Read File

```
# Step 1: Read the file

file = open("sample_text.txt", "r")
text = file.read()
file.close()

print("Original Text:\n")
print(text)
```

```
Original Text:

Natural Language Processing (NLP) is a field of Artificial Intelligence.
NLP helps computers understand human language.
Language is powerful and language is complex.
Machine learning and deep learning are important in NLP.
Artificial Intelligence and machine learning are transforming the world.
```

Convert to Lowercase

```
# Step 2: Lowercase

text = text.lower()

print("After Lowercase:\n")
print(text)
```

```
After Lowercase:

natural language processing (nlp) is a field of artificial intelligence.
nlp helps computers understand human language.
language is powerful and language is complex.
machine learning and deep learning are important in nlp.
artificial intelligence and machine learning are transforming the world.
```

Split into Sentences

```
# Step 3: Split using dot

sentences = text.split(".")

print("After Splitting:")
print(sentences)
#print("Number of raw sentences:", len(sentences))
# Remove empty sentences

sentences = [s.strip() for s in sentences if s.strip() != ""]

print("After Removing Empty:")
for s in sentences:
    print(s)

print("Final sentence count:", len(sentences))
```

```
After Splitting:
['natural language processing (nlp) is a field of artificial intelligence', '\nnlp helps computers understand human language',
After Removing Empty:
natural language processing (nlp) is a field of artificial intelligence
nlp helps computers understand human language
language is powerful and language is complex
machine learning and deep learning are important in nlp
artificial intelligence and machine learning are transforming the world
Final sentence count: 5
```

Clean Each Sentence

```
# Step 4: Clean each sentence

import re

clean_sentences = []

for s in sentences:
    s = re.sub(r'[^a-z\s]', '', s)   # remove special characters
    s = re.sub(r'\s+', ' ', s).strip()

    if s != "":
        clean_sentences.append(s)

print("\nClean Sentences (Line by Line):")
for s in clean_sentences:
    print(s)

print("\nTotal Clean Sentences:", len(clean_sentences))
```

```
Clean Sentences (Line by Line):
natural language processing nlp is a field of artificial intelligence
nlp helps computers understand human language
language is powerful and language is complex
machine learning and deep learning are important in nlp
artificial intelligence and machine learning are transforming the world

Total Clean Sentences: 5
```

Create Vocabulary

```
vocab = set()

for sentence in clean_sentences:
    words = sentence.split()
    for word in words:
        vocab.add(word)

vocab = sorted(list(vocab))

print("\nVocabulary:")
print(vocab)
print("Vocabulary Size:", len(vocab))
```

```
Vocabulary:
['a', 'and', 'are', 'artificial', 'complex', 'computers', 'deep', 'field', 'helps', 'human', 'important', 'in', 'intelligence'
Vocabulary Size: 26
```

Create Word Index

```
word_index = {}

for i, word in enumerate(vocab):
    word_index[word] = i

print("\nWord Index Mapping:")
print(word_index)
```

```
Word Index Mapping:
{'a': 0, 'and': 1, 'are': 2, 'artificial': 3, 'complex': 4, 'computers': 5, 'deep': 6, 'field': 7, 'helps': 8, 'human': 9, 'impo
```

Create Sentence-wise Binary Matrix

```
binary_matrix = []

for sentence in clean_sentences:
    row = [0] * len(vocab)
```

```
    words = sentence.split()

    for word in words:
        index = word_index[word]
        row[index] = 1

    binary_matrix.append(row)

print("\nBinary Matrix:")
for row in binary_matrix:
    print(row)
```

```
Binary Matrix:
[1, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 1, 1, 1, 0, 0, 1, 1, 1, 0, 1, 0, 0, 0, 0]
[0, 0, 0, 0, 0, 1, 0, 0, 1, 1, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0]
[0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0]
[0, 1, 1, 0, 0, 0, 1, 0, 0, 0, 1, 1, 0, 0, 0, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0]
[0, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 1, 0, 0, 0, 0, 0, 1, 1, 0, 1]
```

Create Sentence-wise Frequency Matrix

```
frequency_matrix = []

for sentence in clean_sentences:
    row = [0] * len(vocab)
    words = sentence.split()

    for word in words:
        index = word_index[word]
        row[index] += 1

    frequency_matrix.append(row)

print("\nFrequency Matrix:")
for row in frequency_matrix:
    print(row)
```

```
Frequency Matrix:
[1, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 1, 1, 1, 0, 0, 1, 1, 1, 0, 1, 0, 0, 0, 0]
[0, 0, 0, 0, 0, 1, 0, 0, 1, 1, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0]
[0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 2, 2, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0]
[0, 1, 1, 0, 0, 0, 1, 0, 0, 0, 1, 1, 0, 0, 0, 2, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0]
[0, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 1, 0, 0, 0, 0, 0, 1, 1, 0, 1]
```