In [1]:
```
### simple text convert using BOW(bag of words)
import nltk
```

In [2]:
```
paragraph = """Virat Kohli (Hindi pronunciation: [ʊɪˈɾɑːʈ ˈkoːɦli] ⓘ; born 5 November 1988) is an Indian interr

In 2013, Kohli was ranked number one in the ICC rankings for ODI batsmen. In 2015, he achieved the summit of T2(

Kohli has garnered 10 ICC Awards which is more than any player in International Cricket, making him the most de
"""
```

In [3]:
```
paragraph
```

Out[3]:
```
"Virat Kohli (Hindi pronunciation: [ʊɪˈɾɑːʈ ˈkoːɦli] ⓘ; born 5 November 1988) is an Indian international crick
eter who plays Test and One Day International (ODI) cricket for the Indian national team. A former captain in a
ll formats, Kohli retired from Twenty20 International (T20I) following India's win at the 2024 T20 World Cup. H
e is a right-handed batsman and an occasional unorthodox right arm quick bowler. He represents Royal Challenger
s Bengaluru in the Indian Premier League (IPL) and Delhi in domestic cricket. He holds the record as the highes
t run-scorer in IPL, ranks third in T20I, third in ODI, and stands as the fourth-highest in international crick
et.[5] He also holds the record for scoring the most centuries in ODI cricket and stands second in the list of
most international centuries scored. Kohli is widely regarded as one of the greatest batsmen of all time and th
e greatest batsman in the modern era.[citation needed] Kohli was a key member of the Indian team that won the 2
011 Cricket World Cup, 2013 Champions Trophy and 2024 T20 World Cup and captained India to win the ICC Test mac
e three consecutive times in 2017, 2018, and 2019.[6]\n\nIn 2013, Kohli was ranked number one in the ICC rankin
gs for ODI batsmen. In 2015, he achieved the summit of T20I rankings.[7] In 2018, he was ranked top Test batsma
n, making him the only Indian cricketer to hold the number one spot in all three formats of the game. He is the
first player to score 20,000 runs in a decade. In 2020, the International Cricket Council named him the male cr
icketer of the decade.[8]\n\nKohli has garnered 10 ICC Awards which is more than any player in International Cr
icket, making him the most decorated player in International Cricket history. He won the ICC ODI Player of the
Year award four times in 2012, 2017, 2018, and 2023. He also won the Sir Garfield Sobers Trophy, given to the I
CC Cricketer of the Year, on two occasions, in 2017 and 2018 respectively. In 2018, he became the first player
to win both ICC ODI and Test Player of the Year awards in the same year. Also, he was named the Wisden Leading
Cricketer in the World for three consecutive years, from 2016 to 2018. Kohli has the second most and most 'Play
er of the Match' and 'Player of the Series' awards to his name, respectively, in all three formats combined. At
the national level, Kohli was honoured with the Arjuna Award in 2013, the Padma Shri in 2017, and India's highe
st sporting honour, the Khel Ratna Award, in 2018. In 2018, Time magazine included him on its list of the 100 m
ost influential people in the world.\n"
```

In [4]:
```
### pre-processing steps
import nltk
from nltk.stem import PorterStemmer
from nltk.corpus import stopwords
```

In [5]:
```
### we convert  paragraph into sentance using Tokenzation

sentence = nltk.sent_tokenize(paragraph)
```

In [6]:
```
print(sentence)
```

```
['Virat Kohli (Hindi pronunciation: [ʊɪˈɾɑːʈ ˈkoːɦli] ⓘ; born 5 November 1988) is an Indian international crick
eter who plays Test and One Day International (ODI) cricket for the Indian national team.', "A former captain in
all formats, Kohli retired from Twenty20 International (T20I) following India's win at the 2024 T20 World Cup.",
'He is a right-handed batsman and an occasional unorthodox right arm quick bowler.', 'He represents Royal Challe
ngers Bengaluru in the Indian Premier League (IPL) and Delhi in domestic cricket.', 'He holds the record as the
highest run-scorer in IPL, ranks third in T20I, third in ODI, and stands as the fourth-highest in international
cricket.', '[5] He also holds the record for scoring the most centuries in ODI cricket and stands second in the
list of most international centuries scored.', 'Kohli is widely regarded as one of the greatest batsmen of all t
ime and the greatest batsman in the modern era.', '[citation needed] Kohli was a key member of the Indian team t
hat won the 2011 Cricket World Cup, 2013 Champions Trophy and 2024 T20 World Cup and captained India to win the
ICC Test mace three consecutive times in 2017, 2018, and 2019.', '[6]\n\nIn 2013, Kohli was ranked number one in
the ICC rankings for ODI batsmen.', 'In 2015, he achieved the summit of T20I rankings.', '[7] In 2018, he was ra
nked top Test batsman, making him the only Indian cricketer to hold the number one spot in all three formats of
the game.', 'He is the first player to score 20,000 runs in a decade.', 'In 2020, the International Cricket Coun
cil named him the male cricketer of the decade.', '[8]\n\nKohli has garnered 10 ICC Awards which is more than an
y player in International Cricket, making him the most decorated player in International Cricket history.', 'He
won the ICC ODI Player of the Year award four times in 2012, 2017, 2018, and 2023.', 'He also won the Sir Garfie
ld Sobers Trophy, given to the ICC Cricketer of the Year, on two occasions, in 2017 and 2018 respectively.', 'In
2018, he became the first player to win both ICC ODI and Test Player of the Year awards in the same year.', 'Als
o, he was named the Wisden Leading Cricketer in the World for three consecutive years, from 2016 to 2018.', "Koh
li has the second most and most 'Player of the Match' and 'Player of the Series' awards to his name, respectivel
y, in all three formats combined.", "At the national level, Kohli was honoured with the Arjuna Award in 2013, th
e Padma Shri in 2017, and India's highest sporting honour, the Khel Ratna Award, in 2018.", 'In 2018, Time magaz
ine included him on its list of the 100 most influential people in the world.']
```

In [7]:
```
len(sentence)
```

Out[7]:
```
21
```

In [9]:
```
### we romve umanted words form the sentance

import re
corpus = []
```

```
for i in range(len(sentence)):
    review = re.sub('^a-Az-Z' ,'' , sentence[i])
    review = review.lower()
    corpus.append(review)
```

In [10]:
```
corpus
```

Out[10]:
```
['virat kohli (hindi pronunciation: [ʋɪˈɾɑːʈ ˈkoːɦli] ⓘ; born 5 november 1988) is an indian international cric
keter who plays test and one day international (odi) cricket for the indian national team.',
 "a former captain in all formats, kohli retired from twenty20 international (t20i) following india's win at th
e 2024 t20 world cup.",
 'he is a right-handed batsman and an occasional unorthodox right arm quick bowler.',
 'he represents royal challengers bengaluru in the indian premier league (ipl) and delhi in domestic cricket.',
 'he holds the record as the highest run-scorer in ipl, ranks third in t20i, third in odi, and stands as the fo
urth-highest in international cricket.',
 '[5] he also holds the record for scoring the most centuries in odi cricket and stands second in the list of m
ost international centuries scored.',
 'kohli is widely regarded as one of the greatest batsmen of all time and the greatest batsman in the modern er
a.',
 '[citation needed] kohli was a key member of the indian team that won the 2011 cricket world cup, 2013 champio
ns trophy and 2024 t20 world cup and captained india to win the icc test mace three consecutive times in 2017,
2018, and 2019.',
 '[6]\n\nin 2013, kohli was ranked number one in the icc rankings for odi batsmen.',
 'in 2015, he achieved the summit of t20i rankings.',
 '[7] in 2018, he was ranked top test batsman, making him the only indian cricketer to hold the number one spot
in all three formats of the game.',
 'he is the first player to score 20,000 runs in a decade.',
 'in 2020, the international cricket council named him the male cricketer of the decade.',
 '[8]\n\nkohli has garnered 10 icc awards which is more than any player in international cricket, making him th
e most decorated player in international cricket history.',
 'he won the icc odi player of the year award four times in 2012, 2017, 2018, and 2023.',
 'he also won the sir garfield sobers trophy, given to the icc cricketer of the year, on two occasions, in 2017
and 2018 respectively.',
 'in 2018, he became the first player to win both icc odi and test player of the year awards in the same year.'
,
 'also, he was named the wisden leading cricketer in the world for three consecutive years, from 2016 to 2018.'
,
 "kohli has the second most and most 'player of the match' and 'player of the series' awards to his name, respe
ctively, in all three formats combined.",
 "at the national level, kohli was honoured with the arjuna award in 2013, the padma shri in 2017, and india's
highest sporting honour, the khel ratna award, in 2018.",
 'in 2018, time magazine included him on its list of the 100 most influential people in the world.']
```

In [15]:
```
### stemming

stemmer = PorterStemmer()

for i in corpus:
    words =nltk.word_tokenize(i)
    for word in words :
        if word not in set(stopwords.words('english')):
            print(stemmer.stem(word))
```

```
virat
kohli
(
hindi
pronunci
:
[
ʋɪˈɾɑːʈ
ˈkoːɦli
]
ⓘ
;
born
5
novemb
1988
)
indian
intern
cricket
play
test
one
day
intern
(
odi
)
```

cricket
indian
nation
team
.
former
captain
format
,
kohli
retir
twenty20
intern
(
t20i
)
follow
india
's
win
2024
t20
world
cup
.
right-hand
batsman
occasion
unorthodox
right
arm
quick
bowler
.
repres
royal
challeng
bengaluru
indian
premier
leagu
(
ipl
)
delhi
domest
cricket
.
hold
record
highest
run-scor
ipl
,
rank
third
t20i
,
third
odi
,
stand
fourth-highest
intern
cricket
.
[
5
]
also
hold
record
score
centuri
odi
cricket
stand
second
list
intern
centuri
score
.

kohli
wide
regard
one
greatest
batsmen
time
greatest
batsman
modern
era
.
[
citat
need
]
kohli
key
member
indian
team
2011
cricket
world
cup
,
2013
champion
trophi
2024
t20
world
cup
captain
india
win
icc
test
mace
three
consecut
time
2017
,
2018
,
2019
.
[
6
]
2013
,
kohli
rank
number
one
icc
rank
odi
batsmen
.
2015
,
achiev
summit
t20i
rank
.
[
7
]
2018
,
rank
top
test
batsman
,
make
indian
cricket
hold

number
one
spot
three
format
game
.
first
player
score
20,000
run
decad
.
2020
,
intern
cricket
council
name
male
cricket
decad
.
[
8
]
kohli
garner
10
icc
award
player
intern
cricket
,
make
decor
player
intern
cricket
histori
.
icc
odi
player
year
award
four
time
2012
,
2017
,
2018
,
2023
.
also
sir
garfield
sober
trophi
,
given
icc
cricket
year
,
two
occas
,
2017
2018
respect
.
2018
,
becam
first
player
win
icc

```
odi
test
player
year
award
year
.
also
,
name
wisden
lead
cricket
world
three
consecut
year
,
2016
2018
.
kohli
second
'player
match
'
'player
seri
'
award
name
,
respect
,
three
format
combin
.
nation
level
,
kohli
honour
arjuna
award
2013
,
padma
shri
2017
,
india
's
highest
sport
honour
,
khel
ratna
award
,
2018
.
2018
,
time
magazin
includ
list
100
influenti
peopl
world
.
```

In [27]:
```
### covert the word into vector
### using BOW (bag of words)
from sklearn.feature_extraction.text import CountVectorizer

vc = CountVectorizer(binary=True)
```

In [28]:
```
x = vc.fit_transform(corpus)
```

In [29]:
```
x
```

Out[29]:
```
<21x182 sparse matrix of type '<class 'numpy.int64'>'
        with 389 stored elements in Compressed Sparse Row format>
```

In [30]:
```
### to check unipue value
vc.vocabulary_
```

Out[30]:
```
{'virat': 168,
 'kohli': 90,
 'hindi': 71,
 'pronunciation': 122,
 'ʋɪˈɾɑːʈ': 180,
 "'koːɦli": 181,
 'born': 32,
 'november': 108,
 '1988': 3,
 'is': 86,
 'an': 19,
 'indian': 82,
 'international': 84,
 'cricketer': 45,
 'who': 171,
 'plays': 120,
 'test': 154,
 'and': 20,
 'one': 115,
 'day': 47,
 'odi': 112,
 'cricket': 44,
 'for': 55,
 'the': 157,
 'national': 106,
 'team': 153,
 'former': 57,
 'captain': 35,
 'in': 79,
 'all': 17,
 'formats': 56,
 'retired': 132,
 'from': 60,
 'twenty20': 165,
 't20i': 152,
 'following': 54,
 'india': 81,
 'win': 173,
 'at': 25,
 '2024': 15,
 't20': 151,
 'world': 177,
 'cup': 46,
 'he': 68,
 'right': 133,
 'handed': 66,
 'batsman': 28,
 'occasional': 110,
 'unorthodox': 167,
 'arm': 23,
 'quick': 123,
 'bowler': 34,
 'represents': 130,
 'royal': 134,
 'challengers': 38,
 'bengaluru': 31,
 'premier': 121,
 'league': 92,
 'ipl': 85,
 'delhi': 50,
 'domestic': 51,
 'holds': 75,
 'record': 128,
 'as': 24,
 'highest': 69,
 'run': 135,
 'scorer': 140,
 'ranks': 126,
 'third': 158,
 'stands': 149,
 'fourth': 59,
 'also': 18,
 'scoring': 141,
 'most': 103,
 'centuries': 37,
```

```
'second': 142,
'list': 94,
'of': 113,
'scored': 139,
'widely': 172,
'regarded': 129,
'greatest': 65,
'batsmen': 29,
'time': 160,
'modern': 101,
'era': 52,
'citation': 40,
'needed': 107,
'was': 169,
'key': 88,
'member': 100,
'that': 156,
'won': 176,
'2011': 5,
'2013': 7,
'champions': 39,
'trophy': 164,
'captained': 36,
'to': 162,
'icc': 78,
'mace': 95,
'three': 159,
'consecutive': 42,
'times': 161,
'2017': 10,
'2018': 11,
'2019': 12,
'ranked': 124,
'number': 109,
'rankings': 125,
'2015': 8,
'achieved': 16,
'summit': 150,
'top': 163,
'making': 97,
'him': 70,
'only': 116,
'hold': 74,
'spot': 148,
'game': 61,
'first': 53,
'player': 119,
'score': 138,
'20': 4,
'000': 0,
'runs': 136,
'decade': 48,
'2020': 13,
'council': 43,
'named': 105,
'male': 98,
'has': 67,
'garnered': 63,
'10': 1,
'awards': 27,
'which': 170,
'more': 102,
'than': 155,
'any': 21,
'decorated': 49,
'history': 73,
'year': 178,
'award': 26,
'four': 58,
'2012': 6,
'2023': 14,
'sir': 145,
'garfield': 62,
'sobers': 146,
'given': 64,
'on': 114,
'two': 166,
'occasions': 111,
'respectively': 131,
'became': 30,
'both': 33,
'same': 137,
'wisden': 174,
```

```
 'leading': 91,
 'years': 179,
 '2016': 9,
 'match': 99,
 'series': 143,
 'his': 72,
 'name': 104,
 'combined': 41,
 'level': 93,
 'honoured': 77,
 'with': 175,
 'arjuna': 22,
 'padma': 117,
 'shri': 144,
 'sporting': 147,
 'honour': 76,
 'khel': 89,
 'ratna': 127,
 'magazine': 96,
 'included': 80,
 'its': 87,
 '100': 2,
 'influential': 83,
 'people': 118}
```

In [31]:
```
### to check words in index  0

corpus[0]
```

Out[31]:
```
'virat kohli (hindi pronunciation: [ʋɪˈɾɑːʈ ˈkoːɦli] ⓘ; born 5 november 1988) is an indian international cricketer who plays test and one day international (odi) cricket for the indian national team.'
```

In [32]:
```
x[0].toarray()
```

Out[32]:
```
array([[0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0,
        0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
        1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
        0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 1, 0,
        0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0,
        0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0,
        0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1,
        1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0,
        0, 0, 0, 0, 1, 1]], dtype=int64)
```

In [33]:
```
corpus[1]
```

Out[33]:
```
"a former captain in all formats, kohli retired from twenty20 international (t20i) following india's win at the 2024 t20 world cup."
```

In [34]:
```
x[1].toarray()
```

Out[34]:
```
array([[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0,
        0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0,
        0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 1, 0, 0, 1, 0, 0, 0, 0, 0,
        0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 1, 0, 0, 0,
        0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
        0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
        1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0,
        0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0,
        0, 1, 0, 0, 0, 0]], dtype=int64)
```

In [ ]:
```

```