Lab 1: Analysis of Tokenization and N-grams in Natural Language Processing\

```
import nltk
nltk.download('punkt')

from nltk.tokenize import word_tokenize
from nltk.util import ngrams
from nltk.probability import FreqDist

# Sample document
document = "Natural language processing (NLP) is a subfield of artificial intelligence (AI

# Tokenization
tokens = word_tokenize(document.lower())

# Total number of tokens
total_tokens = len(tokens)

# Number of unique tokens
unique_tokens = set(tokens)
num_unique_tokens = len(unique_tokens)

# Token frequency distribution
token_freq = FreqDist(tokens)

# Generate N-grams
n = 2  # Change to desired n-gram size
n_grams = list(ngrams(tokens, n))

# N-gram frequency distribution
n_gram_freq = FreqDist(n_grams)
```

```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
```

```
print("Total number of tokens:", total_tokens)
print("Number of unique tokens:", num_unique_tokens)
```

```
Total number of tokens: 28
Number of unique tokens: 24
```

```
print("\nToken frequency distribution:")
for token, freq in token_freq.items():
    print(f"{token}: {freq}")
```

```

Token frequency distribution:
natural: 2
language: 2
```

```
processing: 1
(: 2
nlp: 1
): 2
is: 1
a: 1
subfield: 1
of: 1
artificial: 1
intelligence: 1
ai: 1
that: 1
focuses: 1
on: 1
the: 1
interaction: 1
between: 1
computers: 1
and: 1
humans: 1
using: 1
.: 1
```

```
print("\nTop 5 Tokens:")
for token, freq in token_freq.most_common(5):
    print(f"{token}: {freq}")
```

```
Top 5 Tokens:
natural: 2
language: 2
(: 2
): 2
processing: 1
```

```
print("\nN-gram frequency distribution:")
for n_gram, freq in n_gram_freq.items():
    print(f"{' '.join(n_gram)}: {freq}")
```

```
N-gram frequency distribution:
natural language: 2
language processing: 1
processing (: 1
( nlp: 1
nlp ): 1
) is: 1
is a: 1
a subfield: 1
subfield of: 1
of artificial: 1
artificial intelligence: 1
intelligence (: 1
( ai: 1
ai ): 1
) that: 1
that focuses: 1
focuses on: 1
on the: 1
```

```
     the interaction: 1
     interaction between: 1
     between computers: 1
     computers and: 1
     and humans: 1
     humans using: 1
     using natural: 1
     language .: 1
```

```
print("\nTop 5 Bi-grams:")
for n_gram, freq in n_gram_freq.most_common(5):
    print(f"{' '.join(n_gram)}: {freq}")
```

```
     Top 5 Bi-grams:
     natural language: 2
     language processing: 1
     processing (: 1
     ( nlp: 1
     nlp ): 1
```