
Factoide	Solicita información puntual <i>Quien, Que, Cuando, Donde, Por que</i>	Cuál fue el mejor stock del último trimestre?
Subjetiva	Solicita una opinión acerca de un tema	Qué piensas del clima financiero en latinoamérica?

Tipos de Pregunta



Tipo	Descripción	Ejemplos
Factoide	Solicita información puntual <i>Quien, Que, Cuando, Donde, Por que</i>	Cuál fue el mejor stock del último trimestre?
Subjetiva	Solicita una opinión acerca de un tema	Qué piensas del clima financiero en latinoamérica?

Question Answering: enfocado a preguntas Factoides

Datasets QA disponibles



Cuentos, Novelas (texto narrativo)

- Preguntas en base a personajes de una historia
- Fuente: libros, cuentos, novelas
- NarrativeQA, CoQA, Children's Book Test (CBT)

Conocimiento Enciclopédico

- Preguntas acerca de personas, lugares, organizaciones
- Fuente: Wikipedia, Bases de datos, foros, búsquedas en google
- Squad v2, HotpotQA, ComQA, WebQuestions

Noticias

- Preguntas en base a eventos y personajes en un artículo noticiero
- Fuente: paginas de cadenas de noticias, twitter
- CNN / DailyMail, BBC

Conocimiento Enciclopédico (Multilingue)

- XQuad: traducido de Squad v1.1
- 12 lenguajes, incluyendo Espanol
- <https://github.com/google-deepmind/xquad>

Datasets QA disponibles



Contexto

es El calor necesario para hervir el agua y suministrar el vapor puede derivarse de varias fuentes, generalmente de **[la quema de materiales combustibles]**₁ con un suministro adecuado de aire en un espacio cerrado (llamado de varias maneras: **[cámara de combustión]**₂, chimenea...). En algunos casos la fuente de calor es un reactor nuclear, energía geotérmica, **[energía solar]**₃ o calor residual de un motor de combustión interna o proceso industrial. En el caso de modelos o motores de vapor de juguete, la fuente de calor puede ser un calentador **[eléctrico]**₄.

Preguntas

1. ¿Cuál es la fuente de calor habitual para hacer hervir el agua en la máquina de vapor?
2. Aparte de cámara de combustión, ¿qué otro nombre que se le da al espacio en el que se quema el material combustible en el motor?
3. Junto con el calor residual de la energía nuclear, geotérmica y de los motores de combustión interna, ¿qué tipo de energía podría suministrar el calor para una máquina de vapor?
4. ¿Qué tipo de elemento calefactor se utiliza a menudo en las máquinas de vapor de juguete?

QA: Evaluacion



Exact Match (EM)

- 1.0 si la respuesta generada es *exactamente* igual a la referencia gold
- 0.0 de otra forma

Ejemplo

- Q: *Quien es presidente de USA?*
- Referencia Gold: *"Joe Biden"*
- Respuesta Generada: *"Biden"*
- ***EM=0.0***

QA: Evaluacion



F1 Score (F1)

- Promedio de palabras generadas que aparecen también en el gold

$$F_1 = \frac{2 * P * R}{P + R}$$

$$P = \frac{\# \text{Palabras correctas}}{\text{Palabras en respuesta}}$$

$$R = \frac{\# \text{Palabras correctas}}{\text{Palabras en Gold}}$$

QA: Evaluacion



F1 Score (F1)

- Promedio de palabras generadas que aparecen también en el gold

$$F_1 = \frac{2 * P * R}{P + R}$$

Ejemplo

- Q: *Quien es presidente de USA?*
- Referencia Gold: *"Joe Biden"*
- Respuesta Generada: *"Biden"*
- $P = 1 / 1$, $R = 1 / 2$
- **$F1 = 0.66$**

QA: Metodos



Discriminativo o Extractivo

- Extrae una frase contigua del contexto

es El calor necesario para hervir el agua y suministrar el vapor puede derivarse de varias fuentes, generalmente de **[la quema de materiales combustibles]**₁ con un suministro adecuado de aire en un espacio cerrado (llamado de varias maneras: **[cámara de combustión]**₂, chimenea...). En algunos casos la fuente de calor es un reactor nuclear, energía geotérmica, **[energía solar]**₃ o calor residual de un motor de combustión interna o proceso industrial. En el caso de modelos o motores de vapor de juguete, la fuente de calor puede ser un calentador **[eléctrico]**₄.

1. ¿Cuál es la fuente de calor habitual para hacer hervir el agua en la máquina de vapor?

A: la quema de materiales combustibles

QA: Metodos



Generativo o Abstractivo

- Genera un texto nuevo como respuesta

es El calor necesario para hervir el agua y suministrar el vapor puede derivarse de varias fuentes, generalmente de **[la quema de materiales combustibles]**₁ con un suministro adecuado de aire en un espacio cerrado (llamado de varias maneras: **[cámara de combustión]**₂, chimenea...). En algunos casos la fuente de calor es un reactor nuclear, energía geotérmica, **[energía solar]**₃ o calor residual de un motor de combustión interna o proceso industrial. En el caso de modelos o motores de vapor de juguete, la fuente de calor puede ser un calentador **[eléctrico]**₄.

1. ¿Cuál es la fuente de calor habitual para hacer hervir el agua en la máquina de vapor?

A: La combustion de materiales

QA: Desafíos de hoy en día



- Una pregunta puede ser expresada de muchas formas
 - Parafrasis, sinonimos
- Granularidad de información
 - Cuánta información quiere el usuario?
- Topics, áreas de conocimiento
 - Sistemas funcionan solo para los tópicos en los fueron entrenados



Tareas NLP

Resumen Automatico:
Summarization

Resumen Automatico



- Dada un texto **D**
 - Retornar un texto corto **S** con la información general de **D**
- Resumen **S** debe
 - Cubrir todos los eventos y tópicos en **D**
 - Tener informacion *relevante*
 - Ser *coherente*
 - No tener información *redundante*

Formulación de Resumen Automatico



Entrada Texto (documento)

Un niño de tres años ha fallecido esta noche en Gijón tras atragantarse con una de las tradicionales uvas de Nochevieja en su domicilio, en la calle Independencia del barrio del Natahoyo, han informado fuentes municipales. La familia, de origen...

Dataset: **MLSSUM**

Link:

https://elpais.com/politica/2019/01/01/actualidad/1546340064_189703.html

Salida Resumen

Muere un niño de tres años atragantado con una uva en Gijón

Tipos de Resumen



Tipo	Descripción
General <i>General Summarization</i>	<ul style="list-style-type: none">• Abarca todos los tópicos en el documento• Hecho para cualquier usuario
Basado en Aspecto <i>Aspect-based Summ.</i>	<ul style="list-style-type: none">• Abarca sólo ciertos aspectos o tipos de información• Ejemplos<ul style="list-style-type: none">○ Resumir lo que usuarios piensan del precio de un producto○ Resumir la vida personal de un escritor a partir de su biografía
Basado en Query <i>Query-based Summ.</i>	<ul style="list-style-type: none">• Abarca toda la información relacionada al query

Datasets de Summarization



Noticias

- Fuente: artículos de cadenas de noticias
- CNN / DailyMail, XSum, NewsRoom, NYC

Artículos Científicos

- Fuente: PubMed, arXiv, Plos, eLife
- Resumen: formato *Abstract*

Conocimiento Enciclopédico

- Fuente: Wikipedia
- WikiSum
- Resumen: formato

Periodismo Científico

- Fuente: noticieros científicos
- SciTechNews
- Documento: artículo científico
- Resumen: texto noticiero para usuarios no expertos

Noticias Multilingue

- Entrada: artículo en un lenguaje
- Resumen: en otro lenguaje
- MultiNews, MLSum

Datasets de Summarization



– SPANISH –

summary El aeropuerto ha estado hasta las 15.00 con sólo dos pistas por ausencia de 5 de los 18 controladores aéreos.- Varias aerolíneas han denunciado demoras de "hasta 60 minutos con los pasajeros embarcados"

body El espacio hará un repaso cronológico de la vida de la Esteban desde el momento en el que una completa desconocida comenzó a aparecer en los medios en 1998 como la novia de Jesulín de Ubrique hasta llegar a hoy en día, convertida en la princesa del pueblo, en concreto del popular madrileño distrito de San Blas donde vive, tal y como algunos la han calificado, y protagonista de portadas de revistas, diarios y portales web y de aparecer incluso entre los personajes más populares de Google. Junto a María Teresa Campos, estarán en el plató Patricia Pérez, presentadora del programa matinal de los sábados en Telecinco Vuélveme loca, quien ha conducido las campanadas en cuatro ocasiones, y los comentaristas Maribel Escalona, Emilio Pineda y José Manuel Parada. Los vuelos han venido registrando este viernes importantes retrasos en Barajas a pesar de que desde las 15.00 el aeropuerto opera con las cuatro pistas, según han informado fuentes de AENA, mientras las compañías han denunciado demoras por parte de los controladores de hasta 60 minutos con los pasajeros embarcados. Según los datos facilitados por AENA, la ausencia por la mañana de 5 de los 18 controladores que estaban programados en el turno de la torre de control de Barajas obligó a cerrar dos de las pistas del aeropuerto, lo que generó retrasos medios de 30 minutos.

Ejemplo de [MLSum](#)

Summarization: Evaluacion



ROUGE: Basado en n-grams

- Promedio de frases presentes *verbatim* en la referencia
- Indica relevancia, fluidez del resumen
 - ROUGE-1: # frases de una palabra (unigram)
 - ROUGE-2: # frases de dos palabras (bi-gram)
 - ROUGE-L: frase comun mas larga

Summarization: Evaluacion



BERTScore

- Mide similaridad semántica entre resumen generado y referencia
- Usa BERT como modelo semantico
- Mayor similaridad -> más relevancia del resumen

Summarization: Evaluacion



Factualidad

- Busca evaluar que los datos en el resumen sean correctos con respecto al texto documento de entrada
- Ejemplos
 - [QuestEval](#)
 - [SummaC](#)

Summarization: Evaluacion



Anotadores Humanos

- Anotador califica un resumen de acuerdo a criterios
 - Relevancia
 - Coherencia
 - Repetitividad
 - Fluidez
 - Etc
- La escala de calificación depende del método
 - Escala del 1 - 5 o 1-3
 - Comparación en pares
 - etc

Summarization: Metodos

Extractivo

- Extrae una parte del texto y lo presenta como resumen

body El espacio hará un repaso cronológico de la vida de la Esteban desde el momento en el que una completa desconocida comenzó a aparecer en los medios en 1998 como la novia de Jesulín de Ubrique hasta llegar a hoy en día, convertida en la princesa del pueblo, en concreto del popular madrileño distrito de San Blas donde vive, tal y como algunos la han calificado, y protagonista de portadas de revistas, diarios y portales web y de aparecer incluso entre los personajes más populares de Google. Junto a María Teresa Campos, estarán en el plató Patricia Pérez, presentadora del programa matinal de los sábados en Telecinco Vuélveme loca, quien ha conducido las campanadas en cuatro ocasiones, y los comentaristas Maribel Escalona, Emilio Pineda y José Manuel Parada. Los vuelos han venido registrando este viernes importantes retrasos en Barajas a pesar de que desde las 15.00 el aeropuerto opera con las cuatro pistas, según han informado fuentes de AENA, mientras las compañías han denunciado demoras por parte de los controladores de hasta 60 minutos con los pasajeros embarcados. Según los datos facilitados por AENA, la ausencia por la mañana de 5 de los 18 controladores que estaban programados en el turno de la torre de control de Barajas obligó a cerrar dos de las pistas del aeropuerto, lo que generó retrasos medios de 30 minutos.

Resumen:

Los vuelos han venido registrando este viernes importantes retrasos en Barajas a pesar de que desde las 15.00 el aeropuerto opera con las cuatro pistas, según han informado fuentes de AENA,

Summarization: Metodos



Abstractivo

- Genera un texto nuevo

body El espacio hará un repaso cronológico de la vida de la Esteban desde el momento en el que una completa desconocida comenzó a aparecer en los medios en 1998 como la novia de Jesulín de Ubrique hasta llegar a hoy en día, convertida en la princesa del pueblo, en concreto del popular madrileño distrito de San Blas donde vive, tal y como algunos la han calificado, y protagonista de portadas de revistas, diarios y portales web y de aparecer incluso entre los personajes más populares de Google. Junto a María Teresa Campos, estarán en el plató Patricia Pérez, presentadora del programa matinal de los sábados en Telecinco Vuélveme loca, quien ha conducido las campanadas en cuatro ocasiones, y los comentaristas Maribel Escalona, Emilio Pineda y José Manuel Parada. Los vuelos han venido registrando este viernes importantes retrasos en Barajas a pesar de que desde las 15.00 el aeropuerto opera con las cuatro pistas, según han informado fuentes de AENA, mientras las compañías han denunciado demoras por parte de los controladores de hasta 60 minutos con los pasajeros embarcados. Según los datos facilitados por AENA, la ausencia por la mañana de 5 de los 18 controladores que estaban programados en el turno de la torre de control de Barajas obligó a cerrar dos de las pistas del aeropuerto, lo que generó retrasos medios de 30 minutos.

Resumen:

El aeropuerto de Barajas operó con solo dos pistas desde las 15.00 debido a la ausencia de varios operarios.

Summarization: Desafíos de hoy en día



- Longitud del documento a procesar
 - Mientras más largo el documento, más difícil es recopilar toda la información relevante
- Topics especializados
 - Algunos textos requieren conocimiento experto para poder comprenderlos
 - Ejemplo: artículos científicos
- Evaluación
 - Es difícil evaluar si una pieza de información es cubierta por el resumen
 - Parafrasis
 - Evaluación con humanos / anotadores es imprecisa



Tareas NLP

Dialogo

Dialogo Multi-Turno

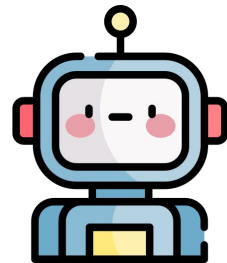
Crea un personaje retado a sobrevivir en una sociedad distópica ...

Wow, la historia de Ava es bien intensa e inspiradora! Me podrias dar mas detalles? ...

Nombre: Ava

Ava tenía solo 16 años cuando el mundo, como ella lo conocía, colapsó dejando atrás una sociedad caótica y...

Por supuesto! ...



Dialogo Multi-Turno



- Extensión natural de la tarea QA a más de una interacción
- Dada una pregunta o query **Q_t** y un historial de interacciones pasadas,
 - Retornar un texto **A_t** con la información requerida

$$Historial = [(Q_1, A_1), \dots (Q_{t-1}, A_{t-1})]$$

Dialogo: Tareas NLP



- **Next Utterance Prediction**
 - Predecir la respuesta correcta de entre varias opciones
- **Conversational QA**
 - Generar respuesta a una pregunta dado el historial y un documento
- **Reading Comprehension**
 - Clasificar la interacción completa en categorías, e.g. *“mesero-comensal”*
- **Coreference Resolution**
 - Para todo pronombre, resolver a qué individuo corresponde en interacciones pasadas.

Dialogo: Reading Comprehension

M

Ma'am, you forgot your phone.

F

Oh, thanks, I couldn't live without this little thing.

M

I know what you mean. It is of great significance to you. So did you enjoy your dinner?

F

Oh yes, everything was just perfect. It's so hard to take the whole family out to eat, but your restaurant was perfect. Johnny had his own place to play in and I had time to talk with my sisters and their husbands.

M

I'm glad to hear it. Our kid area is always popular.

F

Well, you can be sure we'll be back.

What is the probable relationship between the speakers?

✓ A. Waiter and Customer.
✗ B. Brother and Sister.
✗ C. Husband and Wife.

Question & Answer

Dialogo: Next Utterance Prediction

M

Ma'am, you forgot your phone.

F

Oh, thanks, I couldn't live without this little thing.

M

I know what you mean. It is of great significance to you. So did you enjoy your dinner?

F

Oh yes, everything was just perfect. It's so hard to take the whole family out to eat, but your restaurant was perfect. Johnny had his own place to play in and I had time to talk with my sisters and their husbands.

M

✓ A. Thanks for your compliment for the restaurant.

X B. I'm sorry that you don't have a good time.

X C. Goodbye brother! Love you.

X D. Hurry up honey, or we will be late for the dinner.

Response

Dialogo: Conversational QA



Jessica went to sit in her rocking chair. Today was her birthday and she was turning 80. Her granddaughter Annie was coming over in the afternoon and Jessica was very excited to see her. Her daughter Melanie and Melanie's husband Josh were coming as well. Jessica had . . .

Q₁: Who had a birthday?

A₁: Jessica

R₁: Jessica went to sit in her rocking chair. Today was her birthday and she was turning 80.

Q₂: How old would she be?

A₂: 80

R₂: she was turning 80

Q₃: Did she plan to have any visitors?

A₃: Yes

R₃: Her granddaughter Annie was coming over

Q₄: How many?

A₄: Three

R₄: Her granddaughter Annie was coming over in the afternoon and Jessica was very excited to see her. Her daughter Melanie and Melanie's husband Josh were coming as well.

Q₅: Who?

A₅: Annie, Melanie and Josh

R₅: Her granddaughter Annie was coming over in the afternoon and Jessica was very excited to see her. Her daughter Melanie and Melanie's husband Josh were coming as well.

Q: pregunta

A: respuesta

R: explicacion

Dialogo: Tareas NLP

dataset	Task	Reasoning	Domain	Manually
Ubuntu (Lowe et al., 2015)	Next Utterances Prediction	✗	Technique	✗
PERSONA-CHAT (Zhang et al., 2018a)	Next Utterances Prediction	✗	Persona	✓
Dialogue NLI (Welleck et al., 2019)	Next Utterances Prediction	✗	Persona	✗
CoQA (Reddy et al., 2019)	Conversational QA	✓	Diverse	✓
Douban (Wu et al., 2017)	Next Utterances Prediction	✗	Open	✗
DREAM (Sun et al., 2019)	Reading Comprehension	✓	Open	✓
WSC (Levesque et al., 2012)	Coreference Resolution	✓	Open	✗
SWAG (Zellers et al., 2018)	Plausible Inference	✓	Movie	✗
CommonsenseQA (Talmor et al., 2019)	Reading Comprehension	✓	Open	✓
RACE (Lai et al., 2017)	Reading Comprehension	✓	Open	✗
ARC (Clark et al., 2018)	Reading Comprehension	✓	Science	✗
DROP (Dua et al., 2019)	Reading Comprehension	✓	Open	✗
Cosmos (Huang et al., 2019)	Reading Comprehension	✓	Narrative	✓
MuTual	Next Utterances Prediction	✓	Open	✓



Tareas NLP

Traducción Automatica:
Machine Translation

Traducción Automática (Machine Translation)



- Dada una oración S en lenguaje A , retornar una oración T en lenguaje B , tal que
 - T sea una traducción fidedigna de S
- T es traducción de S si
 - T sea semantica y *pragmáticamente* equivalente a S
- Traducción \neq interpretacion?
 - Nombre propios
 - Frases coloquiales
 - Frases con profundo trasfondo cultural




Introduccion a HuggingFace y Google Colab


HuggingFace



- Repositorio de recursos para investigación y desarrollo de NN
 - Datasets
 - Modelos
 - Métricas de evaluación
- Librerías que centraliza el manejo de
 - Datasets: *datasets*
 - Modelos: *transformers*
- Leaderboard para evaluación amplia y estandarizada de LLMs

HuggingFace

 **Hugging Face**

[Models](#) [Datasets](#) [Spaces](#) [Posts](#) [Docs](#) [Enterprise](#) [Pricing](#) [⌵](#) 

To improve the supply chain security of the [huawei-noah](#) organization, we strongly recommend upgrading its subscription to Enterprise Hub, which will unlock SSO, access control, auditability, and much more.

[Upgrade huawei-noah](#)

[+ New](#)

ronaldahmed

- Profile
- Inbox (0)
- Settings
- Billing
- Get **Pro**

Organizations

- HUAWEI Noah's Ark Lab
- GEM benchmark
- Create New


Resources


- Hub guide
- Transformers doc


Following 4 [New Post](#)

All Models Datasets Spaces Papers Collections Community Posts

Upvotes Likes Articles

 **mlabonne** posted an update about 4 hours ago [Suggested for you](#)

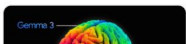
Post 412 

 **Gemini 3 Abliterated**

I noticed that Gemini 3 was much more resilient to refusal removal than other models like Qwen 2.5.

I experimented with different recipes and improved the ablation technique I wrote about last year.

[...read more](#)



Trending last 7 days


All Models Datasets Spaces

sesame/csm-1b
Text-to-Speech • Updated about 1 week ago • 2.84k • 1.02k

google/gemma-3-27b-it
Image-Text-to-Text • Updated 5 days ago • 241k • 708

Qwen/QwQ-32B
Text Generation • Updated about 1 week ago • 417k • 2.29k

deepseek-ai/DeepSeek-R1
Text Generation • Updated about 1 week ago • 1.9M • 11.4k

Sesame CSM  413
Conversational speech generation

FreedomIntelligence/medical-o1-reasoning

<https://huggingface.co/>

HuggingFace: LLM Leaderboard

Spaces | open-llm-leaderboard / open_llm_leaderboard | like 12.8k | Running on CPU UPGRADE | App | Files | Community 1135

Supports strict search and regex • Use semicolons for multiple terms

Quick Filters: For Edge Devices - 786 | **For Consumers - 430** | Mid-range - 3185 | For the GPU-rich - 165 | Only Official Providers - 470

table options | column visibility

	Rank	Type	Model	Average	IFEval	BBH	MATH	GPQA	MUSR
🔼	618	👤	PJMixers-Dev/Qwen2.5-RomboTiesTest-7B	35.29 %	75.58 %	34.93 %	49.62 %	6.38 %	8.72 %
🔼	886	👤	brgx53/3Bgeneralv2-ECE-PRYMMAL-Martial	31.48 %	56.77 %	37.25 %	34.97 %	8.17 %	12.79 %
🔼	1005	👤	DreadPoor/Here_We_Go_Again-8B-SLERP	30.13 %	74.42 %	35.53 %	17.30 %	9.17 %	12.43 %
🔼	1121	💬	microsoft/Phi-4-mini-instruct	29.41 %	73.78 %	38.74 %	16.99 %	7.94 %	6.45 %
🔼	1127	💬	T145/ZEUS-8B-V15	29.37 %	70.13 %	36.18 %	23.04 %	3.47 %	9.42 %
🔼	1131	👤	DreadPoor/Asymmetric_Linearity-8B-Model_Stock	29.35 %	71.74 %	35.44 %	16.47 %	8.61 %	12.23 %
🔼	1200	👤	DreadPoor/VENN_1.2-8B-Model_Stock	28.85 %	72.26 %	35.13 %	17.07 %	6.26 %	12.17 %

https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard#/?params=3%2C7

Colab



- Servicio de Google que provee un notebook (estilo jupyter) en browser
- Instancia corre en un servidor (CPU, GPU, TPU)
 - Instancia gratis con GPUs de 12GB
- Fácil de linkear a google drive para cargar o guardar data o modelos
- Plataforma de desarrollo de las tareas de curso

<https://colab.research.google.com/>

Preguntas?
