1.0 Question 1

This question will be using the Text Corpus data to demonstrate the manual calculation and the calculation using Python programming via Colab platform on unsmoothed unigram probabilities and smoothed unigram probabilities.

1.1 Unsmoothed Unigram Probabilities

Token	**Frequency**	**Probability**
He	2	P(He) = $\frac{Count(He)}{Total\, Token}$ $= \frac{2}{15}$ ≈ 0.1333
read	3	P(read) = $\frac{Count(read)}{Total\, Token}$ $= \frac{3}{15}$ $= \frac{1}{5}$ = 0.2
a	3	P(a) = $\frac{Count(a)}{Total\, Token}$ $= \frac{3}{15}$ $= \frac{1}{5}$ = 0.2
book	3	P(book) = $\frac{Count(book)}{Total\, Token}$ $= \frac{3}{15}$ $= \frac{1}{5}$ = 0.2
I	1	P(I) = $\frac{Count(I)}{Total\, Token}$ $= \frac{1}{15}$ ≈ 0.0666

different	1	P(different) = $\frac{Count(different)}{Total\ Token}$ $= \frac{1}{15}$ ≈ 0.0666
my	1	P(my) = $\frac{Count(my)}{Total\ Token}$ $= \frac{1}{15}$ ≈ 0.0666
Danielle	1	P(Danielle) = $\frac{Count(Danielle)}{Total\ Token}$ $= \frac{1}{15}$ ≈ 0.0666

1.2 Smoothed Unigram Probabilities

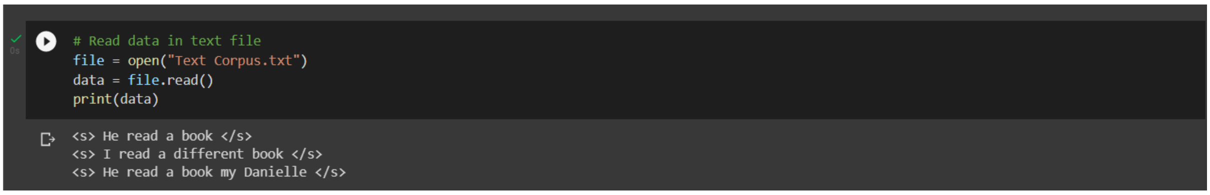
Token	**Frequency**	**Probability**
He	2	P(He) = $\frac{Count(He)+1}{Total\ Token+Unique\ Token+UNK}$ $= \frac{2+1}{15+8+1}$ $= \frac{3}{24}$ $= \frac{1}{8}$ $= 0.125$
read	3	P(read) = $\frac{Count(read)+1}{Total\ Token+Unique\ Token+UNK}$ $= \frac{3+1}{15+8+1}$ $= \frac{4}{24}$ $= \frac{1}{6}$ ≈ 0.1666
a	3	P(a) = $\frac{Count(a)+1}{Total\ Token+Unique\ Token+UNK}$ $= \frac{3+1}{15+8+1}$ $= \frac{4}{24}$

		$= \frac{1}{6}$ ≈ 0.1666
book	3	$P(\text{book}) = \frac{Count(\text{book})+1}{Total\ Token+Unique\ Token+UNK}$ $= \frac{3+1}{15+8+1}$ $= \frac{4}{24}$ $= \frac{1}{6}$ ≈ 0.1666
I	1	$P(\text{I}) = \frac{Count(I)+1}{Total\ Token+Unique\ Token+UNK}$ $= \frac{1+1}{15+8+1}$ $= \frac{2}{24}$ $= \frac{1}{12}$ ≈ 0.0833
different	1	$P(\text{different}) = \frac{Count(different)+1}{Total\ Token+Unique\ Token+UNK}$ $= \frac{1+1}{15+8+1}$ $= \frac{2}{24}$ $= \frac{1}{12}$ ≈ 0.0833
my	1	$P(\text{my}) = \frac{Count(my)+1}{Total\ Token+Unique\ Token+UNK}$ $= \frac{1+1}{15+8+1}$ $= \frac{2}{24}$ $= \frac{1}{12}$ ≈ 0.0833
Danielle	1	$P(\text{Danielle}) = \frac{Count(Danielle)+1}{Total\ Token+Unique\ Token+UNK}$ $= \frac{1+1}{15+8+1}$ $= \frac{2}{24}$

		$= \frac{1}{12}$ ≈ 0.0833

1.3 Python Implementation

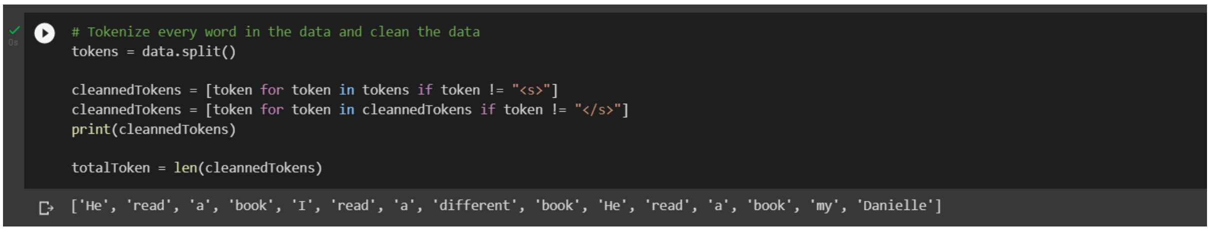
Reading Data



```
# Read data in text file
file = open("Text Corpus.txt")
data = file.read()
print(data)
```

```
<s> He read a book </s>
<s> I read a different book </s>
<s> He read a book my Danielle </s>
```

Data Cleaning



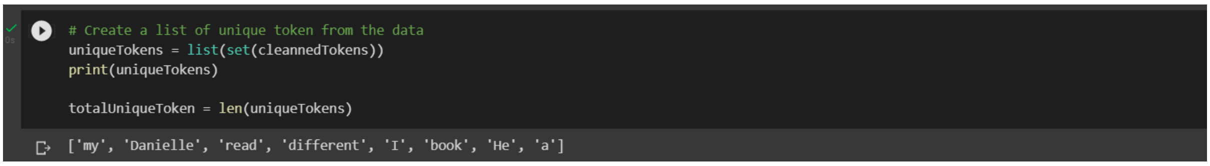
```
# Tokenize every word in the data and clean the data
tokens = data.split()

cleannedTokens = [token for token in tokens if token != "<s>"]
cleannedTokens = [token for token in cleannedTokens if token != "</s>"]
print(cleannedTokens)

totalToken = len(cleannedTokens)
```

```
['He', 'read', 'a', 'book', 'I', 'read', 'a', 'different', 'book', 'He', 'read', 'a', 'book', 'my', 'Danielle']
```

Unique Tokens

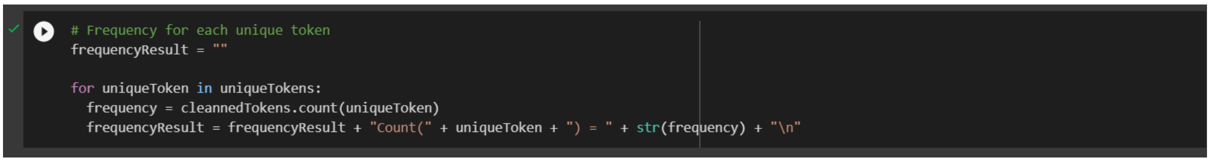


```
# Create a list of unique token from the data
uniqueTokens = list(set(cleannedTokens))
print(uniqueTokens)

totalUniqueToken = len(uniqueTokens)
```

```
['my', 'Danielle', 'read', 'different', 'I', 'book', 'He', 'a']
```

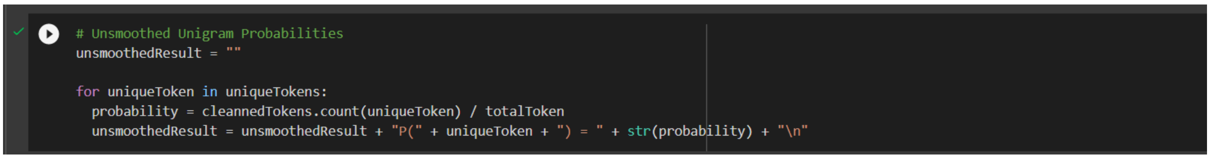
Calculating Frequency



```
# Frequency for each unique token
frequencyResult = ""

for uniqueToken in uniqueTokens:
  frequency = cleannedTokens.count(uniqueToken)
  frequencyResult = frequencyResult + "Count(" + uniqueToken + ") = " + str(frequency) + "\n"
```

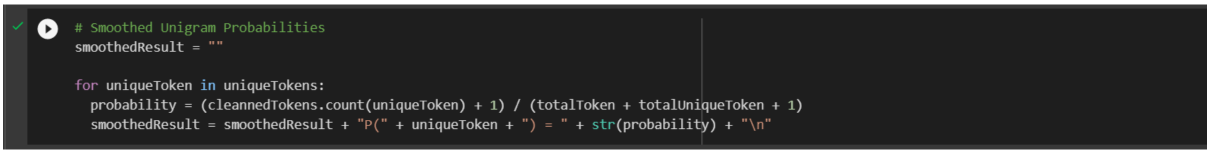
Calculating Unsmoothed Unigram Probabilities



```
# Unsmoothed Unigram Probabilities
unsmoothedResult = ""

for uniqueToken in uniqueTokens:
  probability = cleannedTokens.count(uniqueToken) / totalToken
  unsmoothedResult = unsmoothedResult + "P(" + uniqueToken + ") = " + str(probability) + "\n"
```

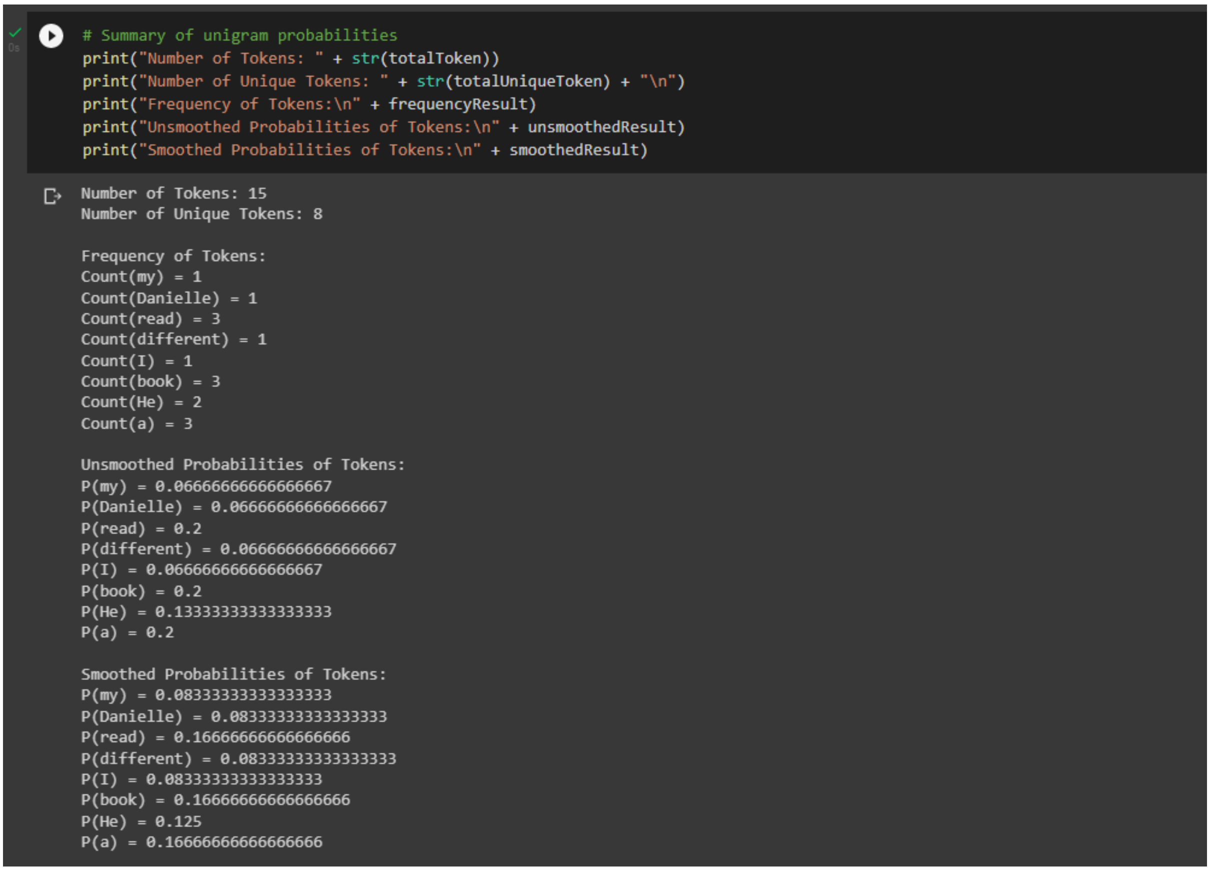
Calculating Smoothed Unigram Probabilities



```
# Smoothed Unigram Probabilities
smoothedResult = ""

for uniqueToken in uniqueTokens:
  probability = (cleannedTokens.count(uniqueToken) + 1) / (totalToken + totalUniqueToken + 1)
  smoothedResult = smoothedResult + "P(" + uniqueToken + ") = " + str(probability) + "\n"
```

Output for Summary of Unigram Probabilities



```
# Summary of unigram probabilities
print("Number of Tokens: " + str(totalToken))
print("Number of Unique Tokens: " + str(totalUniqueToken) + "\n")
print("Frequency of Tokens:\n" + frequencyResult)
print("Unsmoothed Probabilities of Tokens:\n" + unsmoothedResult)
print("Smoothed Probabilities of Tokens:\n" + smoothedResult)
```

```
Number of Tokens: 15
Number of Unique Tokens: 8

Frequency of Tokens:
Count(my) = 1
Count(Danielle) = 1
Count(read) = 3
Count(different) = 1
Count(I) = 1
Count(book) = 3
Count(He) = 2
Count(a) = 3

Unsmoothed Probabilities of Tokens:
P(my) = 0.06666666666666667
P(Danielle) = 0.06666666666666667
P(read) = 0.2
P(different) = 0.06666666666666667
P(I) = 0.06666666666666667
P(book) = 0.2
P(He) = 0.13333333333333333
P(a) = 0.2

Smoothed Probabilities of Tokens:
P(my) = 0.08333333333333333
P(Danielle) = 0.08333333333333333
P(read) = 0.16666666666666666
P(different) = 0.08333333333333333
P(I) = 0.08333333333333333
P(book) = 0.16666666666666666
P(He) = 0.125
P(a) = 0.16666666666666666
```