Assignment1_Porime

October 10, 2024

Assignment #1 - Sub Corpora 2 (Transcript of a YouTube documentary on Jeffery Dahmer - https://www.youtube.com/watch?v=Y1EWmrzD2Mk)

[8]:
```
#loading packages
import nltk
import numpy
import matplotlib

#loading data
with open("/Users/catalinaporime/Desktop/SDA250/Assignment_1/Dahmer.txt", "r",␣
 ↪encoding = "utf8") as f:
    Dahmer = f.read()

#tokenizing the data
DahmerTokens = nltk.word_tokenize(Dahmer)

#printing text (commented due to length)
#print(Dahmer)


#Q1:
print("Question #1")
lengthDahmer = len(DahmerTokens)
print(lengthDahmer)
print("The number of words in the Jeffrey Dahmer documentary on YouTube:",␣
 ↪lengthDahmer)
print()

#Q2:
print("Question #2")
def lexical_diversity(DahmerTokens):
    return len(set(DahmerTokens)) / len(DahmerTokens)
lex_div = lexical_diversity(DahmerTokens)
print(lex_div)
print("The lexical diversity of the text is:", lex_div)
print()

#Q3:
```

```
print("Question #3")
from nltk.probability import FreqDist
fdist1 = FreqDist(DahmerTokens)
fdist1.most_common(13)
print("These are the 10 most common words in the text and their count:")
print(fdist1.most_common(13))
print()

#Q4:
print("Question #4")
dahmerLong = [word for word in DahmerTokens if len(word) >= 10]
longDist = FreqDist(dahmerLong)
print("These are the words that are at least 10 characters long and their count:
 ↪", longDist.most_common())
print()

#Q5:
print("Question #5")
from nltk.tokenize import sent_tokenize, word_tokenize
sentences = sent_tokenize(Dahmer)
longest_sentence = max(sentences, key=lambda sentence:␣
 ↪len(word_tokenize(sentence)))
word_count = len(word_tokenize(longest_sentence))
print("Longest sentence:", longest_sentence)
print("Number of words:", word_count)
print()

#Q6:
print("Question #6")
from nltk.stem.snowball import SnowballStemmer
stemmer = SnowballStemmer ("english")
long1 = word_tokenize(longest_sentence)
stemmed_word = [stemmer.stem(word) for word in long1]
print("This is the stemmed version of the longest sentence:")
print(stemmed_word)
```

```
Question #1
8084
The number of words in the Jeffrey Dahmer documentary on YouTube: 8084

Question #2
0.19594260267194458
The lexical diversity of the text is: 0.19594260267194458

Question #3
These are the 10 most common words in the text and their count:
[(',', 420), ('.', 388), ('the', 301), ('of', 199), ('to', 194), ('was', 189),
```

('he', 171), ('a', 165), ('and', 144), ('that', 134), ('in', 132), ('-', 127), ('Dahmer', 115)]

Question #4
These are the words that are at least 10 characters long and their count:
[('television', 6), ('themselves', 5), ('confession', 5), ('necrophilia', 4), ('detectives', 4), ('cannibalism', 3), ('constantly', 3), ('completely', 3), ('interested', 3), ('remembered', 3), ('grandmother', 3), ('14-year-old', 3), ('photographs', 3), ('essentially', 3), ('immediately', 3), ('one-bedroom', 2), ('investigate', 2), ('31-year-old', 2), ('strangling', 2), ('psychological', 2), ('classmates', 2), ('collection', 2), ('apparently', 2), ('neighborhood', 2), ('hitchhiker', 2), ('ultimately', 2), ('eventually', 2), ('dismembered', 2), ('homosexual', 2), ('opportunity', 2), ('predominantly', 2), ('atrocities', 2), ('13-year-old', 2), ('unconscious', 2), ('necrophiliac', 2), ('refrigerator', 2), ('potentially', 2), ('circumstances', 2), ('incredible', 2), ('controlling', 2), ('assistants', 2), ('whispering', 1), ('synonymous', 1), ('revelations', 1), ('journalist', 1), ('antiseptic', 1), ('throughout', 1), ('playground', 1), ('ostracized', 1), ('surrounding', 1), ('progressed', 1), ('fascinated', 1), ('decapitating', 1), ('disturbing', 1), ('intoxicated', 1), ('experiences', 1), ('ruminating', 1), ('fantasizing', 1), ('19-year-old', 1), ('abandonment', 1), ('pulverized', 1), ('sledgehammer', 1), ('experienced', 1), ('accustomed', 1), ('apprehended', 1), ('relentless', 1), ('Culpability', 1), ('compromise', 1), ('desecrating', 1), ('University', 1), ('colleagues', 1), ('dependency', 1), ('21-year-old', 1), ('discharged', 1), ('approaching', 1), ('frequenting', 1), ('bathhouses', 1), ('stigmatized', 1), ('underground', 1), ('companionship', 1), ('conversations', 1), ('prospective', 1), ('supportive', 1), ('presumably', 1), ('Dissolving', 1), ('identifying', 1), ('committing', 1), ('complaining', 1), ('benzodiazepines', 1), ('sentencing', 1), ('unrelenting', 1), ('unbeknownst', 1), ('authorities', 1), ('29-year-old', 1), ('dismembering', 1), ('constructing', 1), ('diabolical', 1), ('performing', 1), ('sacrificial', 1), ('approached', 1), ('cannibalize', 1), ('metaphorical', 1), ('incorporate', 1), ('experimenting', 1), ('semi-conscious', 1), ('subservient', 1), ('attempting', 1), ('incredibly', 1), ('unpleasant', 1), ('personality', 1), ('manufacturing', 1), ('Midwestern', 1), ('collecting', 1), ('cannibalizing', 1), ('successful', 1), ('African-American', 1), ('infrequently', 1), ('agonizingly', 1), ('Sinthasomphone', 1), ('attractive', 1), ('remarkable', 1), ('characteristic', 1), ('psychopaths', 1), ('extraordinarily', 1), ('high-pressure', 1), ('decomposing', 1), ('minorities', 1), ('continuous', 1), ('fulfillment', 1), ('insatiable', 1), ('dismemberment', 1), ('speculation', 1), ('struggling', 1), ('officially', 1), ('resistance', 1), ('investigation', 1), ('unbelievable', 1), ('fantastical', 1), ('international', 1), ('incredulous', 1), ('everything', 1), ('confessions', 1), ('involvement', 1), ('interesting', 1), ('differentiate', 1), ('investigators', 1), ('preliminary', 1), ('appearance', 1), ('good-looking', 1), ('responsible', 1), ('psychiatry', 1), ('definition', 1), ('paraphilias', 1), ('implications', 1), ('consultant', 1), ('necrophiliacs', 1), ('accomplish', 1), ('necrophilous', 1), ('specialist', 1), ('psychiatrists', 1), ('psychiatric', 1), ('traditional', 1), ('imprisonment', 1), ('especially', 1), ('hoodwinked', 1), ('Correctional', 1), ('disconcerting', 1), ('respectful',

```
1), ('forgiveness', 1), ('Christopher', 1), ('bludgeoned', 1), ('demolished',
1), ('impossible', 1), ('Ultimately', 1), ('inexplicably', 1), ('absolutely',
1), ('perversions', 1), ('nightmares', 1), ('unquestionably', 1)]

Question #5
Longest sentence: - And then when the guy said he wanted to leave, Dahmer
clubbed him on the back of the head with a barbell
and then strangled him, then ultimately disposed of the body,
removed all the flesh, and eventually dissolved it in acid
and pulverized the bones with a sledgehammer.
Number of words: 55

Question #6
This is the stemmed version of the longest sentence:
['-', 'and', 'then', 'when', 'the', 'guy', 'said', 'he', 'want', 'to', 'leav',
',', 'dahmer', 'club', 'him', 'on', 'the', 'back', 'of', 'the', 'head', 'with',
'a', 'barbel', 'and', 'then', 'strangl', 'him', ',', 'then', 'ultim', 'dispos',
'of', 'the', 'bodi', ',', 'remov', 'all', 'the', 'flesh', ',', 'and', 'eventu',
'dissolv', 'it', 'in', 'acid', 'and', 'pulver', 'the', 'bone', 'with', 'a',
'sledgehamm', '.']
```

[ ]: