import nltk

from nltk.corpus import stopwords

from nltk.tokenize import word\_tokenize

from nltk.probability import FreqDist

from nltk.stem import WordNetLemmatizer

*# 1. Read the Moby Dick file*

with open('moby\_dick.txt', 'r') as file:

text = file.read()

*# 2. Tokenization*

tokens = word\_tokenize(text.lower())

*# 3. Stop-words filtering*

stop\_words = set(stopwords.words('english'))

filtered\_tokens = [token for token in tokens if token not in stop\_words]

*# 4. Parts-of-Speech (POS) tagging*

pos\_tags = nltk.pos\_tag(filtered\_tokens)

*# 5. POS frequency*

pos\_counts = {}

for word, pos in pos\_tags:

if pos in pos\_counts:

pos\_counts[pos] += 1

else:

pos\_counts[pos] = 1

# Get the 5 most common parts of speech

top\_pos = sorted(pos\_counts.items(), key=lambda x: x[1], reverse=True)[:5]

print("The 5 most common parts of speech:")

for pos, count in top\_pos:

print(f"{pos}: {count}")

*# 6. Lemmatization*

lemmatizer = WordNetLemmatizer()

lemmatized\_tokens = [lemmatizer.lemmatize(token, pos=pos) for token, pos in pos\_tags[:20]]

print("\nTop 20 lemmatized tokens:")

for token in lemmatized\_tokens:

print(token)

*# Plotting frequency distribution*

freq\_dist = FreqDist(pos\_tags)

freq\_dist.plot(30, cumulative=False)