# Karanjot Singh

In [114]:

```
import string
string.punctuation
text  = """An additional 5,974 infections brought to 92,472, the number of people who h
ave officially tested positive for COVID-19 in Italy since the crisis began last month.
            Italy’s death toll from COVID-19 shot past 10,000 on Saturday with 889 new
 deaths, the country’s Civil Protection Service said.The new fatalities reported in the
world's worst-hit nation came a day after it registered 969 deaths on Friday — the high
est single toll since the COVID-19 virus emerged late last year.
            Italy now looks certain to extend its economically debilitating — and emoti
onally stressful — business closures and the ban on public gatherings past their April
 3 deadline.“Is it time to reopen the country? I think we have to think about it really
carefully,” civil protection service chief Angelo Borrelli told reporters. “The country
is at a standstill and we must maintain the least amount of activity possible to ensure
the survival of all.”
            Italians had begun to hope that their worst disaster in generations was eas
ing after the increase in daily death rates began to slow on March 22.
            But the new surge has changed the Mediterranean nation's mood.     """

punch = [c for c in text if c not in string.punctuation]
punch = ''.join(punch)
print(punch)
```

```
An additional 5974 infections brought to 92472 the number of people who
have officially tested positive for COVID19 in Italy since the crisis be
gan last month
            Italy’s death toll from COVID19 shot past 10000 on Saturday
with 889 new deaths the country’s Civil Protection Service saidThe new f
atalities reported in the worlds worsthit nation came a day after it reg
istered 969 deaths on Friday — the highest single toll since the COVID19
virus emerged late last year
            Italy now looks certain to extend its economically debilitat
ing — and emotionally stressful — business closures and the ban on publi
c gatherings past their April 3 deadline“Is it time to reopen the countr
y I think we have to think about it really carefully” civil protection s
ervice chief Angelo Borrelli told reporters “The country is at a standst
ill and we must maintain the least amount of activity possible to ensure
the survival of all”
            Italians had begun to hope that their worst disaster in gene
rations was easing after the increase in daily death rates began to slow
on March 22
            But the new surge has changed the Mediterranean nations mood
```

In [127]:

```
punch = punch.lower()
print(punch)
```

```
an additional 5974 infections brought to 92472 the number of people who ha
ve officially tested positive for covid19 in italy since the crisis began
last month
            italy’s death toll from covid19 shot past 10000 on saturday wi
th 889 new deaths the country’s civil protection service saidthe new fatal
ities reported in the worlds worsthit nation came a day after it registere
d 969 deaths on friday – the highest single toll since the covid19 virus e
merged late last year
            italy now looks certain to extend its economically debilitatin
g – and emotionally stressful – business closures and the ban on public ga
therings past their april 3 deadline“is it time to reopen the country i th
ink we have to think about it really carefully” civil protection service c
hief angelo borrelli told reporters “the country is at a standstill and we
must maintain the least amount of activity possible to ensure the survival
of all”
            italians had begun to hope that their worst disaster in genera
tions was easing after the increase in daily death rates began to slow on
march 22
            but the new surge has changed the mediterranean nations mood
```

In [128]:

```
import nltk
sentences  = nltk.sent_tokenize(punch)
print(sentences)
```

```
['an additional 5974 infections brought to 92472 the number of people who
have officially tested positive for covid19 in italy since the crisis bega
n last month\n            italy’s death toll from covid19 shot past 10000
on saturday with 889 new deaths the country’s civil protection service sai
dthe new fatalities reported in the worlds worsthit nation came a day afte
r it registered 969 deaths on friday \u2060– the highest single toll since
the covid19 virus emerged late last year\n            italy now looks cert
ain to extend its economically debilitating \u2060– and emotionally stress
ful \u2060– business closures and the ban on public gatherings past their
april 3 deadline“is it time to reopen the country i think we have to think
about it really carefully” civil protection service chief angelo borrelli
told reporters “the country is at a standstill and we must maintain the le
ast amount of activity possible to ensure the survival of all”\n
italians had begun to hope that their worst disaster in generations was ea
sing after the increase in daily death rates began to slow on march 22\n
but the new surge has changed the mediterranean nations mood']
```

In [129]:

```
word = nltk.word_tokenize(punch)
print(word)
```

```
['an', 'additional', '5974', 'infections', 'brought', 'to', '92472', 'th
e', 'number', 'of', 'people', 'who', 'have', 'officially', 'tested', 'posi
tive', 'for', 'covid19', 'in', 'italy', 'since', 'the', 'crisis', 'began',
'last', 'month', 'italy', '’', 's', 'death', 'toll', 'from', 'covid19', 's
hot', 'past', '10000', 'on', 'saturday', 'with', '889', 'new', 'deaths',
'the', 'country', '’', 's', 'civil', 'protection', 'service', 'saidthe',
'new', 'fatalities', 'reported', 'in', 'the', 'worlds', 'worsthit', 'natio
n', 'came', 'a', 'day', 'after', 'it', 'registered', '969', 'deaths', 'o
n', 'friday', '\u2060–', 'the', 'highest', 'single', 'toll', 'since', 'th
e', 'covid19', 'virus', 'emerged', 'late', 'last', 'year', 'italy', 'now',
'looks', 'certain', 'to', 'extend', 'its', 'economically', 'debilitating',
'\u2060–', 'and', 'emotionally', 'stressful', '\u2060–', 'business', 'clos
ures', 'and', 'the', 'ban', 'on', 'public', 'gatherings', 'past', 'their',
'april', '3', 'deadline', '“', 'is', 'it', 'time', 'to', 'reopen', 'the',
'country', 'i', 'think', 'we', 'have', 'to', 'think', 'about', 'it', 'real
ly', 'carefully', '”', 'civil', 'protection', 'service', 'chief', 'angel
o', 'borrelli', 'told', 'reporters', '“', 'the', 'country', 'is', 'at',
'a', 'standstill', 'and', 'we', 'must', 'maintain', 'the', 'least', 'amoun
t', 'of', 'activity', 'possible', 'to', 'ensure', 'the', 'survival', 'of',
'all', '”', 'italians', 'had', 'begun', 'to', 'hope', 'that', 'their', 'wo
rst', 'disaster', 'in', 'generations', 'was', 'easing', 'after', 'the', 'i
ncrease', 'in', 'daily', 'death', 'rates', 'began', 'to', 'slow', 'on', 'm
arch', '22', 'but', 'the', 'new', 'surge', 'has', 'changed', 'the', 'medit
erranean', 'nations', 'mood']
```

In [130]:

```
import nltk
from nltk.stem import PorterStemmer
from nltk.corpus import stopwords

sentences = nltk.sent_tokenize(punch)
stemmer = PorterStemmer()

for i in range(len(sentences)):
    words = nltk.word_tokenize(sentences[i])
    words = [stemmer.stem(word) for word in words if word not in set(stopwords.words('e
nglish'))]
    sentences[i] = ' '.join(words)
print(sentences)
```

```
['addit 5974 infect brought 92472 number peopl offici test posit covid19 i
tali sinc crisi began last month itali ’ death toll covid19 shot past 1000
0 saturday 889 new death countri ’ civil protect servic saidth new fatal r
eport world worsthit nation came day regist 969 death friday \u2060– highe
st singl toll sinc covid19 viru emerg late last year itali look certain ex
tend econom debilit \u2060– emot stress \u2060– busi closur ban public gat
her past april 3 deadlin “ time reopen countri think think realli care ” c
ivil protect servic chief angelo borrelli told report “ countri standstil
must maintain least amount activ possibl ensur surviv ” italian begun hope
worst disast gener eas increas daili death rate began slow march 22 new su
rg chang mediterranean nation mood']
```

In [131]:

```
import nltk
from nltk.corpus import stopwords

sentences = nltk.sent_tokenize(text)

for i in range(len(sentences)):
    words = nltk.word_tokenize(sentences[i])
    words = [word for word in words if word not in set(stopwords.words('english'))]
    sentences[i] = ' '.join(words)
print(sentences)
```

```
['An additional 5,974 infections brought 92,472 , number people officially
tested positive COVID-19 Italy since crisis began last month .', "Italy ’
death toll COVID-19 shot past 10,000 Saturday 889 new deaths , country ’ C
ivil Protection Service said.The new fatalities reported world 's worst-hi
t nation came day registered 969 deaths Friday \u2060— highest single toll
since COVID-19 virus emerged late last year .", 'Italy looks certain exten
d economically debilitating \u2060— emotionally stressful \u2060— business
closures ban public gatherings past April 3 deadline. “ Is time reopen cou
ntry ?', 'I think think really carefully , ” civil protection service chie
f Angelo Borrelli told reporters .', '“ The country standstill must mainta
in least amount activity possible ensure survival all. ” Italians begun ho
pe worst disaster generations easing increase daily death rates began slow
March 22 .', "But new surge changed Mediterranean nation 's mood ."]
```

In [120]:

```
import nltk
from nltk.stem import WordNetLemmatizer
from nltk.corpus import stopwords

sentences = nltk.sent_tokenize(text)
lemmatizer = WordNetLemmatizer()

for i in range(len(sentences)):
    words = nltk.word_tokenize(sentences[i])
    words = [lemmatizer.lemmatize(word) for word in words if word not in set(stopwords.
words('english'))]
    sentences[i] = ' '.join(words)
print(sentences)
```

```
['An additional 5,974 infection brought 92,472 , number people officially
tested positive COVID-19 Italy since crisis began last month .', "Italy ’
death toll COVID-19 shot past 10,000 Saturday 889 new death , country ’ Ci
vil Protection Service said.The new fatality reported world 's worst-hit n
ation came day registered 969 death Friday \u2060— highest single toll sin
ce COVID-19 virus emerged late last year .", 'Italy look certain extend ec
onomically debilitating \u2060— emotionally stressful \u2060— business clo
sure ban public gathering past April 3 deadline. “ Is time reopen country
?', 'I think think really carefully , ” civil protection service chief Ang
elo Borrelli told reporter .', '“ The country standstill must maintain lea
st amount activity possible ensure survival all. ” Italians begun hope wor
st disaster generation easing increase daily death rate began slow March 2
2 .', "But new surge changed Mediterranean nation 's mood ."]
```

In [121]:

```
import nltk
import re
from nltk.corpus import stopwords
from nltk.stem import PorterStemmer
from nltk.stem import WordNetLemmatizer

file = open("data.txt", encoding="utf8")
data = file.read()


ps = PorterStemmer()
wordnet = WordNetLemmatizer()
sentences = nltk.sent_tokenize(data)
corpu = []
for i in range(len(sentences)):
    review = re.sub('[^a-zA-Z]', ' ',sentences[i])
    review = review.lower()
    review = nltk.word_tokenize(review)
    review = [ps.stem(word) for word in review if not word in set(stopwords.words('engl
ish'))]
    review = ' '.join(review)
    corpu.append(review)

from sklearn.feature_extraction.text import CountVectorizer
cv = CountVectorizer(max_features=70)
X = cv.fit_transform(corpu).toarray()
print(X)
```

```
[[0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 1]
 [0 0 0 ... 1 0 0]
 ...
 [0 0 0 ... 1 0 0]
 [0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]]
```

In [122]:

```
import nltk
import pandas as pd
file = open("data.txt", encoding="utf8")
data = file.read()

from sklearn.feature_extraction.text import TfidfVectorizer
cv = TfidfVectorizer()
sentences = nltk.sent_tokenize(data)
x = cv.fit_transform(sentences)
df = pd.DataFrame(x.toarray(),columns= cv.get_feature_names())
df
```

Out[122]:

| | 10 | 13 | 16 | 1978 | 20 | 2009 | 2016 | 2018 | 2019 | 29 | ... | world | woul |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0.0 | 0.168079 | 0.154869 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.168079 | 0.0 | ... | 0.0 | 0.00000 |
| 1 | 0.0 | 0.000000 | 0.000000 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.000000 | 0.0 | ... | 0.0 | 0.00000 |
| 2 | 0.0 | 0.000000 | 0.000000 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.000000 | 0.0 | ... | 0.0 | 0.00000 |
| 3 | 0.0 | 0.000000 | 0.000000 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.000000 | 0.0 | ... | 0.0 | 0.00000 |
| 4 | 0.0 | 0.000000 | 0.000000 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.000000 | 0.0 | ... | 0.0 | 0.00000 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | . |
| 122 | 0.0 | 0.000000 | 0.000000 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.000000 | 0.0 | ... | 0.0 | 0.27894 |
| 123 | 0.0 | 0.000000 | 0.142752 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.000000 | 0.0 | ... | 0.0 | 0.00000 |
| 124 | 0.0 | 0.000000 | 0.000000 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.000000 | 0.0 | ... | 0.0 | 0.00000 |
| 125 | 0.0 | 0.000000 | 0.000000 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.000000 | 0.0 | ... | 0.0 | 0.00000 |
| 126 | 0.0 | 0.000000 | 0.000000 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.000000 | 0.0 | ... | 0.0 | 0.00000 |

127 rows × 1031 columns

In [123]:

```
from gensim.models import Word2Vec
import nltk
import pandas as pd
file = open("data.txt", encoding="utf8")
data = file.read()
review = re.sub('[^a-zA-Z]',' ', data)
sentences = nltk.sent_tokenize(review)
sentences = [nltk.word_tokenize(sentence) for sentence in sentences]
for i in range(len(sentences)):
    sentences[i] = [word for word in sentences[i] if word not in stopwords.words('english')]

model = Word2Vec(sentences,min_count=1)
similar = model.wv.most_similar('language')
print(similar[0])
```

```
('spotting', 0.2968634366989136)
```