Lab Assignment - 7

In [19]:
```
import re

def tokenize(text):
    # Split the text into words using regular expressions
    words = re.findall(r'\b\w+\b', text.lower())
    return words

d = '''Virender Sehwag (pronunciationⓘ, born 20 October 1978) is a former Indian cricketer

# Tokenize the document
tokens = tokenize(d)
print(tokens)
```

```
['virender', 'sehwag', 'pronunciation', 'born', '20', 'october', '1978', 'is', 'a', 'forme
r', 'indian', 'cricketer', 'who', 'represented', 'india', 'from', '1999', 'to', '2013', 'w
idely', 'regarded', 'as', 'one', 'of', 'the', 'most', 'destructive', 'openers', '1', 'an
d', 'one', 'of', 'the', 'greatest', 'batsman', 'of', 'his', 'era', 'he', 'played', 'for',
'delhi', 'capitals', 'in', 'ipl', 'and', 'delhi', 'and', 'haryana', 'in', 'indian', 'domes
tic', 'cricket', 'he', 'played', 'his', 'first', 'one', 'day', 'international', 'in', '199
9', 'and', 'joined', 'the', 'indian', 'test', 'side', 'in', '2001', '2', 'in', 'april', '2
009', 'sehwag', 'became', 'the', 'first', 'indian', 'to', 'be', 'honoured', 'as', 'the',
'wisden', 'leading', 'cricketer', 'in', 'the', 'world', 'for', 'his', 'performance', 'in',
'2008', '3', 'subsequently', 'becoming', 'the', 'first', 'player', 'of', 'any', 'nationali
ty', 'to', 'retain', 'the', 'award', 'for', '2009', '4', 'he', 'worked', 'as', 'stand', 'i
n', 'captain', 'occasionally', 'during', 'absence', 'of', 'main', 'captain', 'of', 'indi
a', 'also', 'worked', 'as', 'vice', 'captain', 'for', 'indian', 'squad', 'he', 'is', 'form
er', 'captain', 'of', 'delhi', 'daredevils', 'and', 'delhi', 'ranji', 'team', 'during', 'h
is', 'time', 'with', 'india', 'sehwag', 'was', 'a', 'member', 'of', 'the', 'team', 'that',
'was', 'one', 'of', 'the', 'joint', 'winners', 'of', 'the', '2002', 'icc', 'champions', 't
rophy', 'the', 'winners', 'of', 'the', '2007', 't20', 'world', 'cup', 'and', 'the', 'winne
rs', 'of', 'the', '2011', 'cricket', 'world', 'cup', 'during', 'the', '2002', 'icc', 'cham
pions', 'trophy', 'sehwag', 'was', 'the', 'highest', 'run', 'scorer', 'with', '271', 'run
s', 'in', '2023', 'he', 'was', 'inducted', 'into', 'icc', 'cricket', 'hall', 'of', 'fame',
'5']
```

In [20]:
```
from collections import Counter

def calculate_tf(tokens):
    # Count the frequency of each word in the document
    tf = Counter(tokens)
    return tf

# Calculate term frequency for the tokens
term_frequency = calculate_tf(tokens)
print(term_frequency)
```

```
Counter({'the': 17, 'of': 13, 'in': 9, 'and': 6, 'indian': 5, 'he': 5, 'sehwag': 4, 'as':
4, 'one': 4, 'his': 4, 'for': 4, 'delhi': 4, 'captain': 4, 'was': 4, 'india': 3, 'to': 3,
'cricket': 3, 'first': 3, 'world': 3, 'during': 3, 'winners': 3, 'icc': 3, 'is': 2, 'a':
2, 'former': 2, 'cricketer': 2, '1999': 2, 'played': 2, '2009': 2, 'worked': 2, 'team': 2,
'with': 2, '2002': 2, 'champions': 2, 'trophy': 2, 'cup': 2, 'virender': 1, 'pronunciatio
n': 1, 'born': 1, '20': 1, 'october': 1, '1978': 1, 'who': 1, 'represented': 1, 'from': 1,
'2013': 1, 'widely': 1, 'regarded': 1, 'most': 1, 'destructive': 1, 'openers': 1, '1': 1,
'greatest': 1, 'batsman': 1, 'era': 1, 'capitals': 1, 'ipl': 1, 'haryana': 1, 'domestic':
1, 'day': 1, 'international': 1, 'joined': 1, 'test': 1, 'side': 1, '2001': 1, '2': 1, 'ap
ril': 1, 'became': 1, 'be': 1, 'honoured': 1, 'wisden': 1, 'leading': 1, 'performance': 1,
'2008': 1, '3': 1, 'subsequently': 1, 'becoming': 1, 'player': 1, 'any': 1, 'nationality':
1, 'retain': 1, 'award': 1, '4': 1, 'stand': 1, 'occasionally': 1, 'absence': 1, 'main':
1, 'also': 1, 'vice': 1, 'squad': 1, 'daredevils': 1, 'ranji': 1, 'time': 1, 'member': 1,
'that': 1, 'joint': 1, '2007': 1, 't20': 1, '2011': 1, 'highest': 1, 'run': 1, 'scorer':
1, '271': 1, 'runs': 1, '2023': 1, 'inducted': 1, 'into': 1, 'hall': 1, 'fame': 1, '5':
1})
```

In [21]:
```
def find_rarest_word(term_frequency):
    # Find the word with the lowest frequency
    rarest_word = min(term_frequency, key=term_frequency.get)
    return rarest_word

# Find the rarest word
rarest_word = find_rarest_word(term_frequency)
print("Rarest word:", rarest_word)
```

```
Rarest word: virender
```

In [ ]:

In [ ]:

In [25]:
```
# Split the document into sentences
sentences = re.split(r'(?<!\w\.\w.)(?<![A-Z][a-z]\.)(?<=\.|\?)\s', d)

# Tokenize the sentences and find the rarest word
min_word_count = float('inf')
rarest_word = None

for sentence in sentences:
    words = re.findall(r'\b\w+\b', sentence.lower())
    for word in words:
        if words.count(word) < min_word_count:
            min_word_count = words.count(word)
            rarest_word = word

# Print the sentences containing the rarest word
for sentence in sentences:
    if rarest_word in sentence.lower():
        print("rarest word: ",rarest_word)
        print("Sentence containing the rarest word:", sentence.strip())
```

```
rarest word:  virender
Sentence containing the rarest word: Virender Sehwag (pronunciationⓘ, born 20 October 197
8) is a former Indian cricketer who represented India from 1999 to 2013.
```

In [ ]: