Understanding Text preprocessing

Tokenization Basic

In [1]:
```
import nltk
nltk.download("wordnet")
nltk.download("punkt")
nltk.download('stopwords')
nltk.download('punkt_tab')
```

```
[nltk_data] Downloading package wordnet to /Users/raman/nltk_data...
[nltk_data]   Package wordnet is already up-to-date!
[nltk_data] Downloading package punkt to /Users/raman/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
[nltk_data] Downloading package stopwords to /Users/raman/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
[nltk_data] Downloading package punkt_tab to /Users/raman/nltk_data...
[nltk_data]   Unzipping tokenizers/punkt_tab.zip.
```

Out[1]: True

In [2]:
```
corpus = """Hello welcome to alwaysoversleeping desktop.
please do watch my body as it fragiule.
you can clean using fibre cloth and keyboard cleaner."""
```

In [3]:
```
print(corpus)
```

```
Hello welcome to alwaysoversleeping desktop.
please do watch my body as it fragiule.
you can clean using fibre cloth and keyboard cleaner.
```

In [4]:
```
# Tokenization -> sentence
from nltk.tokenize import sent_tokenize
```

In [5]:
```
document = sent_tokenize(corpus)
document
```

Out[5]:
```
['Hello welcome to alwaysoversleeping desktop.',
 'please do watch my body as it fragiule.',
 'you can clean using fibre cloth and keyboard cleaner.']
```

In [6]:
```
type(document)
```

Out[6]: list

In [7]:
```
for sentence in document:
    print(sentence)
```

```
Hello welcome to alwaysoversleeping desktop.
please do watch my body as it fragiule.
you can clean using fibre cloth and keyboard cleaner.
```

In [8]:
```
# Tokenization -> word tokens
from nltk.tokenize import word_tokenize
```

In [9]:
```
doc = word_tokenize(corpus)
print(doc)
len(doc)
```

```
['Hello', 'welcome', 'to', 'alwaysoversleeping', 'desktop', '.', 'please',
'do', 'watch', 'my', 'body', 'as', 'it', 'fragiule', '.', 'you', 'can', 'c
lean', 'using', 'fibre', 'cloth', 'and', 'keyboard', 'cleaner', '.']
```

Out[9]: 25

In [10]:
```
from nltk.tokenize import wordpunct_tokenize
```

In [11]:
```
doc1 = wordpunct_tokenize(corpus)
print(doc1)
len(doc)
```

```
['Hello', 'welcome', 'to', 'alwaysoversleeping', 'desktop', '.', 'please',
'do', 'watch', 'my', 'body', 'as', 'it', 'fragiule', '.', 'you', 'can', 'c
lean', 'using', 'fibre', 'cloth', 'and', 'keyboard', 'cleaner', '.']
```

Out[11]: 25

In [12]:
```
from nltk.tokenize import TreebankWordTokenizer  # it place . , etc with
```

In [13]:
```
tokenizer = TreebankWordTokenizer().tokenize(corpus)
print(tokenizer)
len(tokenizer)
```

```
['Hello', 'welcome', 'to', 'alwaysoversleeping', 'desktop.', 'please', 'd
o', 'watch', 'my', 'body', 'as', 'it', 'fragiule.', 'you', 'can', 'clean',
'using', 'fibre', 'cloth', 'and', 'keyboard', 'cleaner', '.']
```

Out[13]: 23

Stemning and It'type

In [14]:
```
words = ['eating','east','eaten','writing','writes','programming','progra
```

PorterStemmer

In [15]:
```
from nltk.stem import PorterStemmer
```

In [16]:
```
stemming = PorterStemmer()
for word in words:
    print(f'{word} ----> {stemming.stem(word)}')
# It may change measning of words
```

```
print(f'congratulation----->{stemming.stem("congratulation")}')
```

```
eating ----> eat
east ----> east
eaten ----> eaten
writing ----> write
writes ----> write
programming ----> program
programs ----> program
finally ----> final
finalize ----> final
history ----> histori
congratulation----->congratul
```

RegexpStemmer

In [17]:
```
from nltk.stem import RegexpStemmer
```

In [18]:
```
reg_exp = RegexpStemmer('ing$|s$|e$|able$', min = 4)
```

In [19]:
```
words1= ['eating', 'unable', 'understanding','readable', 'plants']
```

In [20]:
```
for word in words1:
    print(reg_exp.stem(word))
```

```
eat
un
understand
read
plant
```

Snowball Stemmer

In [21]:
```
from nltk.stem import SnowballStemmer
```

In [22]:
```
Snowball_stemmer = SnowballStemmer("english")
```

In [23]:
```
for word in words:
    print(Snowball_stemmer.stem(word))
```

```
eat
east
eaten
write
write
program
program
final
final
histori
```

In [24]:
```
stemming.stem("fairly"), stemming.stem("sportingly")  # Porter Stemmer
```

Out[24]: ('fairli', 'sportingli')

In [25]:
```
Snowball_stemmer.stem("fairly"), Snowball_stemmer.stem("sportingly")  # S
```

Out[25]: ('fair', 'sport')

Lemmatization

In [26]:
```
from nltk.stem import WordNetLemmatizer  #it's a class which is a thin wr
```

In [27]:
```
lemmatizer = WordNetLemmatizer()
```

In [28]:
```
print(lemmatizer.lemmatize("going"))  # Default pos = n
print(lemmatizer.lemmatize("going", pos = 'v'))
print(lemmatizer.lemmatize("going", pos = 'a')) # Takes 2 parameter (word
# Noun - n
# Verb - v
# Adjective - a
# Adverb - r
```

```
going
go
going
```

In [29]:
```
for word in words:
    print(lemmatizer.lemmatize(word,pos='n'))
```

```
eating
east
eaten
writing
writes
programming
program
finally
finalize
history
```

StopWords

In [30]:
```
## Speech Of DR APJ Abdul Kalam
paragraph = """I have three visions for India. In 3000 years of our histo
               the world have come and invaded us, captured our lands, co
               From Alexander onwards, the Greeks, the Turks, the Moguls,
               the French, the Dutch, all of them came and looted us, too
               Yet we have not done this to any other nation. We have not
               We have not grabbed their land, their culture,
               their history and tried to enforce our way of life on them
               Why? Because we respect the freedom of others.That is why
               first vision is that of freedom. I believe that India got
               this in 1857, when we started the War of Independence. It
```

```
we must protect and nurture and build on. If we are not fr
My second vision for India’s development. For fifty years
It is time we see ourselves as a developed nation. We are
in terms of GDP. We have a 10 percent growth rate in most
Our achievements are being globally recognised today. Yet
see ourselves as a developed nation, self-reliant and self
I have a third vision. India must stand up to the world. B
stands up to the world, no one will respect us. Only stren
strong not only as a military power but also as an economi
My good fortune was to have worked with three great minds.
space, Professor Satish Dhawan, who succeeded him and Dr.
I was lucky to have worked with all three of them closely
I see four milestones in my career"""
```

In [31]:
```
from nltk.corpus import stopwords
```

In [32]:
```
print(len(stopwords.words('english')))
print(len(stopwords.words('german')))
print(len(stopwords.words('french')))
```

```
179
232
157
```

Port Stemmer

In [33]:
```
sentence = sent_tokenize(paragraph)
for sent in range(len(sentence)):
    words = word_tokenize(sentence[sent])
    words = [stemming.stem(word) for word in words if word not in set(sto
    sentence[sent] = ' '.join(words)
sentence
```

Out[33]:
```
['i three vision india .',
 'in 3000 year histori , peopl world come invad us , captur land , conqu
er mind .',
 'from alexand onward , greek , turk , mogul , portugues , british , fre
nch , dutch , came loot us , took .',
 'yet done nation .',
 'we conquer anyon .',
 'we grab land , cultur , histori tri enforc way life .',
 'whi ?',
 'becaus respect freedom others.that first vision freedom .',
 'i believ india got first vision 1857 , start war independ .',
 'it freedom must protect nurtur build .',
 'if free , one respect us .',
 'my second vision india ’ develop .',
 'for fifti year develop nation .',
 'it time see develop nation .',
 'we among top 5 nation world term gdp .',
 'we 10 percent growth rate area .',
 'our poverti level fall .',
 'our achiev global recognis today .',
 'yet lack self-confid see develop nation , self-reli self-assur .',
 'isn ’ incorrect ?',
 'i third vision .',
 'india must stand world .',
 'becaus i believ unless india stand world , one respect us .',
 'onli strength respect strength .',
 'we must strong militari power also econom power .',
 'both must go hand-in-hand .',
 'my good fortun work three great mind .',
 'dr. vikram sarabhai dept .',
 'space , professor satish dhawan , succeed dr. brahm prakash , father n
uclear materi .',
 'i lucki work three close consid great opportun life .',
 'i see four mileston career']
```

SnowSnowball Stemmer

In [34]:
```
sentence = sent_tokenize(paragraph)
for sent in range(len(sentence)):
    words = word_tokenize(sentence[sent])
    words = [Snowball_stemmer.stem(word) for word in words if word not in
    sentence[sent] = ' '.join(words)
sentence
```

Out[34]:
```
['i three vision india .',
 'in 3000 year histori , peopl world come invad us , captur land , conqu
er mind .',
 'from alexand onward , greek , turk , mogul , portugues , british , fre
nch , dutch , came loot us , took .',
 'yet done nation .',
 'we conquer anyon .',
 'we grab land , cultur , histori tri enforc way life .',
 'whi ?',
 'becaus respect freedom others.that first vision freedom .',
 'i believ india got first vision 1857 , start war independ .',
 'it freedom must protect nurtur build .',
 'if free , one respect us .',
 'my second vision india ’ develop .',
 'for fifti year develop nation .',
 'it time see develop nation .',
 'we among top 5 nation world term gdp .',
 'we 10 percent growth rate area .',
 'our poverti level fall .',
 'our achiev global recognis today .',
 'yet lack self-confid see develop nation , self-reli self-assur .',
 'isn ’ incorrect ?',
 'i third vision .',
 'india must stand world .',
 'becaus i believ unless india stand world , one respect us .',
 'onli strength respect strength .',
 'we must strong militari power also econom power .',
 'both must go hand-in-hand .',
 'my good fortun work three great mind .',
 'dr. vikram sarabhai dept .',
 'space , professor satish dhawan , succeed dr. brahm prakash , father n
uclear materi .',
 'i lucki work three close consid great opportun life .',
 'i see four mileston career']
```

Wordnet Lemmatization

In [35]:
```
sentence = sent_tokenize(paragraph)
for sent in range(len(sentence)):
    words = word_tokenize(sentence[sent])
    words = [lemmatizer.lemmatize(word) for word in words if word not in
    sentence[sent] = ' '.join(words)
sentence
```

Out[35]:
```
['I three vision India .',
 'In 3000 year history , people world come invaded u , captured land , c
onquered mind .',
 'From Alexander onwards , Greeks , Turks , Moguls , Portuguese , Britis
h , French , Dutch , came looted u , took .',
 'Yet done nation .',
 'We conquered anyone .',
 'We grabbed land , culture , history tried enforce way life .',
 'Why ?',
 'Because respect freedom others.That first vision freedom .',
 'I believe India got first vision 1857 , started War Independence .',
 'It freedom must protect nurture build .',
 'If free , one respect u .',
 'My second vision India ’ development .',
 'For fifty year developing nation .',
 'It time see developed nation .',
 'We among top 5 nation world term GDP .',
 'We 10 percent growth rate area .',
 'Our poverty level falling .',
 'Our achievement globally recognised today .',
 'Yet lack self-confidence see developed nation , self-reliant self-assu
red .',
 'Isn ’ incorrect ?',
 'I third vision .',
 'India must stand world .',
 'Because I believe unless India stand world , one respect u .',
 'Only strength respect strength .',
 'We must strong military power also economic power .',
 'Both must go hand-in-hand .',
 'My good fortune worked three great mind .',
 'Dr. Vikram Sarabhai Dept .',
 'space , Professor Satish Dhawan , succeeded Dr. Brahm Prakash , father
nuclear material .',
 'I lucky worked three closely consider great opportunity life .',
 'I see four milestone career']
```

In [ ]: