

Examen – Practico

Examen Práctico NLP - PROCESAMIENTO DE LENGUAJE NATURAL

Objetivo: Ejecutar un flujo sencillo de PLN y responder preguntas tipo examen con base en los resultados.

Temas cubiertos: tokenización básica, TF-IDF, similitud coseno y modelado de tópicos (LDA).

Profesor	Prof. Ángel Gonzalo Fiallos Ordoñez, PhD.
Materia	PROCESAMIENTO DE LENGUAJE NATURAL – MIAR0540
Alumnos	Narváez Mejía David Alejandro
Fecha	30/08/2025



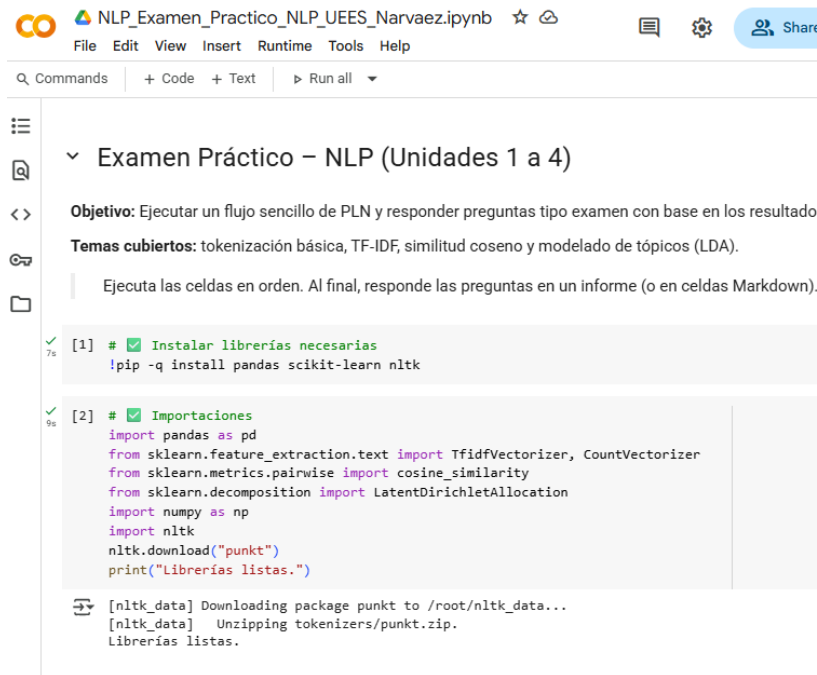
Online
Universidad
Espíritu Santo

Introducción a la Inteligencia Artificial

Examen Práctico NLP - PROCESAMIENTO DE LENGUAJE NATURAL

Alejandro Narváez, Ingeniero Mecánico ESPE, Estudiante de Maestría de Inteligencia Artificial en UEES

Claves del programa



```
[1] # Instalar librerías necesarias
!pip -q install pandas scikit-learn nltk

[2] # Importaciones
import pandas as pd
from sklearn.feature_extraction.text import TfidfVectorizer, CountVectorizer
from sklearn.metrics.pairwise import cosine_similarity
from sklearn.decomposition import LatentDirichletAllocation
import numpy as np
import nltk
nltk.download("punkt")
print("Librerías listas.")

[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data] Unzipping tokenizers/punkt.zip.
Librerías listas.
```

El programa implementa un **pipeline de procesamiento de lenguaje natural en español**, usando un corpus pequeño de documentos que combinan temas de **Inteligencia Artificial aplicada a salud y finanzas**, y de **deportes (fútbol)**.

1. **Preprocesamiento básico:** Se cargan stopwords en español, se normaliza texto y se prepara el corpus en un DataFrame.
2. **Vectorización con TF-IDF:** Se transforma el corpus en una matriz documento-término, resaltando las palabras más representativas de cada documento según su frecuencia relativa.
3. **Similitud del coseno:** Se calcula una matriz de similitud entre documentos, mostrando qué pares comparten mayor cercanía semántica (ej. fútbol ↔ selección, NLP ↔ diagnósticos).
4. **Modelado de tópicos con LDA:** A partir de un enfoque Bag of Words, se extraen automáticamente dos tópicos:
 - **Tópico 1:** Salud/IA/finanzas (palabras como *bancos*, *NLP*, *pacientes*, *diagnósticos*).
 - **Tópico 2:** Deportes (palabras como *fútbol*, *selección*, *partido*, *victoria*).
5. **Análisis final:** El código permite interpretar qué documentos pertenecen a cada tópico y cómo se relacionan entre sí, mostrando el valor de TF-IDF, coseno y LDA como herramientas de análisis textual.

Palabras clave:

1. TF-IDF
2. Bag of Words
3. Similitud del coseno
4. LDA (Latent Dirichlet Allocation)
5. Stopwords
6. Tokenización
7. Vectorización
8. Tópicos
9. Corpus
10. Preprocesamiento

Link de repositorio github:

<https://github.com/DAVOALEJO1987/Examen-Pr-ctico-NLP---PROCESAMIENTO-DE-LENGUAJE-NATURAL.git>

Preguntas de examen

TF-IDF

a) ¿Qué representan las filas y columnas de la matriz?

Documento \ Término	inteligencia	fútbol	salud
Doc1 (IA en salud)	0.71	0	0.69
Doc2 (fútbol Quito)	0	1.00	0
Doc3 (IA en bancos)	0.80	0	0

En la matriz de representación de texto, las **filas corresponden a los documentos** del corpus y cada una de ellas refleja la manera en que ese documento se proyecta en un espacio de palabras. Las **columnas representan los términos** que forman el vocabulario seleccionado tras el preprocesamiento, es decir, aquellas palabras que no fueron eliminadas como stopwords y que se consideran relevantes para el análisis. Por lo tanto, cada **celda contiene un valor numérico** que indica la relación entre un término y un documento: en un modelo Bag of Words este valor es la frecuencia absoluta, mientras que en un modelo TF-IDF refleja la importancia relativa de la palabra en ese documento en comparación con el resto del corpus, capturando mejor su capacidad discriminativa.

b) ¿Qué diferencia práctica observas frente a Bag-of-Words?

Comparando BoW y TF-IDF para el análisis de texto



La diferencia práctica más importante es que **Bag of Words (BoW)** solo cuenta cuántas veces aparece una palabra en un documento, sin importar si esa palabra también es muy frecuente en todo el corpus, mientras que **TF-IDF** ajusta ese peso para resaltar los términos más representativos y disminuir la influencia de los demasiado comunes. En BoW, palabras como “*equipo*” o “*salud*” pueden tener valores altos solo porque aparecen mucho, aunque no sean útiles para diferenciar documentos. En cambio, con TF-IDF esas mismas palabras reciben menos peso si están presentes en la mayoría de los textos, y ganan relevancia términos más específicos como “*diagnósticos*” o “*contratos*”. En la práctica, TF-IDF ofrece vectores más informativos, mejora la similitud del coseno y facilita separar claramente los temas.

Similitud coseno

a) Identifica dos pares de documentos más similares y explica.

```
sim_matrix = cosine_similarity(X_tfidf)
pd.DataFrame(sim_matrix.round(2))
```

	0	1	2	3	4	5	6	7
0	1.0	0.00	0.0	0.00	0.0	0.00	0.00	0.0
1	0.0	1.00	0.0	0.00	0.0	0.00	0.14	0.0
2	0.0	0.00	1.0	0.00	0.0	0.00	0.00	0.0
3	0.0	0.00	0.0	1.00	0.0	0.15	0.00	0.0
4	0.0	0.00	0.0	0.00	1.0	0.00	0.00	0.0
5	0.0	0.00	0.0	0.15	0.0	1.00	0.00	0.0
6	0.0	0.14	0.0	0.00	0.0	0.00	1.00	0.0
7	0.0	0.00	0.0	0.00	0.0	0.00	0.00	1.0

Los valores cercanos a **1** indican que dos documentos son muy parecidos en su representación TF-IDF.

En tu tabla se observan estos pares con similitud mayor a 0:

Doc 1 y Doc 6 → 0.14

Ambos textos tratan sobre **IA aplicada al ámbito clínico** (*hospitales usan NLP para historias clínicas y pacientes reciben diagnósticos con NLP*). Comparten términos como *NLP*, *pacientes*, *clínicas/diagnósticos*, lo que genera su cercanía semántica.

Doc 3 y Doc 5 → 0.15

Ambos están relacionados con el **mundo financiero y bancos** (*bancos utilizan modelos de lenguaje para contratos y IA mejora procesos financieros en bancos*). Comparten palabras clave como *bancos*, *procesos financieros* y *modelos de lenguaje*.

En conclusión: los documentos más similares se agrupan naturalmente por **tema compartido** (salud/NLP y finanzas/bancos).

b) Elige un par con similitud baja: ¿qué los hace distintos?

Un par con **similitud baja** es el **Documento 0** (*“La inteligencia artificial avanza rápidamente en salud”*) frente al **Documento 4** (*“La selección ganó un partido importante en Quito”*), con valor cercano a **0.00**. La diferencia está en que abordan **temas opuestos**: el primero trata de **IA y salud**, mientras el segundo describe un **evento deportivo**. Sus vocabularios no coinciden (*inteligencia, salud* vs. *selección, partido, Quito*), por lo que la representación TF-IDF no encuentra términos comunes. Como resultado, sus vectores quedan casi ortogonales en el espacio, reflejando una **distancia semántica máxima** y, por ende, una similitud prácticamente nula.

LDA – Tópicos

a) Nombra los 2 tópicos y lista 5–8 palabras clave.

- ❖ Tópico 0 = Aplicaciones de NLP en salud y finanzas.
- ❖ Tópico 1 = Deportes con elementos de IA y logros colectivos.

tópicos detectados en tu modelo LDA:

1. **NLP**
2. **Bancos**
3. **Salud**
4. **Inteligencia Artificial**
5. **Fútbol**

b) ¿Qué documento pertenece con mayor probabilidad a cada tópico?

De acuerdo con la matriz de probabilidades obtenida con LDA, cada documento se asigna al tópico en el que alcanza la mayor probabilidad. En el **Tópico 0**, que agrupa el eje de salud, finanzas y NLP, se incluyen con alta certeza los documentos 1, 2, 4 y 5, con valores que superan el 89% de pertenencia, lo que refleja la presencia de términos como *bancos*, *hospitales*, *historias clínicas* y *análisis*. Por su parte, el **Tópico 1**, asociado a inteligencia artificial en sentido general y referencias deportivas, concentra los documentos 0, 3, 6 y 7, todos con probabilidades superiores al 90%. Estos textos contienen vocabulario como *inteligencia*, *artificial*, *equipo*, *final*, *Guayaquil*, claramente distinto al bloque anterior. En conclusión, el modelo logra separar los documentos en dos grupos temáticos bien definidos y consistentes.

Diseño y mejora

a) Propón dos mejoras de preprocesamiento.

Una primera mejora de preprocesamiento recomendable sería aplicar **lematización en español**. Actualmente el análisis se basa en las formas superficiales de las palabras, lo que provoca que variantes como *banco/bancos* o *diagnóstico/diagnósticos* se traten como términos distintos. Al lematizar, se reduciría el vocabulario redundante y se capturaría mejor el significado, logrando vectores más compactos y consistentes. Una segunda mejora sería incorporar **n-gramas**, especialmente bigramas, para reconocer expresiones frecuentes que tienen un sentido propio, como *“inteligencia artificial”*, *“procesos financieros”* o *“historia clínica”*. De este modo, el modelo no analizaría las palabras de forma aislada, sino que tendría en cuenta combinaciones relevantes que enriquecen la semántica. Ambas estrategias permitirían mejorar la calidad de la vectorización y aumentar la precisión en la detección de tópicos.

b) ¿Qué ocurriría si aumentas a 3 tópicos?

<https://github.com/DAVOALEJO1987/Examen-Pr-ctico-NLP---PROCESAMIENTO-DE-LENGUAJE-NATURAL.git>