

# Trabajo Práctico 1

## Aprendizaje Automático Avanzado

Cisnero Matias, Seivane Nicolás, Serafini Franco  
22 de Septiembre de 2025

## Ejercicio 1: Creación de Corpus



## 1.1 Descripción de Librerías Usadas

Se utilizaron las librerías de *r* y *pdfplumber*, en la cual utilizamos la ultima para leer página por página de un pdf y la primera para seleccionar las palabras.

Los links a las librerías son los siguientes.

**pdfplumber**

**r**

### Funciones Utilizadas

- `pdfplumber.open()` as  
pdf
- `pdf.pages[]`
- `.extract_text()`
- `.split('\n')`
- `re.findall()`
- `.endswith()`
- `.strip()`
- `.isdigit()`
- `.split('\n')`

## 1.2.1 Estructura de Código

### Se utiliza la siguiente estructura de código:

Se comienza importando las librerías y creando una lista de palabras, donde se irán agregando las extracciones de texto.

```
import pdfplumber
import re

words = []
```

En lo cual se sigue utilizando la función `pdfplumber.open()` as `pdf`, en la cual se debe especificar la ruta hacia el pdf. El cual nos devuelve *pdf* como una instancia de la clase `pdfplumber.PDF`

```
with pdfplumber.open("ruta") as pdf:
```

## 1.2.2 Estructura de Código

Se continua utilizando una propiedad de la clase `pdfplumber.Page`, de la cual se puede indexar para acceder a las paginas del pdf

```
with pdfplumber.open("ruta") as pdf:
    for page in pdf.pages[:]:
```

En lo cual se utiliza el metodo `.extract_text()`, que recopila todos los objetos de caracteres de la página en un solo string.

```
with pdfplumber.open("ruta") as pdf:
    for page in pdf.pages[:]:
        text = page.extract_text()
        if text:
```

## 1.2.3 Estructura de Código

Se continua diviendo el string segun el metodo `.split('\n')`, el cual devuelve una lista de strings, los cuales fueron separados de acuerdo a `\n`, ergo saltos de linea.

```
with pdfplumber.open("ruta") as pdf:
    for page in pdf.pages[:]:
        text = page.extract_text()
        if text:
            lines = text.split('\n')
```

Luego se sacan las lineas que sean numeros de pagina tanto en el pie de la misma como en el encabezado. La forma de extraccion varia de acuerdo a como es el pdf.

```
if lines[-1].strip().isdigit():
    lines = lines[:-1]
if lines[0].strip().isdigit():
    lines = lines[1:]
```

## 1.2.4 Estructura de Código

Se crea por linea una lista con el método de la librería `re`:

`re.findall(r"_w+|[,!?:;]", line)`, en el cual se separan con expresiones regulares las palabras con `\w+` y aparte los signos de puntuación con `[,!?:;]`, en una lista de strings. Luego para cada palabra se la pasa a minúscula con el método `.lower()`.

```
with pdfplumber.open("ruta") as pdf:
    for page in pdf.pages[:]:
        text = page.extract_text()
        if text:
            lines = text.split('\n')
            if lines[-1].strip().isdigit():
                lines = lines[:-1]
            if lines[0].strip().isdigit():
                lines = lines[1:]
            for line in lines:
                tokens = re.findall(r"\w+|[,!?:;]", line)
                tokens = [token.lower() for token in tokens]
```

## 1.2.5 Estructura de Código

Luego se diferencia por línea los puntos aparte, los cuales consideramos los últimos puntos de las líneas. Cada línea, las cuales fueron convertidas en listas de strings son agregadas a la lista del corpus

```
with pdfplumber.open("ruta") as pdf:
    for page in pdf.pages[:]:
        text = page.extract_text()
        if text:
            lines = text.split('\n')
            if lines[-1].strip().isdigit():
                lines = lines[:-1]
            if lines[0].strip().isdigit():
                lines = lines[1:]
            for line in lines:
                tokens = re.findall(r"\_w+| [.,!?:;]", line)
                tokens = [token.lower() for token in tokens]
            if line.endswith("."):
                tokens[-1] = ". "
            words.extend(tokens)
```



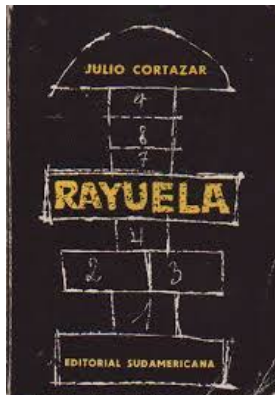
## 1.3 Libros utilizados: Rayuela

**Título:** Rayuela

**Autor:** Julio Cortazar

**Año :** 1963

Se extrayeron 197.342 caracteres y 20.810 caracteres únicos que conforman el vocabulario.



## 1.3.2 Código utilizado: Rayuela

```
with pdfplumber.open("Julio-Cortazar-Rayuela.pdf") as pdf:
    for page in pdf.pages[7:]:
        text = page.extract_text()
        if text:
            lines = text.split('\n')
            if lines[-1].strip().isdigit():
                lines = lines[:-1]
            if lines[0].strip().isdigit():
                lines = lines[1:]
            if lines[0].strip().isdigit():
                lines = lines[1:]
            if lines[-1].strip().isdigit():
                lines = lines[:-1]
            for line in lines:
                tokens = re.findall(r"\w+|[\.,!?:;]", line)
                tokens = [token.lower() for token in tokens]
            if line.endswith("."):
                tokens[-1] = "."
            words.extend(tokens)
```

## 1.3.3 Ejemplo Borrado: Rayuela

ganar de reir, el miedo me hacía una doble llave en la boca del estómago y al final me dio una verdadera desesperación (el mozo se había levantado furioso) y empecé a agarrar los zapatos de las mujeres y a mirar al debajo del arco de la suela no estaría agasapado el azúcar, y las gallinas cacareaban, los gallos gerentes me picoteaban el lomo, oía las carcajadas de Ronald y de Kienne mientras me movía de una mesa a otra hasta encontrar el azúcar escondido detrás de una pata Segundo Imperio. Y todo el mundo enfurecido, hasta yo con el azúcar apretado en la palma de la mano y sintiendo cómo se mezclaba con el sudor de la piel, cómo asquerosamente se deshacía en una especie de venganza pegajosa, esa clase de episodios todos los días.

4º Caso de borrado (-2)

1º Caso de borrado 9

2º Caso de borrado 2

3º Caso de borrado

Aquí había sido primero como una sangría, un vapulco de uso interno, una necesidad de sentir el estúpido pasaporte de tapas azules en el bolsillo del saco, la llave del hotel bien segura en el clavo del tablero. El miedo, la ignorancia, el deslumbramiento: Esto se llama así, eso se pide así, ahora esa mujer va a sonreír, más allá de esa calma empieza el Jardín des Plantes. París, una tarjeta postal con un dibujo de Kien al lado de un espejo vacío.

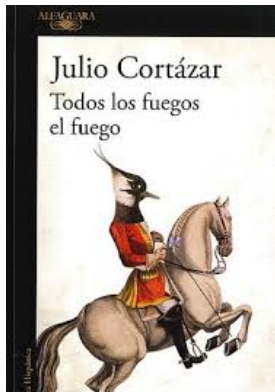
## 1.3 Libros utilizados: Todos los fuegos

**Título:** Todos los fuegos el fuego.

**Autor:** Julio Cortázar

**Año :** 1966

Se extrayeron 55.948 caracteres y 2.828 caracteres únicos que conforman el vocabulario.



## 1.3.2 Código utilizado: Todos los fuegos

```
with pdfplumber.open("Julio Cortazar Todos los fuegos.pdf")
    as pdf:
    for page in pdf.pages[:-1]:
        text = page.extract_text()
        if text:
            lines = text.split('\n')
            if lines[-1].strip().isdigit():
                lines = lines[:-1]
            for line in lines:
                tokens = re.findall(r"\w+|[\.,!?:;]", line)
                tokens = [token.lower() for token in tokens]
            if line.endswith("."):
                tokens[-1] = "."
            words.extend(tokens)
```

## 1.3.3 Ejemplo Borrado: Todos los fuegos

A la cuarta vez de encontrarse con todo eso, de hacer todo eso, el ingeniero había decidido no salir más de su coche, a la espera de que la policía disolviese de alguna manera el embotellamiento. El calor de agosto se sumaba a ese tiempo a ras de neumáticos para que la inmovilidad fuese cada vez más enervante. Todo era olor a gasolina, gritos destemplados de los jovencitos del Simca, brillo del sol rebotando en los cristales y en los bordes cromados, y para colmo la sensación contradictoria del encierro en plena selva de máquinas pensadas para correr. El 404 del ingeniero ocupaba el segundo lugar de la pista de la derecha contando desde la franja divisoria de las dos pistas, con lo cual tenía otros cuatro autos a su derecha y siete a su izquierda, aunque de hecho sólo pudiera ver distintamente los ocho coches que lo rodeaban y sus ocupantes que ya había detallado hasta cansarse. Había charlado con todos, salvo con los muchachos del Simca que le caían antipáticos; entre trecho y trecho se había discutido la situación en sus menores detalles, y la impresión general era que hasta Corbeil-Essones se avanzaría al paso o poco menos, pero que entre Corbeil y Juvisy el ritmo iría acelerándose una vez que los helicópteros y los motociclistas logaran quebrar lo peor del embotellamiento. A nadie le cabía duda de que algún accidente muy grave debía haberse producido en la zona, única explicación de una lentitud tan increíble. Y con eso el gobierno, el calor, los impuestos, la vialidad, un tópico tras otro, tres metros, otro lugar común, cinco metros, una frase sentenciosa o una maldición contenida.

A las dos monjitas del 2HP les hubiera convenido tanto llegar a Milly-la-Forêt antes de las ocho, pues llevaban una cesta de hortalizas para la cocinera. Al matrimonio del Peugeot 203 le importaba sobre todo no perder los juegos televisados de las nueve y media; la muchacha del Dauphine le había dicho al ingeniero que le daba lo mismo llegar más tarde a París pero que se quejaba por principio, porque le parecía un atropello someter a millares de personas a

Único Caso Borrado 3

un régimen de caravana de camellos. En esas últimas horas (debían ser casi las cinco pero el calor los hostigaba insoportablemente) habían avanzado unos cincuenta metros a juicio del ingeniero, aunque uno de los hombres del Taunus que se había acercado a charlar llevando de la mano al niño con su autito, mostró irónicamente la copa de un plátano solitario y la

## 1.3 Libros utilizados: Historias de cronopios y de famas

**Título:** Historias de cronopios y de famas.

**Autor:** Julio Cortázar

**Año :** 1962

Se extrayeron 32.224 caracteres y 2.514 caracteres únicos que conforman el vocabulario.



## 1.3.2 Código utilizado: Historias de cronopios y de famas

En este caso no fue necesario quitar ninguna línea.

```
with pdfplumber.open("Historias-de-Cronopios-y-de-Famas -
    Julio Cortazar.pdf") as pdf:
    for page in pdf.pages[3:-1]:
        text = page.extract_text()
        if text:
            lines = text.split('\n')
            for line in lines:
                tokens = re.findall(r"\w+|[\.,!?:;]", line)
                tokens = [token.lower() for token in tokens]
            if line.endswith("."):
                tokens[-1] = ". "
            words.extend(tokens)
```



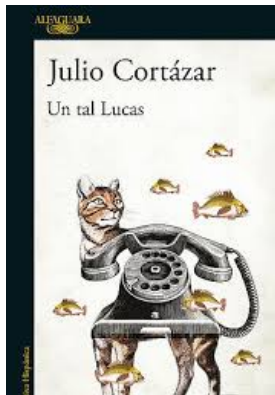
## 1.3 Libros utilizados: Un tal Lucas.

**Título:** Un tal Lucas.

**Autor:** Julio Cortazar

**Año :** 1979

Se extrayeron 32.224 caracteres y 2.514 caracteres únicos que conforman el vocabulario.



## 1.3.2 Código utilizado: Un tal Lucas.

```
with pdfplumber.open("Lucas_Julio_Cortazar.pdf") as pdf:
    for page in pdf.pages[5:]:
        text = page.extract_text()
        if text:
            lines = text.split('\n')
            if lines[-1].strip().isdigit():
                lines = lines[:-1]
            lines = lines[1:]
            for line in lines:
                tokens = re.findall(r"\w+|[\.,!?:;]", line)
                tokens = [token.lower() for token in tokens]
            if line.endswith("."):
                tokens[-1] = "."
            words.extend(tokens)
```

## 1.3.3 Ejemplo Borrado: Un tal Lucas.

2º Caso de Borrado **Un Tal Lucas – Julio Cortázar**

*Lucas, su patriotismo*

De mi pasaporte me gustan las páginas de las renovaciones y los sellos de visados redondos / triangulares / verdes / cuadrados / negros / ovalados / rojos, de mi imagen de Buenos Aires el transbordador sobre el Riachuelo, la plaza Irlanda, los jardines de Agronomía, algunos cafés que acaso ya no están, una cama en un departamento de Maipú casi esquina Córdoba, el olor y el silencio del puerto a medianoche en verano, los árboles de la plaza Lavalle.

Del país me queda un olor de acequias mendocinas, los álamos de Uspallata, el violeta profundo del cerro de Velasco en La Rioja, las estrellas chaqueñas en Pampa de Guanacos yendo de Salta a Misiones en un tren del año cuarenta y dos, un caballo que monté en Saladillo, el sabor del Cinzano con ginebra Gordon en el Boston de Florida, el olor ligeramente alérgico de las plateas del Colón, el superpúlman del Luna Park con Carlos Beulchi y Mario Díaz, algunas lecherías de la madrugada, la fealdad de la Plaza Once, la lectura de *Sur* en los años dulcemente ingenuos, las ediciones a cincuenta centavos de *Claridad*, con Roberto Arlt y Castelnuovo, y también algunos patios, claro, y sombras que me callo, y muertos.

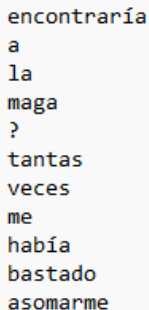
1º Caso de Borrado **10**

## 1.4 Corpus final

Se guarda el corpus final en un archivo llamado **corpus.txt**.

```
with open("corpus.txt", "w", encoding="utf-8") as f:  
    f.write("\n".join(words))
```

Obteniendo un corpus como se ve en la siguiente imagen.

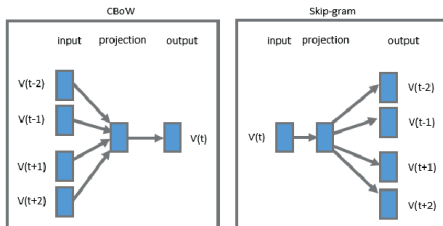


```
encontraría  
a  
la  
maga  
?  
tantas  
veces  
me  
había  
bastado  
asomarme
```

**Vocabulario (único):** 27.971

**Corpus total:** 310.347

## Ejercicio 2: Implementación CBOW y SkipGram



## 2.1 Pasos previos

Se abre y carga el **corpus.txt**.

```
with open("corpus.txt", "r", encoding="utf-8") as f:  
    corpus = f.read().splitlines()
```

Luego se crean los diccionarios que se utilizaran en ambos métodos.

```
vocab = sorted(set(corpus))  
vocab_tamano = len(vocab)  
palabra_a_indice = {palabra: i for i, palabra in  
enumerate(vocab)}  
indice_a_palabra = {i: palabra for i, palabra in  
enumerate(vocab)}
```

## 2.2 CBOW

### Definición: Continuous Bag of Words(CBOW)

**Propósito:** Es un modelo de aprendizaje automático para aprender representaciones de palabras que capturan el "significado" de las palabras basadas en su contexto.

**Principio:** A diferencia de los modelos más simples, CBOW utiliza un contexto de  $C$  palabras para predecir una palabra central

**Contexto vs. Predicción:** A partir de un contexto de  $C$  palabras ( $p_{I,1}, p_{I,2}, \dots, p_{I,C}$ ), se intenta predecir la palabra objetivo ( $p_O$ ), que generalmente es la palabra central

## X.X Skip-gram

### Definición: Skip-gram

**Propósito:** Es un modelo de aprendizaje automático para aprender representaciones de palabras que capturan el significado de las palabras basadas en su contexto. La diferencia con CBOW es que la entrada de la red es una palabra y la salida intenta predecir su contexto.

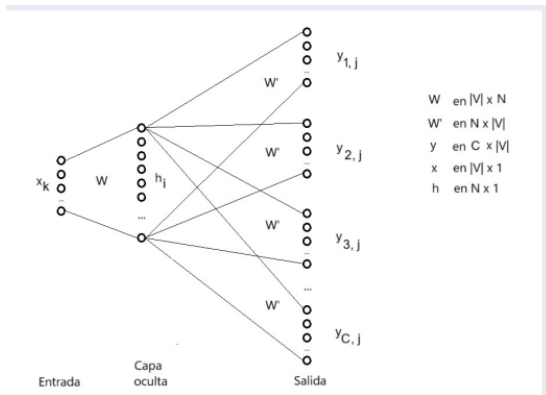
**Principio:** Skip-gram utiliza una palabra central para predecir un contexto de  $C$  palabras.

**Contexto vs. Predicción:** A partir de una palabra central ( $p_O$ ), el modelo intenta predecir las palabras de su contexto ( $p_{I,1}, p_{I,2}, \dots, p_{I,C}$ ).



## X.X Skip-gram

## Arquitectura de Skip-gram



## X.X Skip-gram: Representación de la entrada

### Entrada del modelo

**Palabra central:** se representa como un vector one-hot  $x \in \mathbb{R}^{|V|}$ , donde  $|V|$  es el tamaño del vocabulario.

Ejemplo: si  $|V| = 10\,000$  y la palabra central ocupa la posición 125, entonces  $x$  tiene un 1 en la posición 125 y 0 en las demás.

**Matriz de embeddings de entrada:**

$$W \in \mathbb{R}^{|V| \times N}$$

donde  $N$  es la dimensión del espacio de embeddings (ej.  $N = 300$ ).

## X.X Skip-gram: Representación vectorial

### Cálculo de la representación vectorial de la palabra $p_I$

La capa oculta no tiene función de activación. Se obtiene directamente el embedding de la palabra central:

$$h = W^T x = v_{p_I}, \quad h \in \mathbb{R}^N$$

donde  $v_{p_I}$  es el vector de embeddings asociado a la palabra central  $p_I$ .

Dimensiones:

- $x \in \mathbb{R}^{|V|}$  (one-hot).
- $W \in \mathbb{R}^{|V| \times N}$ .
- $h \in \mathbb{R}^N$  (embedding resultante).

## X.X Skip-gram: Salida y softmax

### Cálculo de probabilidades

Para cada palabra candidata  $j$  en el vocabulario, se calcula:

$$u_j = (v'_j)^T h, \quad u \in \mathbb{R}^{|V|}$$

donde  $v'_j$  es el vector de salida correspondiente a la palabra  $j$ , y  $W' \in \mathbb{R}^{N \times |V|}$  es la matriz de salida.

Luego se aplica softmax:

$$y_j = \frac{\exp(u_j)}{\sum_{k=1}^{|V|} \exp(u_k)}$$

obteniendo la probabilidad de que la palabra  $j$  aparezca en el contexto de  $p_I$ .

## X.X Skip-gram: Función de pérdida

### Pérdida logarítmica

Dado un contexto de  $C$  palabras alrededor de la palabra central  $p_I$ , la función de pérdida se define como:

$$E = - \sum_{c=1}^C \log P(p_{O_c} | p_I)$$

donde  $p_{O_c}$  son las palabras del contexto.

## X.X Skip-gram: Actualización

### Reglas de actualización

Durante el entrenamiento, los vectores de entrada y salida se ajustan.

Para los vectores de salida:

$$v'_j \leftarrow v'_j - \eta E_{Ij} h$$

Para el vector de entrada de la palabra central:

$$v_{p_I} \leftarrow v_{p_I} - \eta E H^T$$

donde:

- $\eta$  es la tasa de aprendizaje.
- $E_{Ij}$  depende del error para cada palabra del contexto.