CSE4022 Natural Language Processing

Digital Assignment -2

19BCE1022 - Parth Gupta

Create a text corpus with minimum 200 words (unique contents).

```
nltk.download('punkt')
nltk.download('wordnet')
nltk.download('omw-1.4')
nltk.download('averaged_perceptron_tagger')
```

```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
[nltk_data] Downloading package wordnet to /root/nltk_data...
[nltk_data]   Package wordnet is already up-to-date!
[nltk_data] Downloading package omw-1.4 to /root/nltk_data...
[nltk_data]   Package omw-1.4 is already up-to-date!
[nltk_data] Downloading package averaged_perceptron_tagger to
[nltk_data]     /root/nltk_data...
[nltk_data]   Unzipping taggers/averaged_perceptron_tagger.zip.
True
```

Creating corpus of the 2 text files file1 and file2

```
import os
from nltk.corpus.reader.plaintext import PlaintextCorpusReader

corpusdir = '/content/corpus'

newcorpus = PlaintextCorpusReader(corpusdir, '.*')
```

```
text=newcorpus.raw().strip()
print(newcorpus.raw().strip())
```

```
A path from a point approximately 330 metres east of the most south westerly corner of 1
Did he look like a doctor?
He ran into debt.
They concluded that he had told a lie.
```

```
I ran into Mary at the party last week.
I said nothing about the matter.
His brother is more patient than he is.
Tom was caught sneaking out of the room.
No one stops to listen to him.
Please wait around for a while.
There is going to be a storm. I clapped my hands. I have just finished my homework. What
I asked him if he knew my name.
```

Paragraphs Seementation

```
print(newcorpus.paras())
```

```
[[['A', 'path', 'from', 'a', 'point', 'approximately', '330', 'metres', 'east', 'of', 't
```

Sentences Segmentation

```
print(newcorpus.sents())
```

```
[['A', 'path', 'from', 'a', 'point', 'approximately', '330', 'metres', 'east', 'of', 'th
```

```
print(nltk.sent_tokenize(text))
```

```
['A path from a point approximately 330 metres east of the most south westerly corner of
```

Words Segementation

```
print(newcorpus.words())
```

```
['A', 'path', 'from', 'a', 'point', 'approximately', ...]
```

```
print(nltk.word_tokenize(text))
```

```
['A', 'path', 'from', 'a', 'point', 'approximately', '330', 'metres', 'east', 'of', 'the
```

Convert to Lowercase

```
text=newcorpus.raw().strip()
text
```

```
'A path from a point approximately 330 metres east of the most south westerly corner of
17 Batherton Close, Widnes and approximately 208 metres east-south-east of the most sou
therly corner of Unit 3 Foundry Industrial Estate, Victoria Street, Widnes, proceeding
in a generally east-north-easterly direction for approximately 28 metres to a point app
roximately 202 metres east-south-east of the most south-easterly corner of Unit 4 Found
ry Industrial Estate, Victoria Street, and approximately 347 metres east of the most so
```

```
text=text.lower()
text
```

```
'a path from a point approximately 330 metres east of the most south westerly corner of
17 batherton close, widnes and approximately 208 metres east-south-east of the most sou
therly corner of unit 3 foundry industrial estate, victoria street, widnes, proceeding
in a generally east-north-easterly direction for approximately 28 metres to a point app
roximately 202 metres east-south-east of the most south-easterly corner of unit 4 found
ry industrial estate, victoria street, and approximately 347 metres east of the most so
```

Stop Words Removal

```
from nltk.corpus import stopwords
stopword = stopwords.words('english')
word_tokens = nltk.word_tokenize(text)
removing_stopwords = [word for word in word_tokens if word not in stopword]
print (removing_stopwords)
```

```
['path', 'point', 'approximately', '330', 'metres', 'east', 'south', 'westerly', 'corner
```

Stemming (Porter Stemmer Algorithm)

```
from nltk.stem import SnowballStemmer
stopword = stopwords.words('english')
snowball_stemmer = SnowballStemmer('english')
word_tokens = nltk.word_tokenize(text)
stemmed_word = [snowball_stemmer.stem(word) for word in word_tokens]
print (stemmed_word)
```

```
['a', 'path', 'from', 'a', 'point', 'approxim', '330', 'metr', 'east', 'of', 'the', 'mos
```

Lemmatization

```
from nltk.stem import WordNetLemmatizer
stopword = stopwords.words('english')
wordnet_lemmatizer = WordNetLemmatizer()
word_tokens = nltk.word_tokenize(text)
lemmatized_word = [wordnet_lemmatizer.lemmatize(word) for word in word_tokens]
print (lemmatized_word)
```

```
['A', 'path', 'from', 'a', 'point', 'approximately', '330', 'metre', 'east', 'of', 'the
```

POS Tagging

```
word = nltk.word_tokenize(text)
pos_tag = nltk.pos_tag(word)
print (pos_tag)
```

```
[('A', 'DT'), ('path', 'NN'), ('from', 'IN'), ('a', 'DT'), ('point', 'NN'), ('approximat
```