CSE4022 NATURAL LANGUAGE PROCESSING

Name: Tanishka Khanolkar

Reg no: 20BCE1511

Utilize Python NLTK (Natural Language Tool Kit) Platform and do the following. Install relevant Packages and Libraries

```
import nltk
nltk.download('brown')
```

```
[nltk_data] Downloading package brown to /root/nltk_data...
[nltk_data]   Unzipping corpora/brown.zip.
True
```

Explore Brown Corpus and find the size, tokens, categories

```
from nltk.corpus import brown
```

```
brown.words()
```

```
['The', 'Fulton', 'County', 'Grand', 'Jury', 'said', ...]
```

Find the size of word tokens?


```
len(brown.words())
```

```
1161192
```

Find the size of word types?

```
len(set(brown.words()))
```

```
56057
```

Find the size of the category “government”

```
len(brown.words(categories="government"))
```

```
70117
```

List the most frequent tokens

```
freq = nltk.FreqDist(brown.words())
print("Common Words:", freq.most_common(10))
```

```
Common Words: [('the', 62713), (',', 58334), ('.', 49346), ('of', 36080), ('and', 27915), ('to', 25732), ('a', 21881), ('in', 19536), (
```

Count the number of sentences

```
len(brown.sents())
```

```
57340
```

Explore the corpora available in NLTK

```
from nltk.corpus import inaugural
nltk.download('inaugural')
from nltk.corpus import shakespeare
nltk.download('shakespeare')
```

```
[nltk_data] Downloading package inaugural to /root/nltk_data...
[nltk_data]   Unzipping corpora/inaugural.zip.
[nltk_data] Downloading package shakespeare to /root/nltk_data...
```

```
[nltk_data]   Package shakespeare is already up-to-date!
True
```

Raw corpus

```
inaugural.raw()
```

```
'Fellow-Citizens of the Senate and of the House of Representatives:\n\nAmong the
vicissitudes incident to life no event could have filled me with greater anxietie
s than that of which the notification was transmitted by your order, and received
on the 14th day of the present month. On the one hand, I was summoned by my Count
ry, whose voice I can never hear but with veneration and love, from a retreat whi
ch I had chosen with the fondest predilection, and, in my flattering hopes, with
an immutable decision, as the asylum of my declining years -- a retreat which was
rendered every day more necessary as well as more dear to me by the addition of h
abit to inclination, and of frequent interruptions in my health to the gradual wa
ste committed on it by time. On the other hand, the magnitude and difficulty of t
```

```
shakespeare.raw()
```

```
'<?xml version="1.0"?>\r\n<?xml-stylesheet type="text/css" href="shakes.css"?>\r
\n<!-- <!DOCTYPE PLAY SYSTEM "play.dtd"> -->\r\n\r\n<PLAY>\r\n<TITLE>The Tragedy
of Antony and Cleopatra</TITLE>\r\n<!--\r\n\r\n<P>Text placed in the public domai
n by Moby Lexical Tools, 1992.</P>\r\n<P>XML markup by Jon Bosak, 1996-1998.</P>
\r\n<P>XML Styling done by Ajay Juneja, 1999.<P>\r\n<P>This work may be freely co
pied and distributed worldwide.</P>\r\n\r\n-->\r\n<PERSONAE>\r\n<TITLE>Dramatis P
ersonae</TITLE>\r\n\r\n\r\n<PGROUP>\r\n<PERSONA>MARK ANTONY</PERSONA>\r\n<PERSONA
>OCTAVIUS CAESAR</PERSONA>\r\n<PERSONA>M. AEMILIUS LEPIDUS</PERSONA>\r\n<GRPDESCR
>triumvirs.</GRPDESCR>\r\n</PGROUP>\r\n\r\n<PERSONA>SEXTUS POMPEIUS</PERSONA>\r\n
```



```
print(brown.tagged_words())
```

```
[('The', 'AT'), ('Fulton', 'NP-TL'), ...]
```

```
from nltk.corpus import conll2000, switchboard
print(conll2000.tagged_words())
```

```
[('Confidence', 'NN'), ('in', 'IN'), ('the', 'DT'), ...]
```

Parsed

```
from nltk.corpus import treebank
print(treebank.parsed_sents('wsj_0003.mrg')[0])
```

```
(S
  (S-TPC-1
    (NP-SBJ
      (NP (NP (DT A) (NN form)) (PP (IN of) (NP (NN asbestos))))
      (RRC
        (ADVP-TMP (RB once))
        (VP
          (VBN used)
          (NP (-NONE- *))
          (S-CLR
            (NP-SBJ (-NONE- *))
            (VP
              (TO to)
              (VP
                (VB make)
                (NP (NNP Kent) (NN cigarette) (NNS filters))))))))))
    (VP
      (VBZ has)
      (VP
        (VBN caused)
        (NP
          (NP (DT a) (JJ high) (NN percentage))
          (PP (IN of) (NP (NN cancer) (NNS deaths)))
          (PP-LOC
            (IN among)
            (NP
              (NP (DT a) (NN group))
              (PP
                (IN of)
                (NP
```

```
                      (NP (NNS workers))
                      (RRC
                        (VP
                          (VBN exposed)
                          (NP (-NONE- *))
                          (PP-CLR (TO to) (NP (PRP it)))
                          (ADVP-TMP
                            (NP
                              (QP (RBR more) (IN than) (CD 30))
                              (NNS years))
                            (IN ago))))))))))))
      (, ,)
      (NP-SBJ (NNS researchers))
      (VP (VBD reported) (SBAR (-NONE- 0) (S (-NONE- *T*-1))))
      (. .))
```

```
from nltk.corpus import conll2007
print(conll2007.parsed_sents('esp.train')[0].tree())
```

```
    (fortaleció
      (aumento El (del (índice (de (desempleo estadounidense)))))
      hoy
      considerablemente
      (al
        (euro
          (cotizaba
            ,
            que
            (a (15.35 las GMT))
            se
```

Automatic saving failed. This file was updated remotely or in another tab. Show diff

```
      .)
```

Multilingual aligned

```
from nltk.corpus import wordnet as wn
nltk.download('omw-1.4')
```

```
    [nltk_data] Downloading package omw-1.4 to /root/nltk_data...
    True
```

```
wn.langs()
```

```
    dict_keys(['eng', 'als', 'arb', 'bul', 'cmn', 'dan', 'ell', 'fin', 'fra', 'heb', 'hrv', 'isl', 'ita', 'ita_iwn', 'jpn', 'cat', 'eus',
    'glg', 'spa', 'ind', 'zsm', 'nld', 'nno', 'nob', 'pol', 'por', 'ron', 'lit', 'slk', 'slv', 'swe', 'tha'])
```

Spoken language

```
from nltk.corpus import indian
indian.raw()
```

```
    '<Corpora type="Monolingual-POS-TAGGED" Language="Bangla">\n<Sentence id=1>\nম
    হিষের_NN সন্তান_NN :_SYM তোড়া_NNP উপজাতি_NN ।_SYM \n</Sentence>\n<Sentence id=2>\n
    বাসস্থান-ঘরগৃহস্থালি_NN তোড়া_NNP ভাষায়_NN গ্রামকেও_NN বলে_VM `_SYM মোদ_NN \'_SYM ।_SYM
    \n</Sentence>\n<Sentence id=3>\nমোদের_NN আয়তন_NN খুব_INTF বড়ো_JJ নয়_VM ।_SYM \n
    </Sentence>\n<Sentence id=4>\nপ্রতি_QF মোদে_NN আছে_VM কিছু_QF কুঁড়েঘর_NN ,_SYM
    সাধারণ_JJ মহিষশালা_NN ।_SYM \n</Sentence>\n<Sentence id=5>\nআর_CC গ্রামের_NN বাইরে_N
    ST থাকে_VM ডেয়ারি-মন্দির_NN ।_SYM \n</Sentence>\n<Sentence id=6>\nআয়তনের_NN তারতম্য
    _NN অনুসারে_PSP গ্রামগুলি_NN দু_QC রকমের_NN :_SYM এতূডমোদ_NNP (_SYM বড়ো_JJ গ্রাম_NN
    )_SYM ওকিনমোদ_NNP (_SYM ছোট_JJ গ্রাম_NN )_SYM ।_SYM \n</Sentence>\n<Sentence id=7>
    \nকোন_DEM কোন_RDP গ্রামের_NN আবার_CC ধর্মীয়_JJ বা_C\ufeffমহিষের_NN সন্তান_NN :_SYM
```

Semantic tagged

```
brown.categories()
```

```
    ['adventure',
     'belles_lettres',
     'editorial',
     'fiction',
     'government',
     'hobbies',
     'humor',
```

```
 'learned',
 'lore',
 'mystery',
 'news',
 'religion',
 'reviews',
 'romance',
 'science_fiction']
```

```
from nltk.corpus import reuters

reuters.categories()
```

```
['acq',
 'alum',
 'barley',
 'bop',
 'carcass',
 'castor-oil',
 'cocoa',
 'coconut',
 'coconut-oil',
 'coffee',
 'copper',
 'copra-cake',
 'corn',
 'cotton',
 'cotton-oil',
 'cpi',
 'cpu',
```

Automatic saving failed. This file was updated remotely or in another tab. Show diff

```
 'dmk',
 'earn',
 'fuel',
 'gas',
 'gnp',
 'gold',
 'grain',
 'groundnut',
 'groundnut-oil',
 'heat',
 'hog',
 'housing',
 'income',
 'instal-debt',
 'interest',
 'ipi',
 'iron-steel',
 'jet',
 'jobs',
 'l-cattle',
 'lead',
 'lei',
 'lin-oil',
 'livestock',
 'lumber',
 'meal-feed',
 'money-fx',
 'money-supply',
 'naphtha',
 'nat-gas',
 'nickel',
 'nkr',
 'nzdlr',
 'oat',
 'oilseed',
 'orange',
 'palladium',
 'palm-oil',
```

Create a text corpus with a minimum of 200 words (unique content). Implement the following text processing

To prevent unwanted access to one's computer and personal information, passwords are the first line of security. The greater the password security, the better the computer's defence against hackers and dangerous software. Therefore, it is important to keep strong passwords for all of your computer accounts, especially in this day and age when technology is widely used and we need to set passwords for a lot of different accounts to complete even the smallest chores. A complex and long password will make it very difficult for a hacker to crack it, whether through a brute-force attack (trying every possible combination of numbers, letters, or special characters) or an automated machine attack

trying thousands of combinations per second to guess your one and only. As a result, the more complex the password, the greater the security for your account. An account is where you keep a lot of sensitive information that you don't want stolen. As a result, safeguarding an account password is critical. Though there are many alternatives to passwords for access control, in many applications, the password is the more compellingly authenticating the identity. Password strength metres provide simple and immediate visual feedback on what constitutes a strong password.

```
import os
PATH = os.getcwd()
FILE_NAME = "samplecorpus.txt"
```

```
from nltk.corpus.reader.plaintext import PlaintextCorpusReader
samplecorpus = PlaintextCorpusReader(PATH,FILE_NAME)
```

```
samplecorpus.raw()
```

```
'To prevent unwanted access to one's computer and personal information, passwords
are the first line of security. The greater the password security, the better the
computer's defence against hackers and dangerous software. Therefore, it is impor
tant to keep strong passwords for all of your computer accounts, especially in th
is day and age when technology is widely used and we need to set passwords for a
lot of different accounts to complete even the smallest chores. A complex and lon
g password will make it very difficult for a hacker to crack it, whether through
a brute-force attack (trying every possible combination of numbers, letters, or s
pecial characters) or an automated machine attack trying thousands of combination
s per second to guess your one and only. As a result, the more complex the passwo
```

Automatic saving failed. This file was updated remotely or in another tab. Show diff

```
samplecorpus.words()
```

```
['To', 'prevent', 'unwanted', 'access', 'to', 'one', ...]
```

Sentence Segmentation

```
samplecorpus.sents()
```

```
[['To', 'prevent', 'unwanted', 'access', 'to', 'one', "'", 's', 'computer', 'and', 'personal', 'information', ',', 'passwords', 'are',
'the', 'first', 'line', 'of', 'security', '.'], ['The', 'greater', 'the', 'password', 'security', ',', 'the', 'better', 'the',
'computer', "'", 's', 'defence', 'against', 'hackers', 'and', 'dangerous', 'software', '.'], ...]
```

Convert to Lowercase

```
text='To prevent unwanted access to ones computer and personal information, passwords are the first line of security. The greater the passwor
```

```
text.lower()
```

```
'to prevent unwanted access to ones computer and personal information, passwords
are the first line of security. the greater the password security, the better the
computers defence against hackers and dangerous software. therefore, it is import
ant to keep strong passwords for all of your computer accounts, especially in thi
s day and age when technology is widely used and we need to set passwords for a l
ot of different accounts to complete even the smallest chores. a complex and long
password will make it very difficult for a hacker to crack it, whether through a
brute-force attack (trying every possible combination of numbers, letters, or spe
cial characters) or an automated machine attack trying thousands of combinations
per second to guess your one and only. as a result, the more complex the passwor
```

Stop words removal

```
nltk.download('stopwords')
```

```
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Unzipping corpora/stopwords.zip.
True
```

```
from nltk.corpus import stopwords
en_stops = set(stopwords.words('english'))
words = []
for x in samplecorpus.words():
```

```
  if x not in en_stops:
    words.append(x)
print(words)
```

```
['To', 'prevent', 'unwanted', 'access', 'one', "'", 'computer', 'personal', 'information', ',', 'passwords', 'first', 'line', 'security
```

Stemming

```
from nltk.stem import PorterStemmer
from nltk.tokenize import word_tokenize

ps = PorterStemmer()


words = word_tokenize(text)
s = ""
l = []
for w in words:
  s = w + " : " + ps.stem(w)
  l.append(s)
print(l)
```

```
['To : to', 'prevent : prevent', 'unwanted : unwant', 'access : access', 'to : to', 'ones : one', 'computer : comput', 'and : and', 'pe
```

Automatic saving failed. This file was updated remotely or in another tab. Show diff

```
import nltk
from nltk.stem import   WordNetLemmatizer
wordnet_lemmatizer = WordNetLemmatizer()

tokenization = nltk.word_tokenize(text)
for w in tokenization:
  print("Lemma for {} is {}".format(w, wordnet_lemmatizer.lemmatize(w)))
```

```
Lemma for is is is
Lemma for critical is critical
Lemma for . is .
Lemma for Though is Though
Lemma for there is there
Lemma for are is are
Lemma for many is many
Lemma for alternatives is alternative
Lemma for to is to
Lemma for passwords is password
Lemma for for is for
Lemma for access is access
Lemma for control is control
Lemma for , is ,
Lemma for in is in
Lemma for many is many
Lemma for applications is application
```

Part of speech tagger

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

```
[('To', 'TO'), ('prevent', 'VB'), ('unwanted', 'JJ'), ('access', 'NN'), ('to', 'TO'), ('ones', 'NNS'), ('computer', 'NN'), ('and', 'CC'
```

Automatic saving failed. This file was updated remotely or in another tab. Show diff

✓ 40s completed at 1:18 AM

Could not connect to the reCAPTCHA service. Please check your internet connection and reload to get a reCAPTCHA challenge.