Sarvagya Singh

- 60009200030 -- K1
- Computer Linguistics -lab1 -- NLTK

In [ ]:
```
import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
import string as st
import re
import nltk
from nltk import PorterStemmer, WordNetLemmatizer, LancasterStemmer
import nltk.corpus
import os
import string
from nltk.corpus import stopwords
```

In [ ]:
```
%%capture
nltk.download('all')
```

In [ ]:
```
text = "Random is a term used in mathematics 464 (and less formally) ' . - ; to mean that there is no way to reliabl
y predict an outcome (to know what will happen before it happens) or sense a pattern. Something that is chosen at ra
ndom is not chosen for any conscious reason, and therefore thought to be purely by chance. An example of a random ev
ent is winning a lottery."
len(text)
```

Out[ ]:
```
365
```

In [ ]:
```
from nltk.tokenize import word_tokenize
from nltk.tokenize import sent_tokenize
tokenized_words = word_tokenize(text)
len(tokenized_words)
```

Out[ ]:
```
78
```

In [ ]:
```
sent_tokenized_words = sent_tokenize(text)
```

In [ ]:
```
#Text Lower Case
def text_lowercase(text):
  return text.lower()

text_lowercase(text)
```

Out[ ]:
```
"random is a term used in mathematics 464 (and less formally) ' . - ; to mean that there is no way to reliably predic
t an outcome (to know what will happen before it happens) or sense a pattern. something that is chosen at random is n
ot chosen for any conscious reason, and therefore thought to be purely by chance. an example of a random event is win
ning a lottery."
```

In [ ]:
```
def remove_numbers(text):
  result = re.sub(r'\d+','',text)
  return result

remove_numbers(text)
```

Out[ ]:
```
"Random is a term used in mathematics  (and less formally) ' . - ; to mean that there is no way to reliably predict a
n outcome (to know what will happen before it happens) or sense a pattern. Something that is chosen at random is not
 chosen for any conscious reason, and therefore thought to be purely by chance. An example of a random event is winnin
g a lottery."
```

In [ ]:
```
def remove_punctuation(text):
  translator = str.maketrans('', '', string.punctuation)
  return text.translate(translator)

remove_punctuation(text)
```

Out[ ]:
```
'Random is a term used in mathematics 464 and less formally     to mean that there is no way to reliably predict an o
utcome to know what will happen before it happens or sense a pattern Something that is chosen at random is not chosen
 for any conscious reason and therefore thought to be purely by chance An example of a random event is winning a lotte
ry'
```

In [ ]:
```
def remove_whitespace(text):
  return " ".join(text.split())

remove_whitespace(text)
```

Out[ ]:
```
"Random is a term used in mathematics 464 (and less formally) ' . - ; to mean that there is no way to reliably predic
t an outcome (to know what will happen before it happens) or sense a pattern. Something that is chosen at random is n
ot chosen for any conscious reason, and therefore thought to be purely by chance. An example of a random event is win
ning a lottery."
```

In [ ]:
```
def remove_stopwords(text):
  stop_words = set(stopwords.words("english"))
  # print(len(stop_words))
  word_tokens = word_tokenize(text)
  filtered_text = [word for word in word_tokens if word not in stop_words]
  return filtered_text

print(len(remove_stopwords(text)))
```

```
45
```

In [ ]:
```
tokens = [t for t in text.split()]
sr= stopwords.words('english')
clean_tokens = tokens[:]
for token in tokens:
  if token in stopwords.words('english'):
    clean_tokens.remove(token)
    freq = nltk.FreqDist(clean_tokens)

print(freq)
```

```
<FreqDist with 37 samples and 39 outcomes>
```

In [ ]:
```
tokens = word_tokenize(text)
porter = PorterStemmer()
lancaster = LancasterStemmer()
def stemming_types(stemmer,text):
  '''
    text: The text is the list which is tokenized, stemmed
  '''
  return stemmer.stem(text)
```

In [ ]:
```
t1 = text_lowercase(text)
t2 = remove_numbers(t1)
t3 = remove_punctuation(t2)
t4 = remove_whitespace(t3)
t5 = remove_stopwords(t4)
stemming_types(porter,text)
```

Out[ ]:
```
"random is a term used in mathematics 464 (and less formally) ' . - ; to mean that there is no way to reliably predic
t an outcome (to know what will happen before it happens) or sense a pattern. something that is chosen at random is n
ot chosen for any conscious reason, and therefore thought to be purely by chance. an example of a random event is win
ning a lottery."
```

In [ ]:
```
stemming_types(lancaster,text)
```

Out[ ]:
```
"random is a term used in mathematics 464 (and less formally) ' . - ; to mean that there is no way to reliably predic
t an outcome (to know what will happen before it happens) or sense a pattern. something that is chosen at random is n
ot chosen for any conscious reason, and therefore thought to be purely by chance. an example of a random event is win
ning a lottery."
```

In [ ]:
```
lemmatizer = WordNetLemmatizer()
# lemmatize string
def lemmatize_word(text):
  word_tokens = word_tokenize(text)
  # provide context i.e. part-of-speech
  lemmas = [lemmatizer.lemmatize(word, pos ='v') for word in word_tokens]
  return lemmas
# text = 'data science uses scientific methods algorithms and many types of processes'
print(lemmatize_word(text))
```

```
['Random', 'be', 'a', 'term', 'use', 'in', 'mathematics', '464', '(', 'and', 'less', 'formally', ')', "'", '.', '-',
';', 'to', 'mean', 'that', 'there', 'be', 'no', 'way', 'to', 'reliably', 'predict', 'an', 'outcome', '(', 'to', 'kno
w', 'what', 'will', 'happen', 'before', 'it', 'happen', ')', 'or', 'sense', 'a', 'pattern', '.', 'Something', 'that',
'be', 'choose', 'at', 'random', 'be', 'not', 'choose', 'for', 'any', 'conscious', 'reason', ',', 'and', 'therefore',
'think', 'to', 'be', 'purely', 'by', 'chance', '.', 'An', 'example', 'of', 'a', 'random', 'event', 'be', 'win', 'a',
'lottery', '.']
```