NLP Project

Details of the book :

- Title: **The Courtship of Animals**
- Author: **William Plane Pycraft**
- Language: **English**
- Number of characters : **594554**
- Number of words : **99525**
- Number of tokens : **112871**

Preprocessing Steps:

- **Text normalization-** normalization generally refers to a series of related tasks meant to put all text on a level playing field.

1. Converting all words to lowercase
2. Removing numbers
3. Removing whitespace

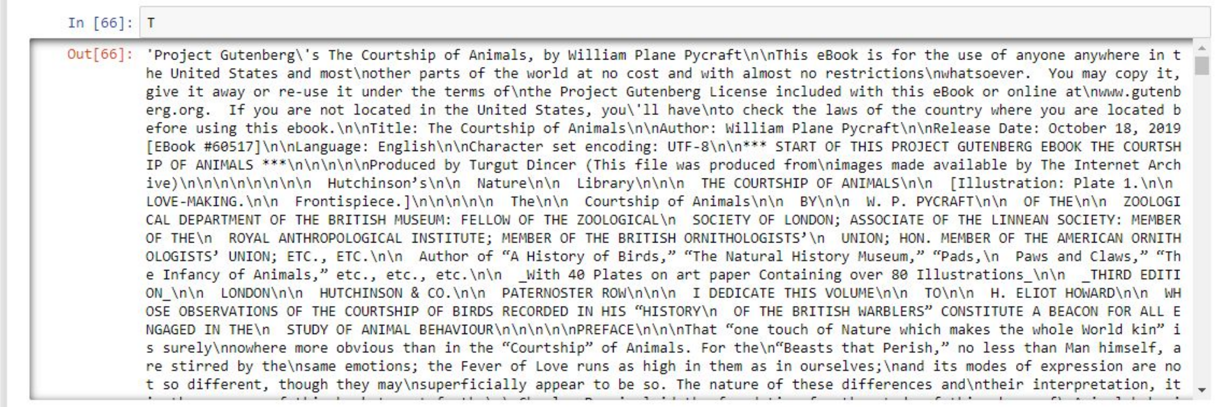
- **Text tokenization**- tokenization is a step which splits longer strings of text into smaller pieces, or tokens.

  1. Stemming- It is the process of eliminating affixes (suffixed, prefixes, infixes, circumfixes) from a word in order to obtain a word stem.
  2. Lemmatization- It is related to stemming, differing in that lemmatization is able to capture canonical forms based on a word's lemma.

Preprocessing helps in getting rid of the less useful parts of text through stopword removal, dealing with capitalization and characters and other details.It consists of the translation (mapping) of terms in the scheme or linguistic reductions.

CODE :

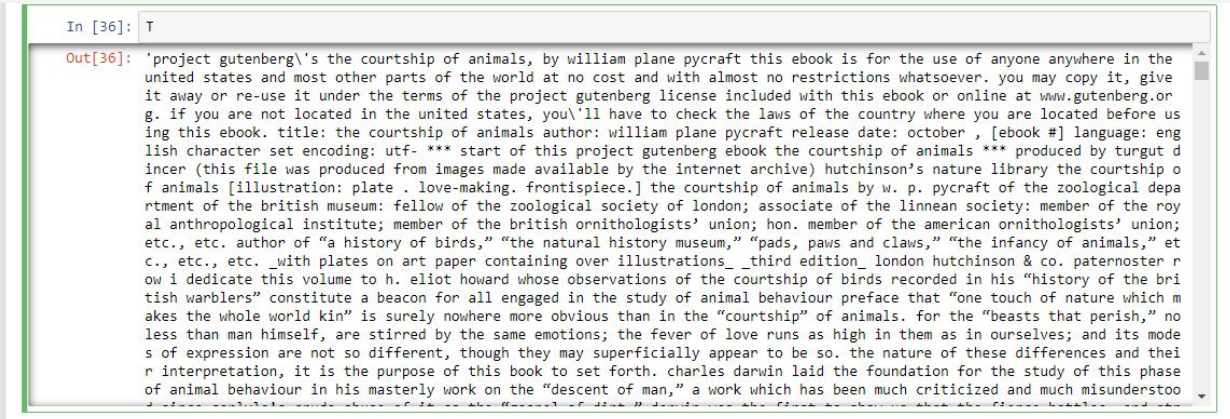
```
import nltk
nltk.download('all')
from nltk.tokenize import word_tokenize,sent_tokenize
from nltk.corpus import stopwords
from nltk.stem import PorterStemmer
from nltk.stem import WordNetLemmatizer
import matplotlib.pyplot as plt
import numpy as np
import re
import string

f = open("Book.txt","r")
T = f.read()
```



```
In [66]: T

Out[66]: 'Project Gutenberg\'s The Courtship of Animals, by William Plane Pycraft\n\nThis eBook is for the use of anyone anywhere in t
         he United States and most\nother parts of the world at no cost and with almost no restrictions\nwhatsoever.  You may copy it,
         give it away or re-use it under the terms of\nthe Project Gutenberg License included with this eBook or online at\nwww.gutenb
         erg.org.  If you are not located in the United States, you\'ll have\nto check the laws of the country where you are located b
         efore using this ebook.\n\nTitle: The Courtship of Animals\n\nAuthor: William Plane Pycraft\n\nRelease Date: October 18, 2019
         [EBook #60517]\n\nLanguage: English\n\nCharacter set encoding: UTF-8\n\n*** START OF THIS PROJECT GUTENBERG EBOOK THE COURTSH
         IP OF ANIMALS ***\n\n\n\n\nProduced by Turgut Dincer (This file was produced from\nimages made available by The Internet Arch
         ive)\n\n\n\n\n\n\n\n  Hutchinson’s\n\n  Nature\n\n  Library\n\n\n  THE COURTSHIP OF ANIMALS\n\n  [Illustration: Plate 1.\n\n
         LOVE-MAKING.\n\n  Frontispiece.]\n\n\n\n\n  The\n\n  Courtship of Animals\n\n  BY\n\n  W. P. PYCRAFT\n\n  OF THE\n\n  ZOOLOGI
         CAL DEPARTMENT OF THE BRITISH MUSEUM: FELLOW OF THE ZOOLOGICAL\n  SOCIETY OF LONDON; ASSOCIATE OF THE LINNEAN SOCIETY: MEMBER
         OF THE\n  ROYAL ANTHROPOLOGICAL INSTITUTE; MEMBER OF THE BRITISH ORNITHOLOGISTS’\n  UNION; HON. MEMBER OF THE AMERICAN ORNITH
         OLOGISTS’ UNION; ETC., ETC.\n\n  Author of “A History of Birds,” “The Natural History Museum,” “Pads,\n  Paws and Claws,” “Th
         e Infancy of Animals,” etc., etc., etc.\n\n  _With 40 Plates on art paper Containing over 80 Illustrations_\n\n  _THIRD EDITI
         ON_\n\n  LONDON\n\n  HUTCHINSON & CO.\n\n  PATERNOSTER ROW\n\n\n  I DEDICATE THIS VOLUME\n\n  TO\n\n  H. ELIOT HOWARD\n\n  WH
         OSE OBSERVATIONS OF THE COURTSHIP OF BIRDS RECORDED IN HIS “HISTORY\n  OF THE BRITISH WARBLERS” CONSTITUTE A BEACON FOR ALL E
         NGAGED IN THE\n  STUDY OF ANIMAL BEHAVIOUR\n\n\n\n\nPREFACE\n\n\nThat “one touch of Nature which makes the whole World kin” i
         s surely\nnowhere more obvious than in the “Courtship” of Animals. For the\n“Beasts that Perish,” no less than Man himself, a
         re stirred by the\nsame emotions; the Fever of Love runs as high in them as in ourselves;\nand its modes of expression are no
         t so different, though they may\nsuperficially appear to be so. The nature of these differences and\ntheir interpretation, it
```

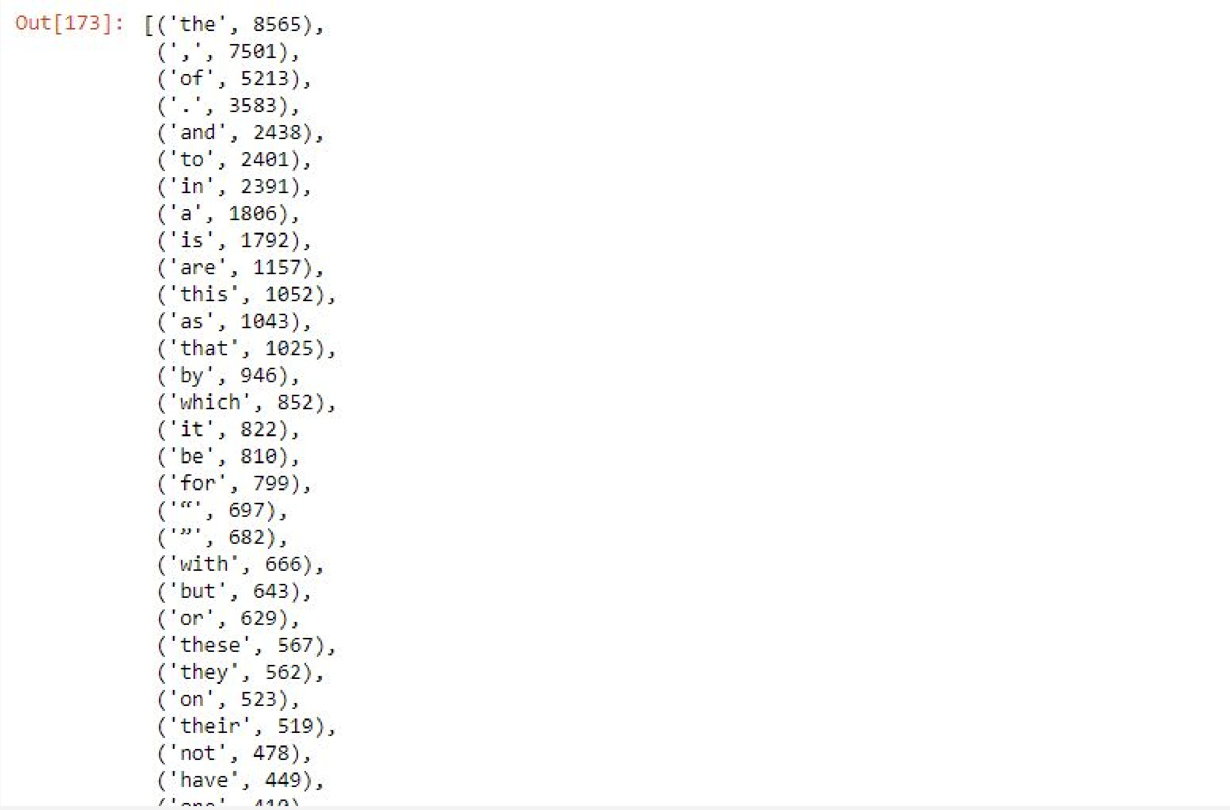
```
#no. of words
no_of_words = len(T.split())


#no. of characters
no_of_char=len(T)


#text normalization

#converting to lower case
T=T.lower()

#removing numbers
T = re.sub(r'\d+','', T)
```

```
#removing whitespaces
T=" ".join(T.split())
```



```
In [36]: T
Out[36]: 'project gutenberg\'s the courtship of animals, by william plane pycraft this ebook is for the use of anyone anywhere in the
         united states and most other parts of the world at no cost and with almost no restrictions whatsoever. you may copy it, give
         it away or re-use it under the terms of the project gutenberg license included with this ebook or online at www.gutenberg.or
         g. if you are not located in the united states, you\'ll have to check the laws of the country where you are located before us
         ing this ebook. title: the courtship of animals author: william plane pycraft release date: october , [ebook #] language: eng
         lish character set encoding: utf- *** start of this project gutenberg ebook the courtship of animals *** produced by turgut d
         incer (this file was produced from images made available by the internet archive) hutchinson’s nature library the courtship o
         f animals [illustration: plate . love-making. frontispiece.] the courtship of animals by w. p. pycraft of the zoological depa
         rtment of the british museum: fellow of the zoological society of london; associate of the linnean society: member of the roy
         al anthropological institute; member of the british ornithologists’ union; hon. member of the american ornithologists’ union;
         etc., etc. author of “a history of birds,” “the natural history museum,” “pads, paws and claws,” “the infancy of animals,” et
         c., etc., etc. _with plates on art paper containing over illustrations_ _third edition_ london hutchinson & co. paternoster r
         ow i dedicate this volume to h. eliot howard whose observations of the courtship of birds recorded in his “history of the bri
         tish warblers” constitute a beacon for all engaged in the study of animal behaviour preface that “one touch of nature which m
         akes the whole world kin” is surely nowhere more obvious than in the “courtship” of animals. for the “beasts that perish,” no
         less than man himself, are stirred by the same emotions; the fever of love runs as high in them as in ourselves; and its mode
         s of expression are not so different, though they may superficially appear to be so. the nature of these differences and thei
         r interpretation, it is the purpose of this book to set forth. charles darwin laid the foundation for the study of this phase
         of animal behaviour in his masterly work on the “descent of man,” a work which has been much criticized and much misunderstoo
```

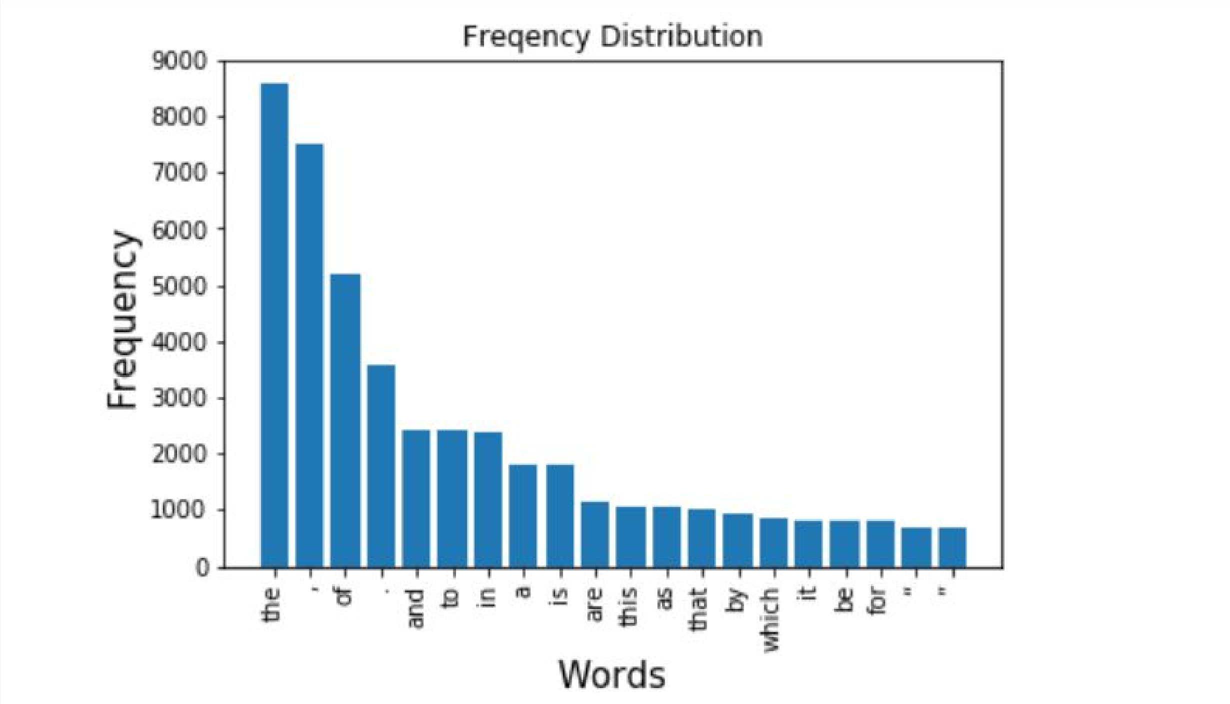
```
#text Tokenization
token=word_tokenize(T)
no_of_tokens=len(token)
frequency=nltk.FreqDist(set(token))
frequency
```



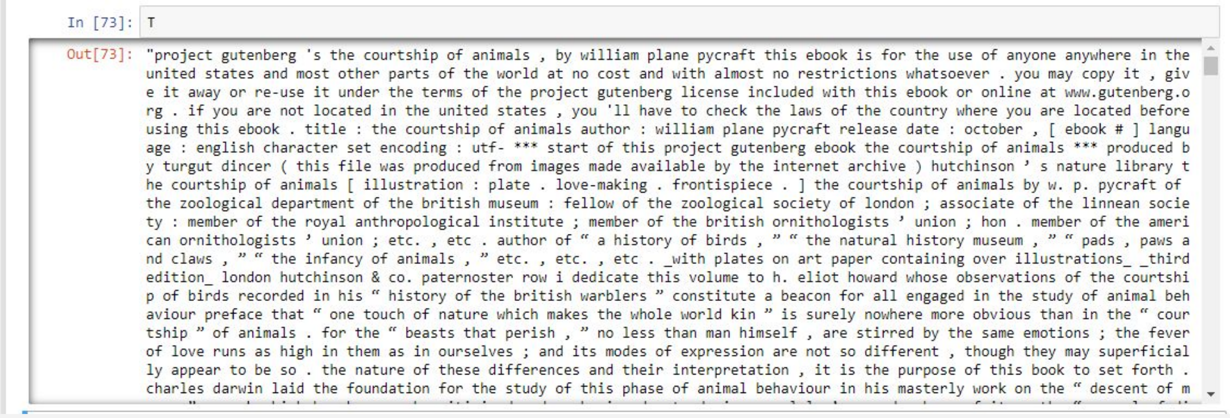
```
Out[173]: [('the', 8565),
           (',', 7501),
           ('of', 5213),
           ('.', 3583),
           ('and', 2438),
           ('to', 2401),
           ('in', 2391),
           ('a', 1806),
           ('is', 1792),
           ('are', 1157),
           ('this', 1052),
           ('as', 1043),
           ('that', 1025),
           ('by', 946),
           ('which', 852),
           ('it', 822),
           ('be', 810),
           ('for', 799),
           ('“', 697),
           ('”', 682),
           ('with', 666),
           ('but', 643),
           ('or', 629),
           ('these', 567),
           ('they', 562),
           ('on', 523),
           ('their', 519),
           ('not', 478),
           ('have', 449),
           ('one', 410),
```

```
f=frequency.most_common(20)
word=[]
freq=[]
```

```
for r,p in f:
    word.append(r)
    freq.append(p)
index = np.arange(len(word))
plt.bar(index, freq)
plt.xlabel('Words', fontsize=15)
plt.ylabel('Frequency', fontsize=15)
plt.xticks(index, word, fontsize=10, rotation=90)
plt.title('Freqency Distribution')
plt.show()
```

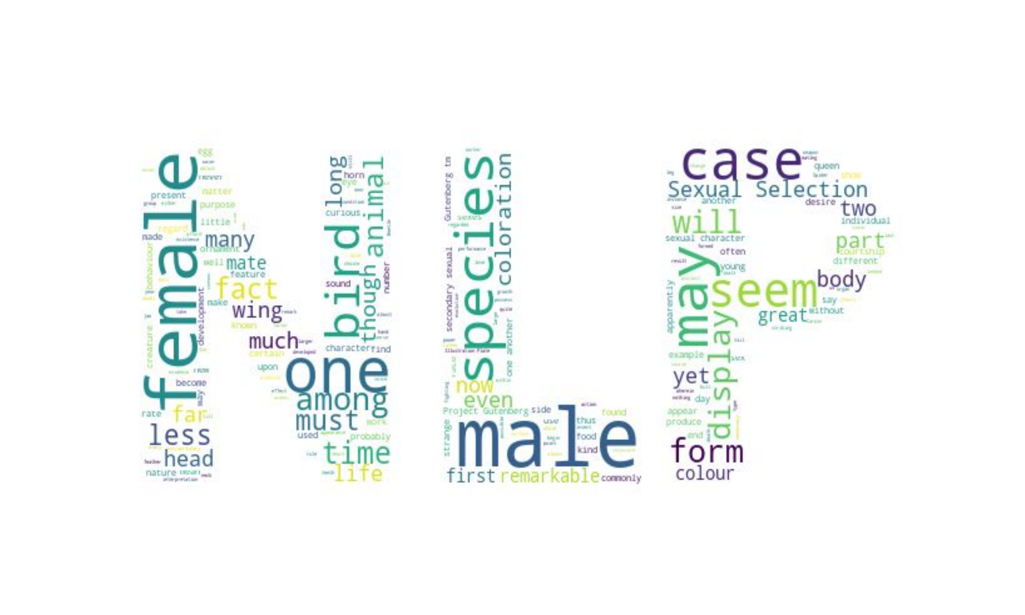


```
#stemming
stemmer= PorterStemmer()
for word in token:    stemmer.stem(word)

#lemmatization
lemmatizer=WordNetLemmatizer()
for word in token:
    lemmatizer.lemmatize(word)
s=' '
T=s.join(token)
```



```
In [73]: T

Out[73]: "project gutenberg 's the courtship of animals , by william plane pycraft this ebook is for the use of anyone anywhere in the
         united states and most other parts of the world at no cost and with almost no restrictions whatsoever . you may copy it , giv
         e it away or re-use it under the terms of the project gutenberg license included with this ebook or online at www.gutenberg.o
         rg . if you are not located in the united states , you 'll have to check the laws of the country where you are located before
         using this ebook . title : the courtship of animals author : william plane pycraft release date : october , [ ebook # ] langu
         age : english character set encoding : utf- *** start of this project gutenberg ebook the courtship of animals *** produced b
         y turgut dincer ( this file was produced from images made available by the internet archive ) hutchinson ’ s nature library t
         he courtship of animals [ illustration : plate . love-making . frontispiece . ] the courtship of animals by w. p. pycraft of
         the zoological department of the british museum : fellow of the zoological society of london ; associate of the linnean socie
         ty : member of the royal anthropological institute ; member of the british ornithologists ’ union ; hon . member of the ameri
         can ornithologists ’ union ; etc. , etc . author of “ a history of birds , ” “ the natural history museum , ” “ pads , paws a
         nd claws , ” “ the infancy of animals , ” etc. , etc. , etc . _with plates on art paper containing over illustrations_ _third
         edition_ london hutchinson & co. paternoster row i dedicate this volume to h. eliot howard whose observations of the courtshi
         p of birds recorded in his “ history of the british warblers ” constitute a beacon for all engaged in the study of animal beh
         aviour preface that “ one touch of nature which makes the whole world kin ” is surely nowhere more obvious than in the “ cour
         tship ” of animals . for the “ beasts that perish , ” no less than man himself , are stirred by the same emotions ; the fever
         of love runs as high in them as in ourselves ; and its modes of expression are not so different , though they may superficial
         ly appear to be so . the nature of these differences and their interpretation , it is the purpose of this book to set forth .
         charles darwin laid the foundation for the study of this phase of animal behaviour in his masterly work on the “ descent of m
```

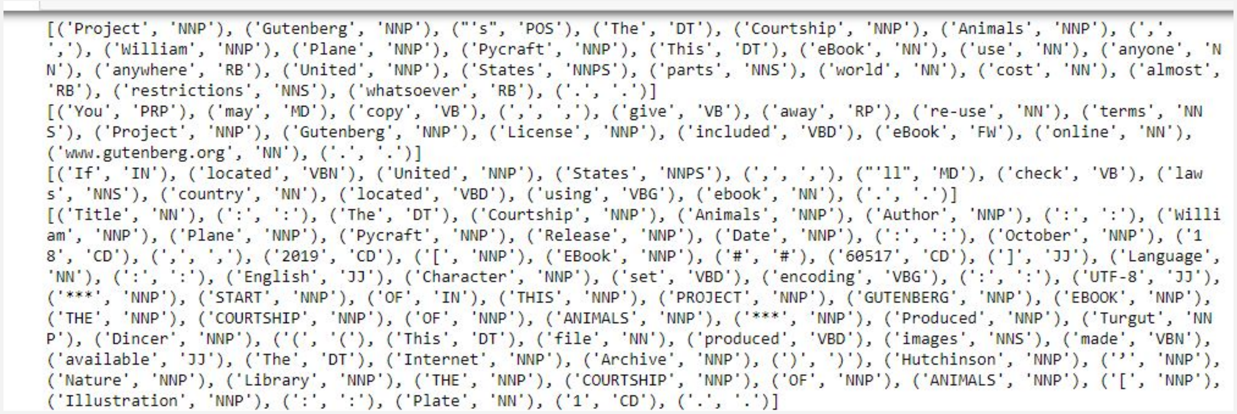
```
#initial wordcloud
wc = WordCloud(background_color="white",mask=mask)
wc.generate(T)
wc.to_file("wc1.png")
```



```
#wordcloud after removing stopwords
wc=WordCloud(background_color="white",mask=mask,stopwords=set(STOPWO
RDS))
wc.generate(T)
wc.to_file("wc2.png")
```

one
form
seem
fact
must
may
bird
will
display
animal
species
male
female
case
time
among
development
courtship
character
sexual character
secondary sexual
Sexual Selection
Project Gutenberg
less
life
present
coloration
colour
remarkable

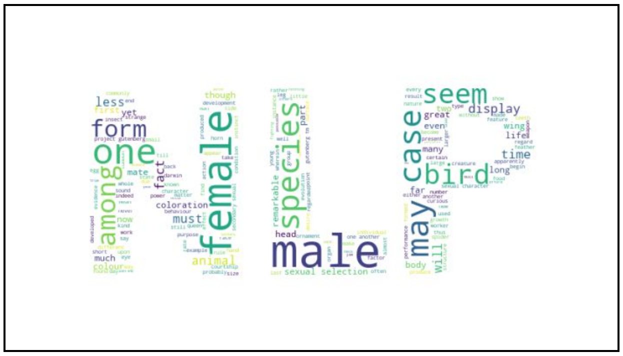
```
#pos tagging
stop_words = set(stopwords.words('english'))
tokenized = sent_tokenize(T)
for i in tokenized:
     wordsList = nltk.word_tokenize(i)

#removing stop words from wordList
  wordsList = [w for w in wordsList if not w in stop_words]
  tagged = nltk.pos_tag(wordsList)
```



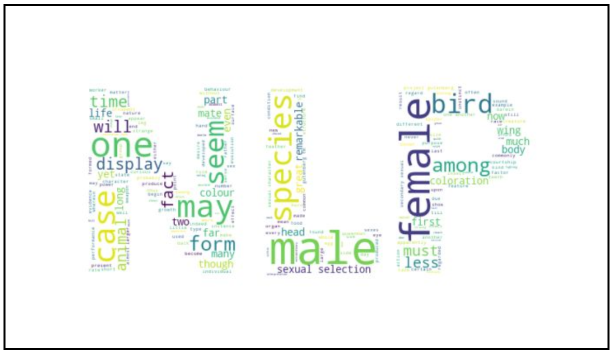
```
[('Project', 'NNP'), ('Gutenberg', 'NNP'), ("'s", 'POS'), ('The', 'DT'), ('Courtship', 'NNP'), ('Animals', 'NNP'), (',',
','), ('William', 'NNP'), ('Plane', 'NNP'), ('Pycraft', 'NNP'), ('This', 'DT'), ('eBook', 'NN'), ('use', 'NN'), ('anyone', 'N
N'), ('anywhere', 'RB'), ('United', 'NNP'), ('States', 'NNPS'), ('parts', 'NNS'), ('world', 'NN'), ('cost', 'NN'), ('almost',
'RB'), ('restrictions', 'NNS'), ('whatsoever', 'RB'), ('.', '.')]
[('You', 'PRP'), ('may', 'MD'), ('copy', 'VB'), (',', ','), ('give', 'VB'), ('away', 'RP'), ('re-use', 'NN'), ('terms', 'NN
S'), ('Project', 'NNP'), ('Gutenberg', 'NNP'), ('License', 'NNP'), ('included', 'VBD'), ('eBook', 'FW'), ('online', 'NN'),
('www.gutenberg.org', 'NN'), ('.', '.')]
[('If', 'IN'), ('located', 'VBN'), ('United', 'NNP'), ('States', 'NNPS'), (',', ','), ("'ll", 'MD'), ('check', 'VB'), ('law
s', 'NNS'), ('country', 'NN'), ('located', 'VBD'), ('using', 'VBG'), ('ebook', 'NN'), ('.', '.')]
[('Title', 'NN'), (':', ':'), ('The', 'DT'), ('Courtship', 'NNP'), ('Animals', 'NNP'), ('Author', 'NNP'), (':', ':'), ('Willi
am', 'NNP'), ('Plane', 'NNP'), ('Pycraft', 'NNP'), ('Release', 'NNP'), ('Date', 'NNP'), (':', ':'), ('October', 'NNP'), ('1
8', 'CD'), (',', ','), ('2019', 'CD'), ('[', 'NNP'), ('EBook', 'NNP'), ('#', '#'), ('60517', 'CD'), (']', 'JJ'), ('Language',
'NN'), (':', ':'), ('English', 'JJ'), ('Character', 'NNP'), ('set', 'VBD'), ('encoding', 'VBG'), (':', ':'), ('UTF-8', 'JJ'),
('***', 'NNP'), ('START', 'NNP'), ('OF', 'IN'), ('THIS', 'NNP'), ('PROJECT', 'NNP'), ('GUTENBERG', 'NNP'), ('EBOOK', 'NNP'),
('THE', 'NNP'), ('COURTSHIP', 'NNP'), ('OF', 'NNP'), ('ANIMALS', 'NNP'), ('***', 'NNP'), ('Produced', 'NNP'), ('Turgut', 'NN
P'), ('Dincer', 'NNP'), ('(', '('), ('This', 'DT'), ('file', 'NN'), ('produced', 'VBD'), ('images', 'NNS'), ('made', 'VBN'),
('available', 'JJ'), ('The', 'DT'), ('Internet', 'NNP'), ('Archive', 'NNP'), (')', ')'), ('Hutchinson', 'NNP'), ('’', 'NNP'),
('Nature', 'NNP'), ('Library', 'NNP'), ('THE', 'NNP'), ('COURTSHIP', 'NNP'), ('OF', 'NNP'), ('ANIMALS', 'NNP'), ('[', 'NNP'),
('Illustration', 'NNP'), (':', ':'), ('Plate', 'NN'), ('1', 'CD'), ('.', '.')]
```

A **word cloud** is a popular visualization of words typically associated with keywords and text data. They are most commonly used to highlight popular or trending terms based on frequency of use and prominence.
It is a collection, or cluster, of words depicted in different sizes. The bigger and bolder the word appears, the more often it’s mentioned within a given text and the more important it is.

Following are the wordclouds for the given dataset:



Wc1



Wc2

Wc1- wordcloud with stopwords
Wc2- wordcloud without stopwords

There is a difference between the wordclouds due to change in the wordset of the text , because of the removal of stopwords from the wordset.
Stop words are generally the most common words in a language. These stopwords might not add much value to the meaning of the document.For eg : *the*, *is*, *at*, *which*, and *on .*

These stopwords are removed from the given text so that more focus can be given to those words which define the meaning of the text.

Therefore , there is a slight difference between the wordclouds as there is a change in size of some words. This is so because the size depends on the frequency and prominence of words and stopwords had a considerable frequency.

RELATIONSHIP BETWEEN WORD LENGTH AND FREQUENCY:

According to Zipf's law, the frequency of a given word is inversely proportional to its length.

Following table shows some words and their frequency taken from the data:

WORD	FREQUENCY	WORD	FREQUENCY	WORD	FREQUENCY
the	8565	males	312	selection	181
of	5213	females	278	remarkable	105
and	2438	species	264	development	66

From the above table, we can conclude that it follows Zipf's law to some extent.