Shri Ramdeobaba College of Engineering and Management, Nagpur

Department of Computer Science and Engineering

Natural Language Processing Lab

Name : Shantanu Mane
Branch : CSE - AIML (VIth SEM)
Roll Num : E-63

AIM :

1. Write a Python NLTK program to split the text sentence/paragraph from genesis corpus and display it into a list of words. Remove the Punctuation and Stopwords from the given text and perform Stemming and POS tagging.
2. Write a Python program to tokenize sentences with nltk, spacy and gensim in language other than English (german).

1. Importing the Dependencies

```
from nltk.corpus import genesis
from nltk.corpus import stopwords
from nltk.tokenize import RegexpTokenizer
from nltk.stem import PorterStemmer
import spacy
```

Part A

2. Working with the Genesis Corpus

2.1 Analyzing the Genesis Corpus

```
genesis.fileids()
```

```
['english-kjv.txt',
 'english-web.txt',
 'finnish.txt',
 'french.txt',
 'german.txt',
 'lolcat.txt',
 'portuguese.txt',
 'swedish.txt']
```

```
genesis.words('english-web.txt')
```

```
['In', 'the', 'beginning', 'God', 'created', 'the', ...]
```

```
genesisWords = genesis.words('english-web.txt')

len(genesisWords)
```

```
44054
```

2.2 Removing StopWords from the Genesis Corpus Text

2.2.1 List of Predefined Stopwords

```
stopwordsList = stopwords.words('english')

print(
    f"Some of the stop words : {stopwordsList[:15]}, length of the stopwords list in english language :
        {len(stopwordsList)}"
)
```

```
Some of the stop words : ['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're", "you've", "you'll",
"you'd", 'your', 'yours'], length of the stopwords list in english language : 179
```

```
genesisWords_wo_stopwords = [word for word in genesisWords if not word in stopwordsList]

len(genesisWords_wo_stopwords)
```

```
25756
```

As we can see, we have reduced the length of genesis words list from 44054 to 25756, which implies that stop words are removed successfully.

```
genesisWords_wo_stopwords
```

```
['In',
 'beginning',
 'God',
 'created',
 'heavens',
 'earth',
 '.',
 'Now',
 'earth',
 'formless',
 'empty',
 '.',
 'Darkness',
 'surface',
 'deep',
 ...]
```

2.3 Removing the Punctuations from Genesis Corpus

2.3.1 Using NLTK's RegExpTokeniser

```
regexTokenizer = RegexpTokenizer(r'\w+')
```

```
' '.join(genesisWords_wo_stopwords)
```

```
'In beginning God created heavens earth . Now earth formless empty . Darkness surface deep . God \' Spirit hovering surface
waters . God said , " Let light ," light . God saw light , saw good . God divided light darkness . God called light Day ,
darkness called Night . There evening morning , one day . God said , " Let expanse middle waters , let divide waters waters
." God made expanse , divided waters expanse waters expanse ; . God called expanse sky . There evening morning , second day .
God said , " Let waters sky gathered together one place , let dry land appear ;" . God called dry land Earth , gathering
together waters called Seas . God saw good . God said , " Let earth put forth grass , herbs yielding seed , fruit trees
bearing fruit kind , seed , earth ;" . The earth brought forth grass , herbs yielding seed kind , trees bearing fruit , seed
, kind ; God saw good . There evening morning , third day . God said , " Let lights expanse sky divide day night ; let signs
, seasons , days years ; let lights expanse sky give light earth ;" . God made two great lights : greater light rule day ,
lesser light rule night . He also made stars . God set expanse sky give light earth , rule day night , divide light darkness
. God saw good . There evening morning , fourth day . God said , " Let waters swarm swarms living creatures , let birds fly
earth open expanse sky ." God created large sea creatures , every living creature moves , waters swarmed , kind , every
winged bird kind . God saw good . God blessed , saying , " Be fruitful , multiply , fill waters seas , let birds multiply
earth ." There evening morning , fifth day . God said , " Let earth bring forth living creatures kind , livestock , creeping
things , animals earth kind ;" . God made animals earth kind , livestock kind , everything creeps ground kind . God saw good
. God said , " Let us make man image , likeness : let dominion fish sea , birds sky , livestock , earth , every creeping
thing creeps earth ." God created man image . In God \' image created ; male female created . God blessed . God said , " Be
fruitful , multiply , fill earth , subdue . Have dominion fish sea , birds sky , every living thing moves earth ." God said ,
" Behold , I given every herb yielding seed , surface earth , every tree , bears fruit yielding seed . It food . To every
```

```
animal earth , every bird sky , everything creeps earth , life , I given every green herb food ;" . God saw everything made ,
, behold , good . There evening morning , sixth day . The heavens earth finished , vast array . On seventh day God finished
work made ; rested seventh day work made . God blessed seventh day , made holy , rested work created made . This history
generations heavens earth created , day Yahweh God made earth heavens . No plant field yet earth , herb field yet sprung ;
Yahweh God caused rain earth . There man till ground , mist went earth , watered whole surface ground . Yahweh God formed man
dust ground , breathed nostrils breath life ; man became living soul . Yahweh God planted garden eastward , Eden , put man
formed . Out ground Yahweh God made every tree grow pleasant sight , good food ; tree life also middle garden , tree
knowledge good evil . A river went Eden water garden ; parted , became four heads . The name first Pishon : one flows whole
land Havilah , gold ; gold land good . There aromatic resin onyx stone . The name second river Gihon : river flows whole land
Cush . The name third river Hiddekel : one flows front Assyria . The fourth river Euphrates . Yahweh God took man , put
garden Eden dress keep . Yahweh God commanded man , saying , " Of every tree garden may freely eat ; tree knowledge good evil
, shall eat ; day eat surely die ." Yahweh God said , " It good man alone ; I make helper suitable ." Out ground Yahweh God
formed every animal field , every bird sky , brought man see would call . Whatever man called every living creature , name .
The man gave names livestock , birds sky , every animal field ; man found helper suitable . Yahweh God caused deep sleep fall
man , slept ; took one ribs , closed flesh place . He made rib , Yahweh God taken man , woman , brought man . The man said ,
" This bone bones , flesh flesh . She called Woman , taken Man ." Therefore man leave father mother , join wife , one
```

```python
genesisWords_wo_punctuations = regexTokenizer.tokenize(' '.join(genesisWords_wo_stopwords))

genesisWords_wo_punctuations[:50]
```

```
['In',
 'beginning',
 'God',
 'created',
 'heavens',
 'earth',
 'Now',
 'earth',
 'formless',
 'empty',
 'Darkness',
 'surface',
 'deep',
 'God',
 'Spirit',
 'hovering',
 'surface',
 'waters',
 'God',
 'said',
 'Let',
 'light',
 'light',
 'God',
 'saw',
 'light',
 'saw',
 'good',
 'God',
 'divided',
 'light',
 'darkness',
 'God',
 'called',
 'light',
 'Day',
 'darkness',
 'called',
 'Night',
 'There',
 'evening',
 'morning',
 'one',
 'day',
 'God',
 'said',
 'Let',
 'expanse',
 'middle',
 'waters']
```

3. Stemming and POS Tagging of Genesis Corpus

3.1 Stemming Using `PorterStemmer`

```python
pStemmer = PorterStemmer()

print(f"Word".ljust(15), f"Stem Word")
print("-" * 25)
for word in genesisWords_wo_punctuations[:25]:
    print(f"{word.lower()}".ljust(15), f"{pStemmer.stem(word)}")
```

```
Word            Stem Word
-------------------------
in              in
beginning       begin
god             god
created         creat
heavens         heaven
earth           earth
now             now
earth           earth
formless        formless
empty           empti
darkness        dark
surface         surfac
```

```
deep            deep
god             god
spirit          spirit
hovering        hover
surface         surfac
waters          water
god             god
said            said
let             let
light           light
light           light
god             god
saw             saw
```

3.2 Part Of Speech Tagging Using SpaCy

```
spacyObj = spacy.load('en_core_web_sm')
```

```
genesisWords_Doc = spacyObj(" ".join(genesisWords_wo_punctuations))

genesisWords_Doc[:100]
```

```
In beginning God created heavens earth Now earth formless empty Darkness surface deep God Spirit hovering surface waters God
said Let light light God saw light saw good God divided light darkness God called light Day darkness called Night There
evening morning one day God said Let expanse middle waters let divide waters waters God made expanse divided waters expanse
waters expanse God called expanse sky There evening morning second day God said Let waters sky gathered together one place
let dry land appear God called dry land Earth gathering together waters called Seas God saw good God said Let
```

```
print(f"Word".ljust(15), "Alpha", "Space", "Stop", "Punctuation")
print("-" * 47)
for word in genesisWords_Doc[:25]:
    print(f"{word}".ljust(15), word.is_alpha, '', word.is_space, word.is_stop, word.is_punct)
```

```
Word            Alpha Space Stop Punctuation
-----------------------------------------------
In              True  False True False
beginning       True  False False False
God             True  False False False
created         True  False False False
heavens         True  False False False
earth           True  False False False
Now             True  False True False
earth           True  False False False
formless        True  False False False
empty           True  False True False
Darkness        True  False False False
surface         True  False False False
deep            True  False False False
God             True  False False False
Spirit          True  False False False
hovering        True  False False False
surface         True  False False False
waters          True  False False False
God             True  False False False
said            True  False False False
Let             True  False False False
light           True  False False False
light           True  False False False
God             True  False False False
saw             True  False False False
```

Part B

4. Tokenizing German Language Using NLTK

4.1 Analyzing the German language from Genesis Corpus

```
# German language corpus is inbuilt in Genesis Corpus

genesis.fileids()[4]
```

```
'german.txt'
```

```
germanTokens = genesis.words('german.txt')

germanTokens
```

```
['Am', 'Anfang', 'schuf', 'Gott', 'Himmel', 'und', ...]
```

```
len(germanTokens)
```

```
43941
```

4.2 Removing stopwords from text

```
germanStopWords = set(stopwords.words('german'))

germanTokens[:10]
```

```
['Am', 'Anfang', 'schuf', 'Gott', 'Himmel', 'und', 'Erde', '.', 'Und', 'die']
```

```
germanTokens_wo_stopwords = [word for word in germanTokens if not word in germanStopWords]

print(f"text without stopwords : {germanTokens_wo_stopwords[:10]}, length of text : {len(germanTokens_wo_stopwords)}")
```

```
text without stopwords : ['Am', 'Anfang', 'schuf', 'Gott', 'Himmel', 'Erde', '.', 'Und', 'Erde', 'wüst'], length of text :
26699
```

As we can see, we have reduced the length of genesis words list from 43941 to 26699, which implies that stop words are removed successfully.

4.3 Removing punctuations using `RegexpTokenizer`

```
punctTokenizer = RegexpTokenizer(r'\w+')

germanWords_wo_punct = punctTokenizer.tokenize(" ".join(germanTokens_wo_stopwords))

print(f"text without punctuations : {germanWords_wo_punct[:10]}, length of text : {len(germanWords_wo_punct)}")
```

```
text without punctuations : ['Am', 'Anfang', 'schuf', 'Gott', 'Himmel', 'Erde', 'Und', 'Erde', 'wüst', 'leer'], length of
text : 18831
```

4.4 POS Tagging using SpaCy

```
spacyObjGerman = spacy.load("de_core_news_sm")

germanWords_Doc = spacyObjGerman(" ".join(germanTokens))

germanWords_Doc[:10]
```

```
Am Anfang schuf Gott Himmel und Erde . Und die
```

```
print(f"Word".ljust(15), "Alpha", "Space", "Stop", "Punctuation")
print("-" * 47)
for word in germanWords_Doc[:25]:
    print(f"{word}".ljust(15), word.is_alpha, '', word.is_space, word.is_stop, word.is_punct)
```

```
Word            Alpha Space Stop Punctuation
-----------------------------------------------
Am              True  False True False
```

```
Anfang          True  False False False
schuf           True  False False False
Gott            True  False False False
Himmel          True  False False False
und             True  False True False
Erde            True  False False False
.               False  False False True
Und             True  False True False
die             True  False True False
Erde            True  False False False
war             True  False True False
wüst            True  False False False
und             True  False True False
leer            True  False False False
,               False  False False True
und             True  False True False
es              True  False True False
war             True  False True False
finster         True  False False False
auf             True  False True False
der             True  False True False
Tiefe           True  False False False
;               False  False False True
und             True  False True False
```