# Sprint 8 : Tasca M9 T01

## Exercici 1

### Agafa un text en anglès que vulguis, i calcula'n la freqüència de les paraules.

In [224…

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import MinMaxScaler
from sklearn.model_selection import train_test_split, GridSearchCV
from sklearn.pipeline import Pipeline
```

https://www.supersummary.com/one-hundred-years-of-solitude/summary/

In [225…

```
text=""" One Hundred Years of Solitude tells the story of five generations of the Buendía family from the founding of both their lineage and the town
of Macondo until both are wiped out by a hurricane. The novel is written in third-person omniscient past tense. Time shifts constantly: chapters are
narrated out of chronological order, and the relative importance of events has little relation to how much attention they receive on the page.
Rather than following a single protagonist through a linear set of events, the novel contains many plots and subplots revolving around both the
central characters of the Buendía family across generations and the side characters who interact with them.

The book opens as Colonel Aureliano Buendía remembers his childhood in Macondo, when his father José Arcadio Buendía was interested in all of the
inventions brought to town by travelers. Subsequently, José Arcadio Buendía experiments with new inventions, including alchemy and the astrolabe.
Úrsula Iguaràn, his wife and the Buendía matriarch, becomes frustrated with his behavior and his preoccupation with tinkering over more practical
projects.

Úrsula Iguaràn and José Arcadio Buendía have three children: Colonel Aureliano Buendía, José Arcadio, and Amaranta. Colonel Aureliano Buendía lives
a vibrant and exciting life as a Liberal general during wartime: he fights for Colombian independence from the Conservative government, and only stops
fighting when it's time for an armistice. He has 17 children by women he meets on his military campaigns, and he spends his retirement making tiny
gold fishes in the Buendía house. José Arcadio is the father of Arcadio with Pilar Ternera; when she is pregnant he leaves to become a sailor. He
briefly returns, hale and hearty; while he is in Macondo he marries Rebeca and stops the execution of his brother the Colonel. Amaranta outlasts
her brothers; she dies single and unhappy at home.

Arcadio has three children with Santa Sofía de Piedad: Remedios the Beauty, Aureliano Segundo, and José Arcadio Segundo. Arcadio dies after becoming
a dictatorial ruler of Macondo; Santa Sofía retreats to the background for a few decades and then leaves Macondo to live with a cousin. Remedios the
Beauty lives a short but mysterious life before she ascends into the sky. José Arcadio Segundo develops purpose to his life as a leader of a workers'
strike at a banana company that is exploiting the people of Macondo; he is the sole survivor of a massacre of those same workers.

Aureliano Segundo marries Fernanda del Carpio and takes Petra Cotes as a lover. Aureliano Segundo and Fernanda del Carpio have three children:
Renata Remedios (Meme), José Arcadio, and Amaranta Úrsula. Meme falls in love with a mechanic, and Fernanda sends her to a convent; a nun drops her
son Aureliano off later that year. José Arcadio goes to religious school and meets the Pope before coming home to be murdered by teens in his bed for
gold. Amaranta Úrsula is educated in Europe and returns to Macondo with her Belgian husband, Gaston; they are not completely happy together. When
Gaston returns to Europe, Amaranta Úrsula begins a sexual relationship with her nephew, Aureliano, with whom she has the final child of the Buendía
family.

At the end of the novel, Macondo and the Buendía family home have fallen into ruin. The child of Amaranta Úrsula and Aureliano is born with a pig's
tail because it is the product of incest, the fulfillment of an omen that appeared in the first generation of the Buendía family. Amaranta Úrsula
dies in childbirth, and Aureliano takes the infant into the town square, where he drinks and loses the child. Later, he finds it being devoured by
ants. He returns to the crumbling Buendía home, where he decodes the ancient writings of the wanderer Melquíades. As he reads the parchments, he
realizes that the writings have predicted all of the events that have befallen the Buendía family. A windstorm rises as he is reading, and as he
reads the last line, the storm wipes Macondo off the face of the earth. """
```

In [226…

```
from nltk.tokenize import word_tokenize

# break the text into words
token_word = word_tokenize(text)

# remove punctuation
token_word = [word for word in token_word if word.isalnum()]
```

In [227…

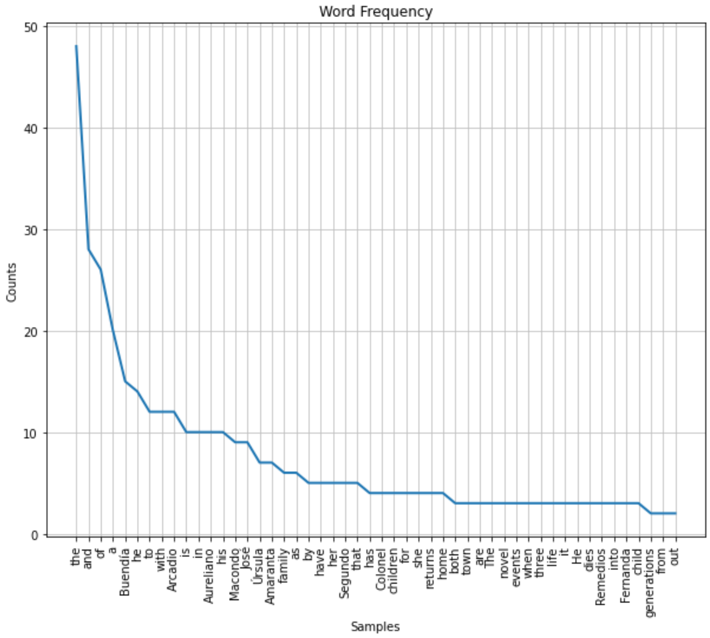
```
from nltk.probability import FreqDist

word_freq = FreqDist(token_word)
word_freq
```

Out[227]:

```
FreqDist({'the': 48, 'and': 28, 'of': 26, 'a': 20, 'Buendía': 15, 'he': 14, 'to': 12, 'with': 12, 'Arcadio': 12, 'is': 10, ...})
```

In [249…

```
#plotting the word frequency
plt.figure(figsize = (10, 8))
plt.title("Word Frequency")
plt.xlabel("Words")
plt.ylabel("Counts")
word_freq.plot(50, cumulative = False)
plt.show()
```



- Las palabras con mayor frecuencia son: the,and, of, a, etc..., seguidas de los nombres de los personajes del libro, Buendía, Arcadio, Aureliano, Macondo, etc...

## Exercici 2

**Treu les stopwords i realitza stemming al teu conjunt de dades.**

## 2.1 Stopwords

- Es una metodología para ignorar las palabras vacías o palabra de uso común (como "el", "un", "una", "en", etc.. ) en un motor de búsqueda , tanto al indexar entradas para la búsqueda como al recuperarlas como resultado final.

In [229…

```
from nltk.corpus import stopwords

#import english stopwords
stop_words = list(stopwords.words("english"))
```

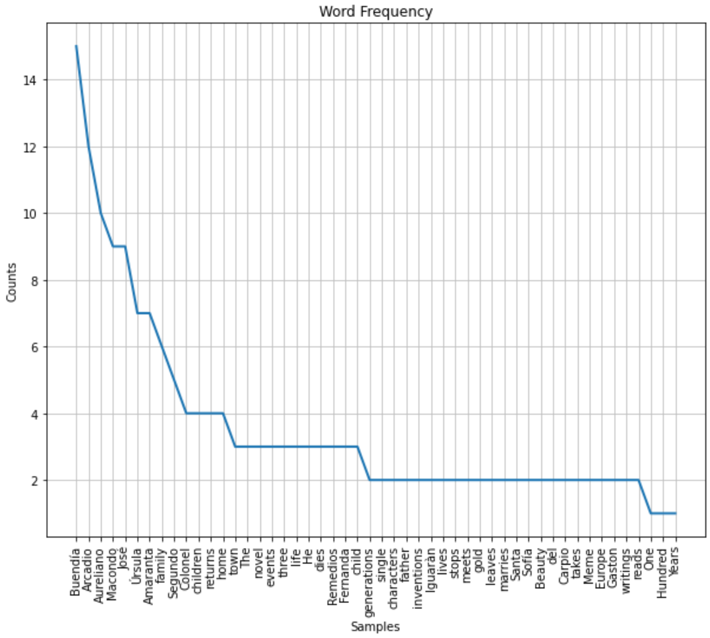
In [230…

```
tokenized_sent = token_word
filtered_sent=[]
for w in tokenized_sent:
    if w not in stop_words:
        filtered_sent.append(w)

print("Tokenized Sentence:",tokenized_sent)
print("")
print("Filterd Sentence:",filtered_sent)
```

```
Tokenized Sentence: ['One', 'Hundred', 'Years', 'of', 'Solitude', 'tells', 'the', 'story', 'of', 'five', 'generations', 'of', 'the', 'Buendía', 'famil
y', 'from', 'the', 'founding', 'of', 'both', 'their', 'lineage', 'and', 'the', 'town', 'of', 'Macondo', 'until', 'both', 'are', 'wiped', 'out', 'by',
'a', 'hurricane', 'The', 'novel', 'is', 'written', 'in', 'omniscient', 'past', 'tense', 'Time', 'shifts', 'constantly', 'chapters', 'are', 'narrated',
'out', 'of', 'chronological', 'order', 'and', 'the', 'relative', 'importance', 'of', 'events', 'has', 'little', 'relation', 'to', 'how', 'much', 'attent
ion', 'they', 'receive', 'on', 'the', 'page', 'Rather', 'than', 'following', 'a', 'single', 'protagonist', 'through', 'a', 'linear', 'set', 'of', 'event
s', 'the', 'novel', 'contains', 'many', 'plots', 'and', 'subplots', 'revolving', 'around', 'both', 'the', 'central', 'characters', 'of', 'the', 'Buendí
a', 'family', 'across', 'generations', 'and', 'the', 'side', 'characters', 'who', 'interact', 'with', 'them', 'The', 'book', 'opens', 'as', 'Colonel',
'Aureliano', 'Buendía', 'remembers', 'his', 'childhood', 'in', 'Macondo', 'when', 'his', 'father', 'José', 'Arcadio', 'Buendía', 'was', 'interested', 'i
n', 'all', 'of', 'the', 'inventions', 'brought', 'to', 'town', 'by', 'travelers', 'Subsequently', 'José', 'Arcadio', 'Buendía', 'experiments', 'with',
'new', 'inventions', 'including', 'alchemy', 'and', 'the', 'astrolabe', 'Úrsula', 'Iguaràn', 'his', 'wife', 'and', 'the', 'Buendía', 'matriarch', 'becom
es', 'frustrated', 'with', 'his', 'behavior', 'and', 'his', 'preoccupation', 'with', 'tinkering', 'over', 'more', 'practical', 'projects', 'Úrsula', 'Ig
uaràn', 'and', 'José', 'Arcadio', 'Buendía', 'have', 'three', 'children', 'Colonel', 'Aureliano', 'Buendía', 'José', 'Arcadio', 'and', 'Amaranta', 'Colo
nel', 'Aureliano', 'Buendía', 'lives', 'a', 'vibrant', 'and', 'exciting', 'life', 'as', 'a', 'Liberal', 'general', 'during', 'wartime', 'he', 'fights',
'for', 'Colombian', 'independence', 'from', 'the', 'Conservative', 'government', 'and', 'only', 'stops', 'fighting', 'when', 'it', 'time', 'for', 'an',
'armistice', 'He', 'has', '17', 'children', 'by', 'women', 'he', 'meets', 'on', 'his', 'military', 'campaigns', 'and', 'he', 'spends', 'his', 'retiremen
t', 'making', 'tiny', 'gold', 'fishes', 'in', 'the', 'Buendía', 'house', 'José', 'Arcadio', 'is', 'the', 'father', 'of', 'Arcadio', 'with', 'Pilar', 'Te
rnera', 'when', 'she', 'is', 'pregnant', 'he', 'leaves', 'to', 'become', 'a', 'sailor', 'He', 'briefly', 'returns', 'hale', 'and', 'hearty', 'while', 'h
e', 'is', 'in', 'Macondo', 'he', 'marries', 'Rebeca', 'and', 'stops', 'the', 'execution', 'of', 'his', 'brother', 'the', 'Colonel', 'Amaranta', 'outlast
s', 'her', 'brothers', 'she', 'dies', 'single', 'and', 'unhappy', 'at', 'home', 'Arcadio', 'has', 'three', 'children', 'with', 'Santa', 'Sofía', 'de',
'Piedad', 'Remedios', 'the', 'Beauty', 'Aureliano', 'Segundo', 'and', 'José', 'Arcadio', 'Segundo', 'Arcadio', 'dies', 'after', 'becoming', 'a', 'dictat
orial', 'ruler', 'of', 'Macondo', 'Santa', 'Sofía', 'retreats', 'to', 'the', 'background', 'for', 'a', 'few', 'decades', 'and', 'then', 'leaves', 'Macon
do', 'to', 'live', 'with', 'a', 'cousin', 'Remedios', 'the', 'Beauty', 'lives', 'a', 'short', 'but', 'mysterious', 'life', 'before', 'she', 'ascends',
'into', 'the', 'sky', 'José', 'Arcadio', 'Segundo', 'develops', 'purpose', 'to', 'his', 'life', 'as', 'a', 'leader', 'of', 'a', 'strike', 'at', 'a', 'ba
nana', 'company', 'that', 'is', 'exploiting', 'the', 'people', 'of', 'Macondo', 'he', 'is', 'the', 'sole', 'survivor', 'of', 'a', 'massacre', 'of', 'tho
se', 'same', 'workers', 'Aureliano', 'Segundo', 'marries', 'Fernanda', 'del', 'Carpio', 'and', 'takes', 'Petra', 'Cotes', 'as', 'a', 'lover', 'Aurelian
o', 'Segundo', 'and', 'Fernanda', 'del', 'Carpio', 'have', 'three', 'children', 'Renata', 'Remedios', 'Meme', 'José', 'Arcadio', 'and', 'Amaranta', 'Úrs
ula', 'Meme', 'falls', 'in', 'love', 'with', 'a', 'mechanic', 'and', 'Fernanda', 'sends', 'her', 'to', 'a', 'convent', 'a', 'nun', 'drops', 'her', 'so
n', 'Aureliano', 'off', 'later', 'that', 'year', 'José', 'Arcadio', 'goes', 'to', 'religious', 'school', 'and', 'meets', 'the', 'Pope', 'before', 'comin
g', 'home', 'to', 'be', 'murdered', 'by', 'teens', 'in', 'his', 'bed', 'for', 'gold', 'Amaranta', 'Úrsula', 'is', 'educated', 'in', 'Europe', 'and', 're
turns', 'to', 'Macondo', 'with', 'her', 'Belgian', 'husband', 'Gaston', 'they', 'are', 'not', 'completely', 'happy', 'together', 'When', 'Gaston', 'retu
rns', 'to', 'Europe', 'Amaranta', 'Úrsula', 'begins', 'a', 'sexual', 'relationship', 'with', 'her', 'nephew', 'Aureliano', 'with', 'whom', 'she', 'has',
'the', 'final', 'child', 'of', 'the', 'Buendía', 'family', 'At', 'the', 'end', 'of', 'the', 'novel', 'Macondo', 'and', 'the', 'Buendía', 'family', 'hom
e', 'have', 'fallen', 'into', 'ruin', 'The', 'child', 'of', 'Amaranta', 'Úrsula', 'and', 'Aureliano', 'is', 'born', 'with', 'a', 'pig', 's', 'tail', 'be
cause', 'it', 'is', 'the', 'product', 'of', 'incest', 'the', 'fulfillment', 'of', 'an', 'omen', 'that', 'appeared', 'in', 'the', 'first', 'generation',
'of', 'the', 'Buendía', 'family', 'Amaranta', 'Úrsula', 'dies', 'in', 'childbirth', 'and', 'Aureliano', 'takes', 'the', 'infant', 'into', 'the', 'town',
'square', 'where', 'he', 'drinks', 'and', 'loses', 'the', 'child', 'Later', 'he', 'finds', 'it', 'being', 'devoured', 'by', 'ants', 'He', 'returns', 't
o', 'the', 'crumbling', 'Buendía', 'home', 'where', 'he', 'decodes', 'the', 'ancient', 'writings', 'of', 'the', 'wanderer', 'Melquíades', 'As', 'he', 'r
eads', 'the', 'parchments', 'he', 'realizes', 'that', 'the', 'writings', 'have', 'predicted', 'all', 'of', 'the', 'events', 'that', 'have', 'befallen',
'the', 'Buendía', 'family', 'A', 'windstorm', 'rises', 'as', 'he', 'is', 'reading', 'and', 'as', 'he', 'reads', 'the', 'last', 'line', 'the', 'storm',
'wipes', 'Macondo', 'off', 'the', 'face', 'of', 'the', 'earth']

Filterd Sentence: ['One', 'Hundred', 'Years', 'Solitude', 'tells', 'story', 'five', 'generations', 'Buendía', 'family', 'founding', 'lineage', 'town',
'Macondo', 'wiped', 'hurricane', 'The', 'novel', 'written', 'omniscient', 'past', 'tense', 'Time', 'shifts', 'constantly', 'chapters', 'narrated', 'chro
nological', 'order', 'relative', 'importance', 'events', 'little', 'relation', 'much', 'attention', 'receive', 'page', 'Rather', 'following', 'single',
'protagonist', 'linear', 'set', 'events', 'novel', 'contains', 'many', 'plots', 'subplots', 'revolving', 'around', 'central', 'characters', 'Buendía',
'family', 'across', 'generations', 'side', 'characters', 'interact', 'The', 'book', 'opens', 'Colonel', 'Aureliano', 'Buendía', 'remembers', 'childhoo
d', 'Macondo', 'father', 'José', 'Arcadio', 'Buendía', 'interested', 'inventions', 'brought', 'town', 'travelers', 'Subsequently', 'José', 'Arcadio', 'B
uendía', 'experiments', 'new', 'inventions', 'including', 'alchemy', 'astrolabe', 'Úrsula', 'Iguaràn', 'wife', 'Buendía', 'matriarch', 'becomes', 'frust
rated', 'behavior', 'preoccupation', 'tinkering', 'practical', 'projects', 'Úrsula', 'Iguaràn', 'José', 'Arcadio', 'Buendía', 'three', 'children', 'Colo
nel', 'Aureliano', 'Buendía', 'José', 'Arcadio', 'Amaranta', 'Colonel', 'Aureliano', 'Buendía', 'lives', 'vibrant', 'exciting', 'life', 'Liberal', 'gene
ral', 'wartime', 'fights', 'Colombian', 'independence', 'Conservative', 'government', 'stops', 'fighting', 'time', 'armistice', 'He', '17', 'children',
'women', 'meets', 'military', 'campaigns', 'spends', 'retirement', 'making', 'tiny', 'gold', 'fishes', 'Buendía', 'house', 'José', 'Arcadio', 'father',
'Arcadio', 'Pilar', 'Ternera', 'pregnant', 'leaves', 'become', 'sailor', 'He', 'briefly', 'returns', 'hale', 'hearty', 'Macondo', 'marries', 'Rebeca',
'stops', 'execution', 'brother', 'Colonel', 'Amaranta', 'outlasts', 'brothers', 'dies', 'single', 'unhappy', 'home', 'Arcadio', 'three', 'children', 'Sa
nta', 'Sofía', 'de', 'Piedad', 'Remedios', 'Beauty', 'Aureliano', 'Segundo', 'José', 'Arcadio', 'Segundo', 'Arcadio', 'dies', 'becoming', 'dictatorial',
'ruler', 'Macondo', 'Santa', 'Sofía', 'retreats', 'background', 'decades', 'leaves', 'Macondo', 'live', 'cousin', 'Remedios', 'Beauty', 'lives', 'shor
t', 'mysterious', 'life', 'ascends', 'sky', 'José', 'Arcadio', 'Segundo', 'develops', 'purpose', 'life', 'leader', 'strike', 'banana', 'company', 'explo
iting', 'people', 'Macondo', 'sole', 'survivor', 'massacre', 'workers', 'Aureliano', 'Segundo', 'marries', 'Fernanda', 'del', 'Carpio', 'takes', 'Petr
a', 'Cotes', 'lover', 'Aureliano', 'Segundo', 'Fernanda', 'del', 'Carpio', 'three', 'children', 'Renata', 'Remedios', 'Meme', 'José', 'Arcadio', 'Amaran
ta', 'Úrsula', 'Meme', 'falls', 'love', 'mechanic', 'Fernanda', 'sends', 'convent', 'nun', 'drops', 'son', 'Aureliano', 'later', 'year', 'José', 'Arcadi
o', 'goes', 'religious', 'school', 'meets', 'Pope', 'coming', 'home', 'murdered', 'teens', 'bed', 'gold', 'Amaranta', 'Úrsula', 'educated', 'Europe', 'r
eturns', 'Macondo', 'Belgian', 'husband', 'Gaston', 'completely', 'happy', 'together', 'When', 'Gaston', 'returns', 'Europe', 'Amaranta', 'Úrsula', 'beg
ins', 'sexual', 'relationship', 'nephew', 'Aureliano', 'final', 'child', 'Buendía', 'family', 'At', 'end', 'novel', 'Macondo', 'Buendía', 'family', 'hom
e', 'fallen', 'ruin', 'The', 'child', 'Amaranta', 'Úrsula', 'Aureliano', 'born', 'pig', 'tail', 'product', 'incest', 'fulfillment', 'omen', 'appeared',
'first', 'generation', 'Buendía', 'family', 'Amaranta', 'Úrsula', 'dies', 'childbirth', 'Aureliano', 'takes', 'infant', 'town', 'square', 'drinks', 'los
es', 'child', 'Later', 'finds', 'devoured', 'ants', 'He', 'returns', 'crumbling', 'Buendía', 'home', 'decodes', 'ancient', 'writings', 'wanderer', 'Melq
uíades', 'As', 'reads', 'parchments', 'realizes', 'writings', 'predicted', 'events', 'befallen', 'Buendía', 'family', 'A', 'windstorm', 'rises', 'readin
g', 'reads', 'last', 'line', 'storm', 'wipes', 'Macondo', 'face', 'earth']
```

In [250…

```python
word_freq_stemmed = FreqDist(filtered_sent)

#plotting the word frequency
plt.figure(figsize = (10, 8))
plt.title("Word Frequency")
plt.xlabel("Words")
plt.ylabel("Counts")
word_freq_stemmed.plot(50, cumulative = False)
plt.show()
```



- Una vez eliminadas las stopwords, las palabras con mayor frecuencia son los nombres de los personajes del libro, como era previsible, Buendía, Arcadio, Aureliano, Macondo, José, etc..

## 2.2 Normalización de Léxico: Stemming

- Es un proceso de normalización lingüística, que reduce las palabras a su raíz o elimina los afijos derivativos. Por ejemplo, conexión, conectado, palabra de conexión se reduce a una palabra común "conectar".

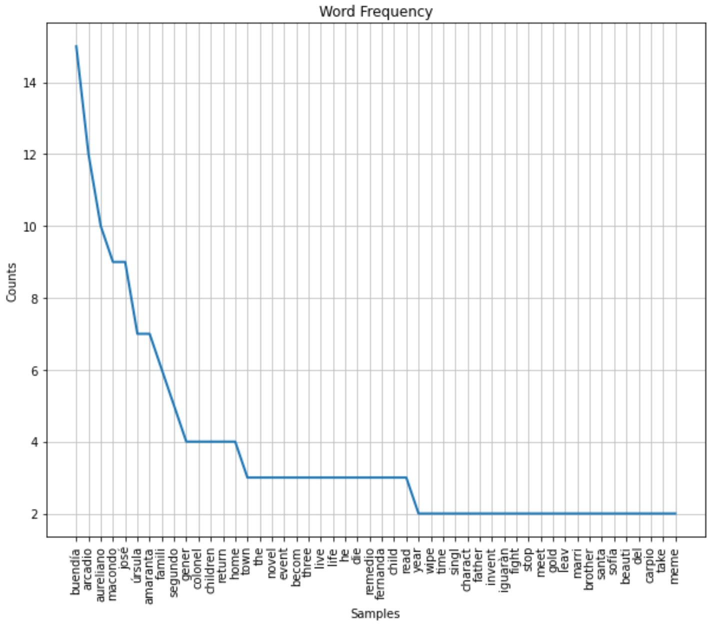
In [235…

```
# Stemming
from nltk.stem import PorterStemmer
from nltk.tokenize import sent_tokenize, word_tokenize

ps = PorterStemmer()

stemmed_words=[]
for w in filtered_sent:
    stemmed_words.append(ps.stem(w))

print("Filtered Sentence:",filtered_sent)
print(" ")
print("Stemmed Sentence:",stemmed_words)
```

```
Filtered Sentence: ['One', 'Hundred', 'Years', 'Solitude', 'tells', 'story', 'five', 'generations', 'Buendía', 'family', 'founding', 'lineage', 'town',
'Macondo', 'wiped', 'hurricane', 'The', 'novel', 'written', 'omniscient', 'past', 'tense', 'Time', 'shifts', 'constantly', 'chapters', 'narrated', 'chro
nological', 'order', 'relative', 'importance', 'events', 'little', 'relation', 'much', 'attention', 'receive', 'page', 'Rather', 'following', 'single',
'protagonist', 'linear', 'set', 'events', 'novel', 'contains', 'many', 'plots', 'subplots', 'revolving', 'around', 'central', 'characters', 'Buendía',
'family', 'across', 'generations', 'side', 'characters', 'interact', 'The', 'book', 'opens', 'Colonel', 'Aureliano', 'Buendía', 'remembers', 'childhoo
d', 'Macondo', 'father', 'José', 'Arcadio', 'Buendía', 'interested', 'inventions', 'brought', 'town', 'travelers', 'Subsequently', 'José', 'Arcadio', 'B
uendía', 'experiments', 'new', 'inventions', 'including', 'alchemy', 'astrolabe', 'Úrsula', 'Iguaràn', 'wife', 'Buendía', 'matriarch', 'becomes', 'frust
rated', 'behavior', 'preoccupation', 'tinkering', 'practical', 'projects', 'Úrsula', 'Iguaràn', 'José', 'Arcadio', 'Buendía', 'three', 'children', 'Colo
nel', 'Aureliano', 'Buendía', 'José', 'Arcadio', 'Amaranta', 'Colonel', 'Aureliano', 'Buendía', 'lives', 'vibrant', 'exciting', 'life', 'Liberal', 'gene
ral', 'wartime', 'fights', 'Colombian', 'independence', 'Conservative', 'government', 'stops', 'fighting', 'time', 'armistice', 'He', '17', 'children',
'women', 'meets', 'military', 'campaigns', 'spends', 'retirement', 'making', 'tiny', 'gold', 'fishes', 'Buendía', 'house', 'José', 'Arcadio', 'father',
'Arcadio', 'Pilar', 'Ternera', 'pregnant', 'leaves', 'become', 'sailor', 'He', 'briefly', 'returns', 'hale', 'hearty', 'Macondo', 'marries', 'Rebeca',
'stops', 'execution', 'brother', 'Colonel', 'Amaranta', 'outlasts', 'brothers', 'dies', 'single', 'unhappy', 'home', 'Arcadio', 'three', 'children', 'Sa
nta', 'Sofía', 'de', 'Piedad', 'Remedios', 'Beauty', 'Aureliano', 'Segundo', 'José', 'Arcadio', 'Segundo', 'Arcadio', 'dies', 'becoming', 'dictatorial',
'ruler', 'Macondo', 'Santa', 'Sofía', 'retreats', 'background', 'decades', 'leaves', 'Macondo', 'live', 'cousin', 'Remedios', 'Beauty', 'lives', 'shor
t', 'mysterious', 'life', 'ascends', 'sky', 'José', 'Arcadio', 'Segundo', 'develops', 'purpose', 'life', 'leader', 'strike', 'banana', 'company', 'explo
iting', 'people', 'Macondo', 'sole', 'survivor', 'massacre', 'workers', 'Aureliano', 'Segundo', 'marries', 'Fernanda', 'del', 'Carpio', 'takes', 'Petr
a', 'Cotes', 'lover', 'Aureliano', 'Segundo', 'Fernanda', 'del', 'Carpio', 'three', 'children', 'Renata', 'Remedios', 'Meme', 'José', 'Arcadio', 'Amaran
ta', 'Úrsula', 'Meme', 'falls', 'love', 'mechanic', 'Fernanda', 'sends', 'convent', 'nun', 'drops', 'son', 'Aureliano', 'later', 'year', 'José', 'Arcadi
o', 'goes', 'religious', 'school', 'meets', 'Pope', 'coming', 'home', 'murdered', 'teens', 'bed', 'gold', 'Amaranta', 'Úrsula', 'educated', 'Europe', 'r
eturns', 'Macondo', 'Belgian', 'husband', 'Gaston', 'completely', 'happy', 'together', 'When', 'Gaston', 'returns', 'Europe', 'Amaranta', 'Úrsula', 'beg
ins', 'sexual', 'relationship', 'nephew', 'Aureliano', 'final', 'child', 'Buendía', 'family', 'At', 'end', 'novel', 'Macondo', 'Buendía', 'family', 'hom
e', 'fallen', 'ruin', 'The', 'child', 'Amaranta', 'Úrsula', 'Aureliano', 'born', 'pig', 'tail', 'product', 'incest', 'fulfillment', 'omen', 'appeared',
'first', 'generation', 'Buendía', 'family', 'Amaranta', 'Úrsula', 'dies', 'childbirth', 'Aureliano', 'takes', 'infant', 'town', 'square', 'drinks', 'los
es', 'child', 'Later', 'finds', 'devoured', 'ants', 'He', 'returns', 'crumbling', 'Buendía', 'home', 'decodes', 'ancient', 'writings', 'wanderer', 'Melq
uíades', 'As', 'reads', 'parchments', 'realizes', 'writings', 'predicted', 'events', 'befallen', 'Buendía', 'family', 'A', 'windstorm', 'rises', 'readin
g', 'reads', 'last', 'line', 'storm', 'wipes', 'Macondo', 'face', 'earth']

Stemmed Sentence: ['one', 'hundr', 'year', 'solitud', 'tell', 'stori', 'five', 'gener', 'buendía', 'famili', 'found', 'lineag', 'town', 'macondo', 'wip
e', 'hurrican', 'the', 'novel', 'written', 'omnisci', 'past', 'tens', 'time', 'shift', 'constantli', 'chapter', 'narrat', 'chronolog', 'order', 'rel',
'import', 'event', 'littl', 'relat', 'much', 'attent', 'receiv', 'page', 'rather', 'follow', 'singl', 'protagonist', 'linear', 'set', 'event', 'novel',
'contain', 'mani', 'plot', 'subplot', 'revolv', 'around', 'central', 'charact', 'buendía', 'famili', 'across', 'gener', 'side', 'charact', 'interact',
'the', 'book', 'open', 'colonel', 'aureliano', 'buendía', 'rememb', 'childhood', 'macondo', 'father', 'josé', 'arcadio', 'buendía', 'interest', 'inven
t', 'brought', 'town', 'travel', 'subsequ', 'josé', 'arcadio', 'buendía', 'experi', 'new', 'invent', 'includ', 'alchemi', 'astrolab', 'úrsula', 'iguarà
n', 'wife', 'buendía', 'matriarch', 'becom', 'frustrat', 'behavior', 'preoccup', 'tinker', 'practic', 'project', 'úrsula', 'iguaràn', 'josé', 'arcadio',
'buendía', 'three', 'children', 'colonel', 'aureliano', 'buendía', 'josé', 'arcadio', 'amaranta', 'colonel', 'aureliano', 'buendía', 'live', 'vibrant',
'excit', 'life', 'liber', 'gener', 'wartim', 'fight', 'colombian', 'independ', 'conserv', 'govern', 'stop', 'fight', 'time', 'armistic', 'he', '17', 'ch
ildren', 'women', 'meet', 'militari', 'campaign', 'spend', 'retir', 'make', 'tini', 'gold', 'fish', 'buendía', 'hous', 'josé', 'arcadio', 'father', 'arc
adio', 'pilar', 'ternera', 'pregnant', 'leav', 'becom', 'sailor', 'he', 'briefli', 'return', 'hale', 'hearti', 'macondo', 'marri', 'rebeca', 'stop', 'ex
ecut', 'brother', 'colonel', 'amaranta', 'outlast', 'brother', 'die', 'singl', 'unhappi', 'home', 'arcadio', 'three', 'children', 'santa', 'sofía', 'd
e', 'piedad', 'remedio', 'beauti', 'aureliano', 'segundo', 'josé', 'arcadio', 'segundo', 'arcadio', 'die', 'becom', 'dictatori', 'ruler', 'macondo', 'sa
nta', 'sofía', 'retreat', 'background', 'decad', 'leav', 'macondo', 'live', 'cousin', 'remedio', 'beauti', 'live', 'short', 'mysteri', 'life', 'ascend',
'sky', 'josé', 'arcadio', 'segundo', 'develop', 'purpos', 'life', 'leader', 'strike', 'banana', 'compani', 'exploit', 'peopl', 'macondo', 'sole', 'survi
vor', 'massacr', 'worker', 'aureliano', 'segundo', 'marri', 'fernanda', 'del', 'carpio', 'take', 'petra', 'cote', 'lover', 'aureliano', 'segundo', 'fern
anda', 'del', 'carpio', 'three', 'children', 'renata', 'remedio', 'meme', 'josé', 'arcadio', 'amaranta', 'úrsula', 'meme', 'fall', 'love', 'mechan', 'fe
rnanda', 'send', 'convent', 'nun', 'drop', 'son', 'aureliano', 'later', 'year', 'josé', 'arcadio', 'goe', 'religi', 'school', 'meet', 'pope', 'come', 'h
ome', 'murder', 'teen', 'bed', 'gold', 'amaranta', 'úrsula', 'educ', 'europ', 'return', 'macondo', 'belgian', 'husband', 'gaston', 'complet', 'happi',
'togeth', 'when', 'gaston', 'return', 'europ', 'amaranta', 'úrsula', 'begin', 'sexual', 'relationship', 'nephew', 'aureliano', 'final', 'child', 'buendí
a', 'famili', 'at', 'end', 'novel', 'macondo', 'buendía', 'famili', 'home', 'fallen', 'ruin', 'the', 'child', 'amaranta', 'úrsula', 'aureliano', 'born',
'pig', 'tail', 'product', 'incest', 'fulfil', 'omen', 'appear', 'first', 'gener', 'buendía', 'famili', 'amaranta', 'úrsula', 'die', 'childbirth', 'aurel
iano', 'take', 'infant', 'town', 'squar', 'drink', 'lose', 'child', 'later', 'find', 'devour', 'ant', 'he', 'return', 'crumbl', 'buendía', 'home', 'deco
d', 'ancient', 'write', 'wander', 'melquíad', 'as', 'read', 'parchment', 'realiz', 'write', 'predict', 'event', 'befallen', 'buendía', 'famili', 'a', 'w
indstorm', 'rise', 'read', 'read', 'last', 'line', 'storm', 'wipe', 'macondo', 'face', 'earth']
```

In [251…

```python
word_freq_stemmed = FreqDist(stemmed_words)

#plotting the word frequency
plt.figure(figsize = (10, 8))
plt.title("Word Frequency")
plt.xlabel("Words")
plt.ylabel("Counts")
word_freq_stemmed.plot(50, cumulative = False)
plt.show()
```



- El gráfico muestra que las palabras más frecuentes siguen siendo los nombres de los protagonistas del libro y se ha eliminado algunas de las que aparecian en el gráfico anterior después del steamming.

## Exercici 3

**Realitza sentiment analysis al teu conjunt de dades.**

In [252…

```python
import nltk
```

```
from nltk.sentiment.vader import SentimentIntensityAnalyzer
```

In [253…

```
nltk.download('vader_lexicon')
```

```
[nltk_data] Downloading package vader_lexicon to
[nltk_data]     C:\Users\hecto\AppData\Roaming\nltk_data...
[nltk_data]   Package vader_lexicon is already up-to-date!
```

Out[253]:

```
True
```

### 3.1 Sentiment Analysis con SentimentIntensityAnalyzer():

In [254…

```
# sentiment of the original text
sentiment = SentimentIntensityAnalyzer()
sentiment.polarity_scores(text)
```

Out[254]:

```
{'neg': 0.057, 'neu': 0.896, 'pos': 0.048, 'compound': -0.7369}
```

SentimentIntensityAnalyzer() es una función que nos proporciona polarity_scores, que es un método que facilita las puntuaciones para las siguientes categorías:

- Negativo
- Neutral
- Positivo
- Compuesto

La puntuación compuesta es la suma de puntuaciones positivas, negativas y neutras, normalizadas entre -1 (negativo más extremo) y +1 (positivo más extremo).

In [255…

```
sent_dict = sentiment.polarity_scores(text)
print(type(sent_dict))
```

```
<class 'dict'>
```

In [256…

```
def sent_score():

    if sent_dict['compound'] >= 0.05 :
        return 'Positive'
    elif sent_dict['compound'] <= - 0.05 :
        return 'Negative'
    else :
        return 'Neutral'

result = sent_score()

print("Overall text is : ", sent_dict)
print("Text was rated as ", sent_dict['neg']*100, "% Negative")
print("Text was rated as ", round(sent_dict['neu']*100,1), "% Neutral")
print("Text was rated as ", sent_dict['pos']*100, "% Positive")

print("The overall text was rated as",result, "with a score of", sent_dict['compound']*100, "% Compound")
```

```
Overall text is :  {'neg': 0.057, 'neu': 0.896, 'pos': 0.048, 'compound': -0.7369}
Text was rated as  5.7 % Negative
Text was rated as  89.6 % Neutral
Text was rated as  4.8 % Positive
The overall text was rated as Negative with a score of -73.69 % Compound
```

### 3.2 Sentiment Analysis de las frases individuales del texto:

In [257…

```
# first we split the text into sentences
sentences = nltk.sent_tokenize(text)
```

In [258…

```
# we apply Sentiment analysis individually on each sentence of the text
control=0
resultados=pd.DataFrame()

for sentence in sentences:
    control=control+1
    sentiment = SentimentIntensityAnalyzer()
    print( "Sentencia número ", control,":", end="")
    scores = sentiment.polarity_scores(sentence)
    print (scores)

    resul= pd.DataFrame.from_dict(scores, orient='index')
    resultados = resultados.append(resul, ignore_index=False)

    print('')
```

```
Sentencia número  1 :{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0}

Sentencia número  2 :{'neg': 0.205, 'neu': 0.598, 'pos': 0.197, 'compound': -0.0258}

Sentencia número  3 :{'neg': 0.0, 'neu': 0.915, 'pos': 0.085, 'compound': 0.3612}

Sentencia número  4 :{'neg': 0.0, 'neu': 0.94, 'pos': 0.06, 'compound': 0.3182}

Sentencia número  5 :{'neg': 0.0, 'neu': 0.915, 'pos': 0.085, 'compound': 0.4019}

Sentencia número  6 :{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0}

Sentencia número  7 :{'neg': 0.139, 'neu': 0.861, 'pos': 0.0, 'compound': -0.5267}

Sentencia número  8 :{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0}

Sentencia número  9 :{'neg': 0.168, 'neu': 0.668, 'pos': 0.163, 'compound': 0.2023}

Sentencia número  10 :{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0}

Sentencia número  11 :{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0}

Sentencia número  12 :{'neg': 0.068, 'neu': 0.932, 'pos': 0.0, 'compound': -0.1531}

Sentencia número  13 :{'neg': 0.219, 'neu': 0.781, 'pos': 0.0, 'compound': -0.4215}

Sentencia número  14 :{'neg': 0.0, 'neu': 0.817, 'pos': 0.183, 'compound': 0.5859}

Sentencia número  15 :{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0}

Sentencia número  16 :{'neg': 0.0, 'neu': 0.844, 'pos': 0.156, 'compound': 0.34}

Sentencia número  17 :{'neg': 0.116, 'neu': 0.818, 'pos': 0.066, 'compound': -0.2263}

Sentencia número  18 :{'neg': 0.0, 'neu': 0.743, 'pos': 0.257, 'compound': 0.5859}

Sentencia número  19 :{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0}

Sentencia número  20 :{'neg': 0.0, 'neu': 0.826, 'pos': 0.174, 'compound': 0.6369}

Sentencia número  21 :{'neg': 0.167, 'neu': 0.833, 'pos': 0.0, 'compound': -0.6597}

Sentencia número  22 :{'neg': 0.138, 'neu': 0.862, 'pos': 0.0, 'compound': -0.4964}

Sentencia número  23 :{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0}

Sentencia número  24 :{'neg': 0.292, 'neu': 0.602, 'pos': 0.106, 'compound': -0.6124}

Sentencia número  25 :{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0}

Sentencia número  26 :{'neg': 0.103, 'neu': 0.897, 'pos': 0.0, 'compound': -0.3182}

Sentencia número  27 :{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0}

Sentencia número  28 :{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0}

Sentencia número  29 :{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0}

Sentencia número  30 :{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0}
```

In [259…

```
resultados['index']=resultados.index
```

In [260…

```
resultados.columns=['valores','score_sentiment']
resultados.columns
```

Out[260]:

```
Index(['valores', 'score_sentiment'], dtype='object')
```

In [261…

```
status= resultados.groupby("score_sentiment")[["valores"]].sum()
status
```

Out[261]:

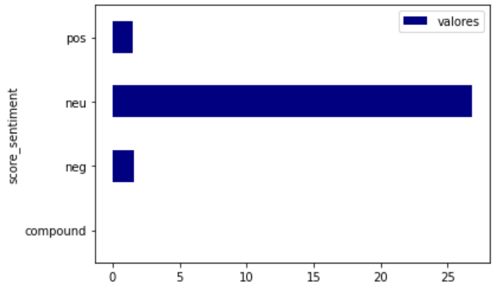
| | **valores** |
|---|---|
| **score_sentiment** | |
| **compound** | -0.0078 |
| **neg** | 1.6150 |
| **neu** | 26.8520 |
| **pos** | 1.5320 |

In [262…

```
status.plot(kind="barh", color="navy")
```

Out[262]:

```
<AxesSubplot:ylabel='score_sentiment'>
```



- La suma de los varores de cada una de las 30 sentencias regleja que en general el texto es neutro, y que las puntuaciones positivas y negativas se compensan, por lo que el "compound" global es -0.0078

### 3.3 Sentiment Analysis a partir del léxico con TextBlob:

- El segundo enfoques para el análisis de sentimiento se basa en el léxico y ea un modelo de de aprendizaje no supervisado. En los análisis anteriores necesitamos datos preetiquetados, pero el segundo no, lo cual es una ventaja.
- Textblob es una biblioteca utilizada para NLP (Natural Language Processing) que tiene diccionarios especiales (léxico) con información detallada, especialmente precisa y preparada para el análisis de sentimientos. Estos léxicos tienen una lista de palabras polares positivas y negativas con una puntuación asociada y usando varias técnicas, como la posición de las palabras, las palabras circundantes, el contexto, las partes del discurso, frases, etc., se asignan puntuaciones a los texto para el que queremos calcular el sentimiento, la suma total es el sentimiento final del texto.

In [263…

```
from textblob import TextBlob
sent = TextBlob(text)
print(sent.sentiment)
```

```
Sentiment(polarity=0.02462876271015806, subjectivity=0.3963127957314004)
```

- La polaridad es una métrica utilizada para estimar el sentimiento general de un texto y varía de -1 a 1, donde -1 sería completamente negativo, 0 es neutral y 1 es completamente positivo. En este caso, está muy cerca de 0, por lo que el sentimiento general es neutral.

- La subjetividad es una métrica que va de 0 a 1, donde 0 es muy objetivo y 1 es muy subjetivo. En este caso está más próximo a la objetividad, 0.39, al tratarse de una narración donde el autor no tiene, evidentemente una opinioón o posición personal.