

# Python

## Procesamiento de Lenguaje Natural

Nubes de Palabras



Rogelio Ferreira Escutia

Profesor / Investigador  
Tecnológico Nacional de México  
Campus Morelia



# Nubes de Palabras



# Nube de Palabras (1)

- Se hará la nube de palabras a partir del texto de una página Web (Wikipedia: Ciencia de Datos):



The screenshot shows the Wikipedia page for 'Ciencia de datos' in Spanish. The browser address bar displays 'es.wikipedia.org/wiki/Ciencia\_de\_datos'. The page features the Wikipedia logo on the left and navigation tabs for 'Artículo' and 'Discusión'. The main content area contains the title 'Ciencia de datos' and several paragraphs of text. The first paragraph defines data science as an interdisciplinary field. The second paragraph provides a definition of data science. The third paragraph mentions Jim Gray and the Turing Award. The fourth paragraph discusses the new paradigm of data science.

← → ↻ es.wikipedia.org/wiki/Ciencia\_de\_datos 🔍 📄 ☆ 🏠 👤

No has accedido [Discusión](#) [Contribuciones](#) [Crear una cuenta](#) [Accede](#)

Artículo **Discusión** Leer Editar Ver historial

## Ciencia de datos

La **ciencia de datos** es un campo interdisciplinario que involucra métodos científicos, procesos y sistemas para extraer conocimiento o un mejor entendimiento de datos en sus diferentes formas, ya sea estructurados o no estructurados,<sup>1</sup> lo cual es una continuación de algunos campos de análisis de datos como la [estadística](#), la [minería de datos](#), el [aprendizaje automático](#), y la [analítica predictiva](#).<sup>1</sup>

También se define La ciencia de datos como "un concepto para unificar estadísticas, análisis de datos, aprendizaje automático, y sus métodos relacionados, a efectos de comprender y analizar los fenómenos reales",<sup>2</sup> empleando técnicas y teorías extraídas de muchos campos dentro del contexto de las matemáticas, la estadística, la ciencia de la información, y la informática.

El ganador del [premio Turing](#), [Jim Gray](#), imaginó la ciencia de datos como un "cuarto paradigma" de la ciencia (empírico, teórico, computacional, y ahora basado en datos), y afirmó que "todo lo relacionado con la ciencia está cambiando debido al impacto de la tecnología de la información y el diluvio de datos".<sup>3</sup>

En este nuevo paradigma, los investigadores se apoyan de sistemas y procesos que son muy diferentes a los utilizados en el pasado, como son modelos, ecuaciones, algoritmos, así como evaluación e interpretación de resultados.<sup>1</sup>

# Nube de Palabras (2)

- **Bibliotecas:**

```
# Bibliotecas a utilizar
import wikipedia
import re
import nltk
from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords
from wordcloud import WordCloud, STOPWORDS
import matplotlib.pyplot as plt

# Extraer y procesar de Wikipedia
# Manejo de expresiones regulares
# Para procesar lenguaje natural
# Para Tokenizar un texto
# Cargar las "Stopwords" del español
# Crear nubes de ideas
# Graficar
```

# Nube de Palabras (3)

- Se seleccionará Wikipedia en Español, se cargará la página seleccionada y se imprimirá en pantalla

```
# Seleccionar Wikipedia en Español
wikipedia.set_lang("es")

# Extraer información de una página de Wikipedia
texto = wikipedia.summary("Ciencia de Datos")

# Impresión completa del texto
print("\nTexto Completo:\n\n", texto)
```

# Nube de Palabras (4)

- Eliminamos símbolos y caracteres especiales y “tokenizamos” el texto:

```
# Eliminar símbolos y caracteres especiales usando expresiones regulares
texto_sin_simbolos = re.sub(r'[^\w\s]', '', texto)

# Convertimos a tokens todo el texto y lo imprimimos en pantalla
tokens_de_mi_texto = word_tokenize(texto_sin_simbolos)
print('\nImpresión de todos los tokens del texto:\n\n', tokens_de_mi_texto)
print('\n Tokens Totales: ', len(tokens_de_mi_texto))
```

# Nube de Palabras (5)

- Cargamos las “StopWords” y hacemos el filtrado:

```
# Cargamos las "stopwords" del español (las palabras que no nos aportan información)
palabras_vacias = set(stopwords.words('spanish'))

# Filtramos los tokens eliminando las "stopwords"
lista_final = []
for palabra in tokens_de_mi_texto:
    if palabra not in palabras_vacias:
        lista_final.append(palabra)
```

- Imprimimos los Tokens:

```
# Impresión final de las palabras relevantes del texto (Tokenizar)
print('\nLista Final eliminando las palabras vacías (no relevantes):\n\n', lista_final)
print("\nTotal de Tokens sin Stopwords: ", len(lista_final))
```

# Nube de Palabras (6)

- Convertimos toda nuestra lista de “tokens” en un solo texto, ya que para generar la nube (con la biblioteca “WordCloud”) se le tienen que dar los datos en este formato y lo imprimimos en pantalla:

```
# Convertir la lista de tokens a un solo texto separados por un espacio
texto_final = " ".join(lista_final)

# Imprimir el texto final
print("texto final:\n", texto_final)
```

# Nube de Palabras (7)

- De acuerdo a los parámetros que nos pide la biblioteca “WordCloud”, configuramos la pantalla de salida de nuestra “Nube de Palabras” con los siguientes datos:

```
# Formato de la "Nube de Ideas"
nube_de_ideas = WordCloud(
    width = 500,
    height = 500,
    random_state = 1,
    background_color = "salmon",
    colormap= "Pastel1",
    collocations = False,
    stopwords = STOPWORDS,
).generate(texto_final)
```





## Rogelio Ferreira Escutia

Profesor / Investigador  
Tecnológico Nacional de México  
Campus Morelia



[rogelio.fe@morelia.tecnm.mx](mailto:rogelio.fe@morelia.tecnm.mx)



[rogeplus@gmail.com](mailto:rogeplus@gmail.com)



[xumarhu.net](http://xumarhu.net)



[@rogeplus](https://twitter.com/rogeplus)



[https://www.youtube.com/  
channel/UC0on88n3LwTKxJb8T09sGjg](https://www.youtube.com/channel/UC0on88n3LwTKxJb8T09sGjg)



[rogelioferreiraescutia](https://www.linkedin.com/in/rogelioferreiraescutia)

