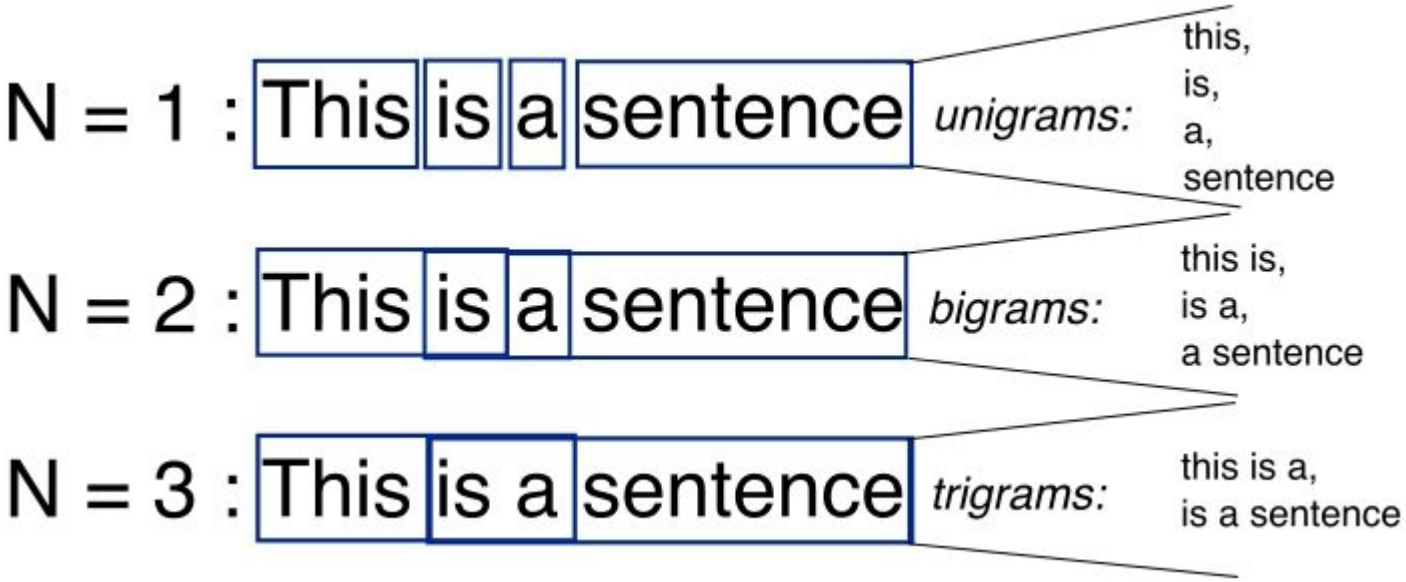
N-Gramas (Fraseado)



Uso de n-gramas

Consiste en unir grupos de 2 o más palabras que, debido a su naturaleza, adquieren mucho más valor cuando están juntas (como si se tratara de una sola palabra), generalmente son usadas en **negaciones**, o palabras que **siempre van juntas** para expresar una idea concreta.

Quiero realizar un retiro sin tarjeta → Quiero realizar un retiro sin-tarjeta

¿Cuál es el saldo en mi tarjeta oro? → Cuál es el saldo en mi tarjeta-oro?

No quiero realizar la operación → No-quiero realizar la operación

¿Qué otros ejemplos se te ocurren de posible n-gramas?

Negación

Producto específico

Determinación de N-gramas

$$\text{N-Grams}_K = K - (N - 1)$$

Donde:
N-Grams: Cantidad de n-gramas posibles
K: Cantidad de tokens en el texto
N: Tamaño del n-grama

Ejemplo: Calcular la cantidad de 3-gramas en el texto:

"La vida es un regalo y no pienso desperdiciarla. Nunca se sabe qué cartas repartirá la próxima vez."

N-Grams = 20 - (3 - 1) = 18

(La vida es), (vida es un), (es un regalo), (un regalo y) (regalo y no), (y no pienso), (no pienso desperdiciarla), ...

Ejemplo y ejercicio:
https://colab.research.google.com/drive/1c8w46awtxzrpDDmODE0KocMuGQcpJKVC?usp=sharing

Stopwords (Palabras vacías)

[“This”, “is”, “a”, “test”]

StopWords

Las **StopWords (Palabras vacías)** son palabras que vuelven difícil el análisis para un sistema de PLN. Pueden ser palabras **muy poco comunes** o **demasiado comunes**, que generen **ambigüedad**.

Por definición, siempre son StopWords:

- Artículos definidos (El, La, Los, Las)
- Artículos indefinidos (Un, Una, Unos, Unas)
- Adj. Posesivos (Mi, Tu, Su, Nuestro, Nuestra)
- Preposiciones (a, con, de, en, para, por, etc)
- Pronombres demostrativos (este, esa, aquel, etc)

*Son las decisiones las que nos hacen ser quienes **somos**, y siempre podemos optar por hacer lo correcto" (Spiderman 3).*

↓

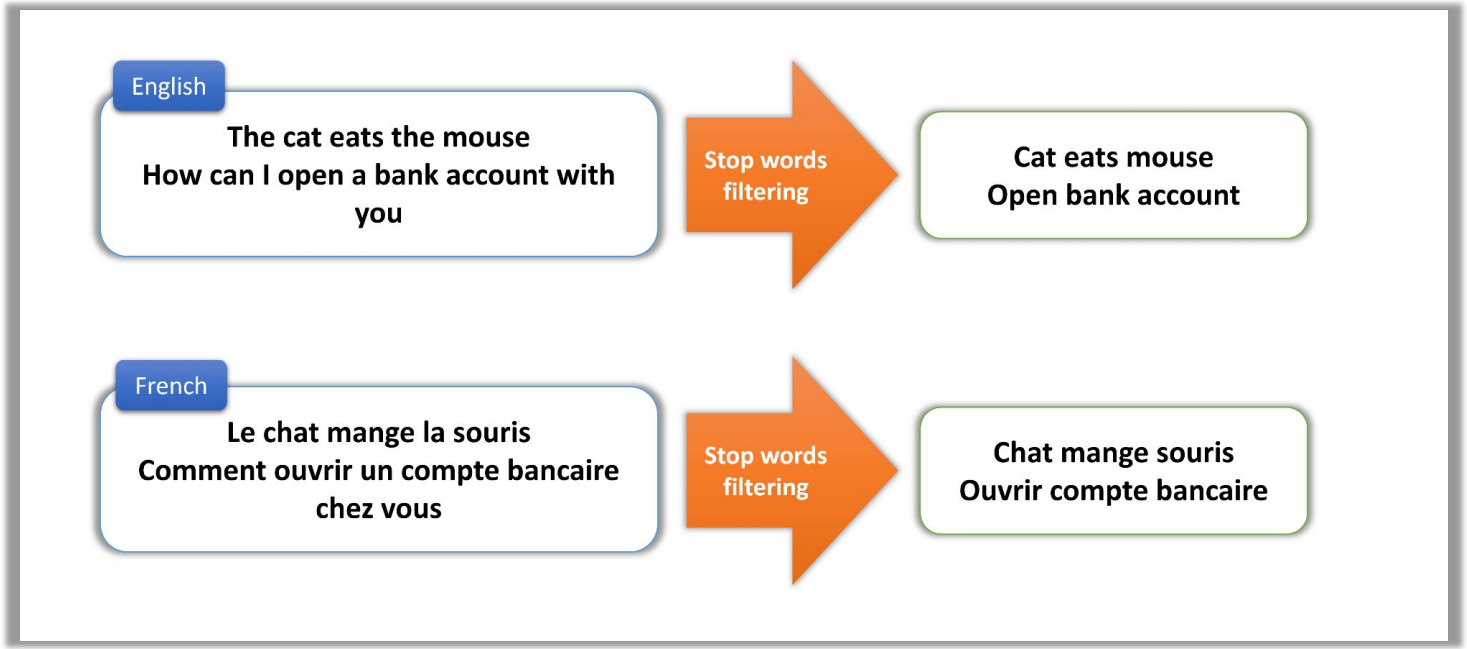
*"**Decisiones nos hacen ser quienes somos**, siempre podemos hacer correcto" (Spiderman 3).*

DATO: *En el idioma japonés algunas de estas palabras no las consideran necesarias en la comprensión de ideas*

UNIVERSIDAD Panamericana

Stopwords

Dependiendo del idioma, la cantidad de StopWords por defecto puede ser menor o mayor.



Ejercicio para determinar StopWords:

A) De los textos de libros, en la carpeta ***"Textos_Stopwords"***, encontrar todas las palabras diferentes y eliminar aquellas que la librería NLTK maneje como *stopwords*. Quitar signos de puntuación y aplicar el método *lower()* para pasar todo a minúsculas.

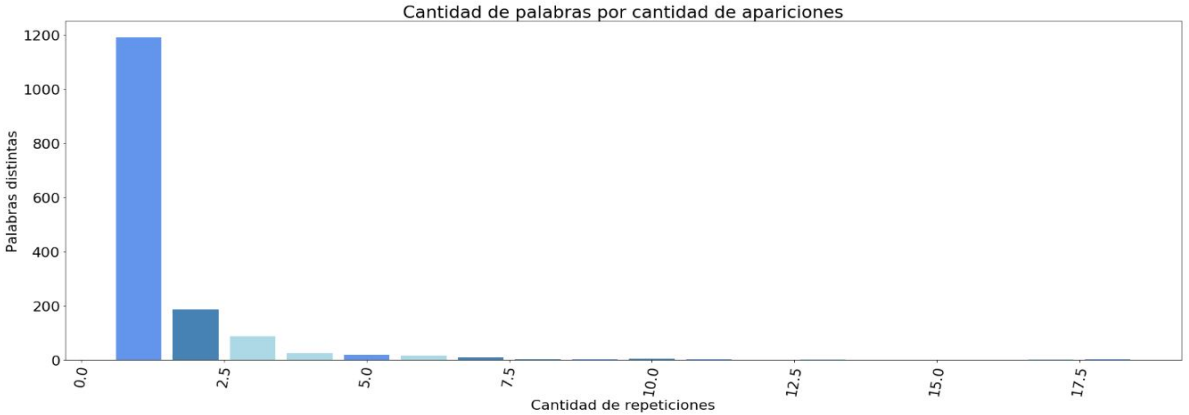
B) Con las palabras únicas, realizar una gráfica de barras (Histograma) mostrando cuántas palabras existen que aparezcan solo una vez, cuantas aparecen 2 veces, cuantas 3, etc…

C) Imprimir cuales son las palabras correspondientes a cada frecuencia y analizar los resultados para determinar hasta qué grado las palabras son relevantes

D) Añadir StopWords personalizadas por el usuario y repetir el proceso de graficado. Observar diferencias

```
Palabras que se repiten 5 veces:
{'cualquier', 'lado', 'mañana', 'luego',
o', 'segura', 'sabía', 'hacía', 'piel',

Palabras que se repiten 6 veces:
{'casi', 'abuela', 'bien', 'momento', 'a
ía', 'fuego'}
```



Tips:

- De la librería *nltk.corpus*, importar *stopwords*, ver documentación de NLTK en: "https://www.nltk.org/"
- Crear un diccionario en el que se almacenen todas las palabras diferentes contenidas en los textos, con base en su frecuencia de aparición.

Solución (Parte 1):

```
# Importar librerías de NLTK
from nltk.tokenize import RegexpTokenizer
from nltk.tokenize.treebank import TreebankWordDetokenizer
from nltk.corpus import stopwords
```

Uso de la librería NLTK para invocar un diccionario de Stopwords en Español predefinido

```
# Asignación de StopWords predefinidas para idioma Español
import nltk
stop_words = nltk.corpus.stopwords.words('spanish')
print(stop_words)
```