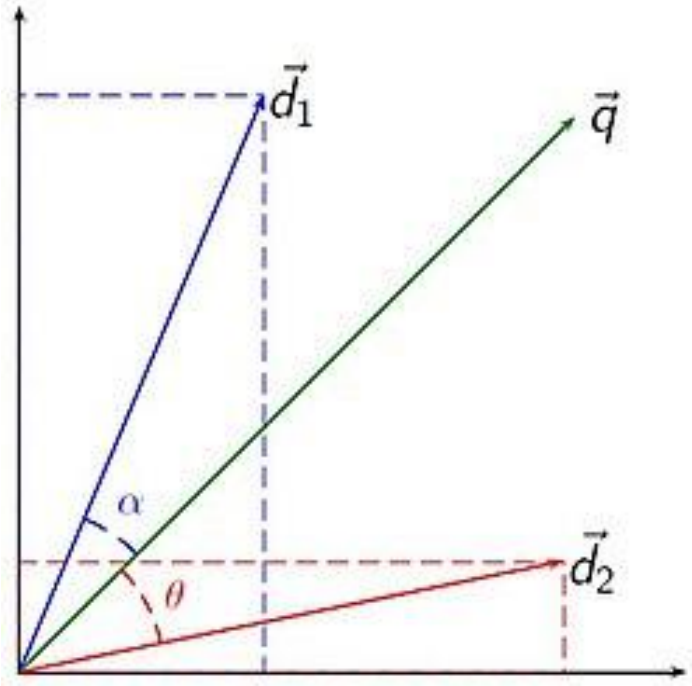
INSTITUTO TECNOLÓGICO BELTRÁN
Centro de Tecnología e Innovación

# MODELO ESPACIO VECTORIAL

*Trabajo practico N°9*




**Alumno** = Nicolas Mesquiatti

**Materia** = Técnicas de procesamiento del habla

## Que es un modelo de espacio vectorial

*Un **modelo espacio vectorial** representa documentos como vectores de números, para poder medir la similitud entre ellos, por ejemplo, con la distancia o el ángulo entre vectores*

***EJERCICIO***

*Crea un programa en Python que busque las similitudes en un conjunto de documentos con modelo de*
*espacio vectorial utilizando sklearn.*

🔻 Pasos a seguir:

-Se tienen 3 documentos con información sobre animales.

"doc1": "El veloz zorro marrón salta sobre el perro perezoso.",

"doc2": "Un perro marrón persiguió al zorro.",

"doc3": "El perro es perezoso.",

**-Convertir documentos a vectores usando TF-IDF.**

**-Calcular la similitud del coseno entre los documentos.**

**-Graficar la matriz de similitud.**

### 1. Descripción de los Documentos Analizados

Los textos utilizados fueron:

- **doc1:** "El veloz zorro marrón salta sobre el perro perezoso."
- **doc2:** "Un perro marrón persiguió al zorro."
- **doc3:** "El perro es perezoso."

Se siguieron los siguientes pasos:

1. **Limpieza de texto:** Se eliminaron palabras vacías y se lematizo a su palabra raíz (stopwords, lematización) en español usando `nltk`.
2. **Vectorización TF-IDF:** Cada documento se convirtió en un vector numérico que representa la relevancia de sus palabras.
3. **Cálculo de similitud del coseno:** Se midió el ángulo entre los vectores para determinar la similitud entre cada par de documentos.
4. **Visualización:** Se construyó un mapa de calor con la matriz de similitud.

**-Convertir documentos a vectores usando TF-IDF.**

```
# Vectorización con TF-IDF y stopwords en español
stopwords_es = stopwords.words('spanish')
vectorizador = TfidfVectorizer(stop_words=stopwords_es)
tfidf_matrix = vectorizador.fit_transform(textos)
```

**-Calcular la similitud del coseno entre los documentos.**

```
# Similaridad coseno
similitud = cosine_similarity(tfidf_matrix)
```

Compara cada documento con los otros y da un número entre 0 y 1:

- **1** = muy parecidos
- **0** = nada parecidos

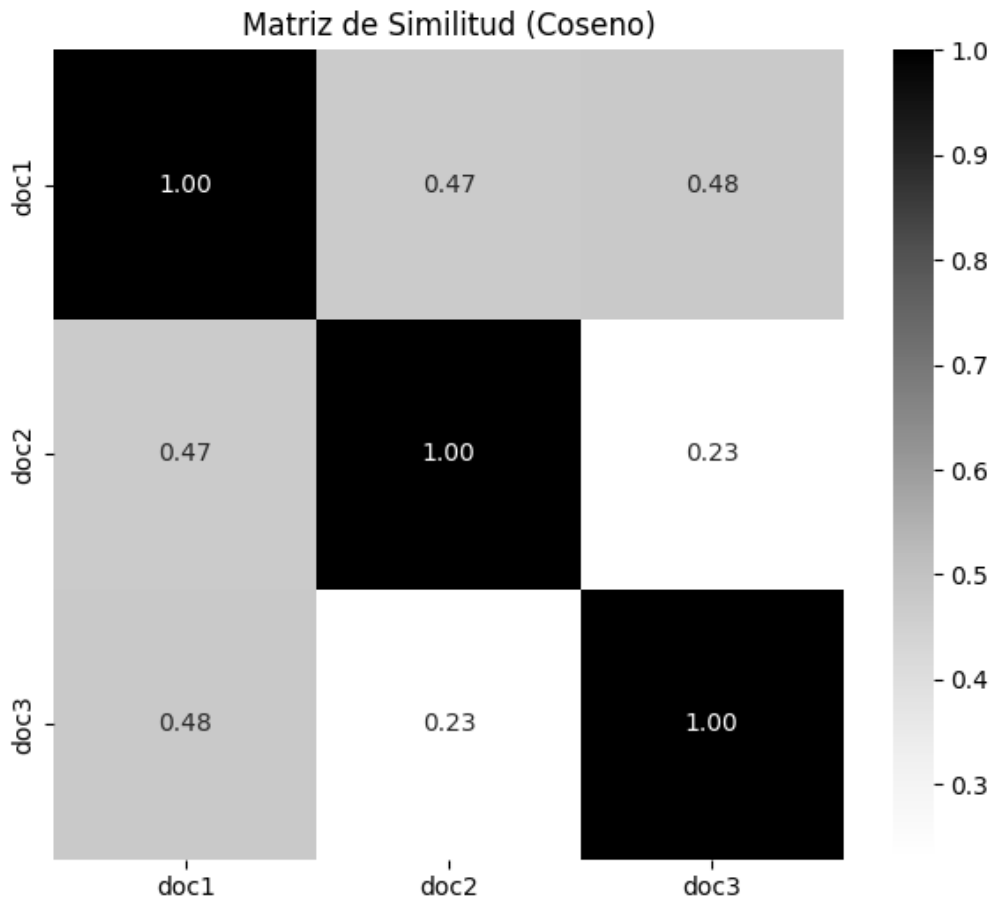
**-Graficar la matriz de similitud.**

```python
# Mostrar resultados
print("📊 Matriz de similitud:")
print(np.round(similitud, 2))

# Graficar
# Graficar
plt.figure(figsize=(6, 5))

sns.heatmap(similitud, annot=True,
            xticklabels=nombres_docs,
            yticklabels=nombres_docs,
            cmap="Greys",  # <- cambio hecho acá
            fmt=".2f")

plt.title("Matriz de Similitud (Coseno)")
plt.show()
```



## *Conclusión*

### *Tonos oscuros: mayor similitud*
### *Tonos claros: menor similitud*

**doc1 y doc2** presentan una similitud del 47**%**, indicando que comparten varios términos significativos (como "zorro", "perro", "marrón").

**doc1 y doc3** tienen una similitud de 48%, ya que el tercero es más corto y, pero tiene más palabras en común como (“El”, “perro”,” perezoso”)

**doc2 y doc3** tienen la menor similitud (23%), lo cual es esperable dado que uno habla de una acción ("persiguió") y el otro de un estado ("es perezoso").