assignment1Emma

October 10, 2024

Assignment #1 - Sub Corpora 3 (A Reddit Post that discusses: "What Do You Think of Jeffery Dahmer?" - https://www.reddit.com/r/AskALiberal/comments/xr64cy/what_do_you_think_of_jeffery_dahmer/

[37]:
```
import nltk
nltk.download()

from nltk.book import *
import numpy
import matplotlib

import os
from nltk.corpus import PlaintextCorpusReader
corpus_root = "C:/Users/emmad/SDA250_labs/Notebook/assignment1/redditData/
 ↪Dahmer_Reddit.txt"
corpus_root

# Question 1
print("Question 1")
with open(corpus_root, "r", encoding = "utf8") as f:
    Dahmer = f.read()
DahmerTokens = nltk.word_tokenize(Dahmer)
lengthDahmer = len(DahmerTokens)

print("The number of words in the 'Dahmer Reddit Post' is:" ,lengthDahmer)
print()
print()

# Question 2
print("Question 2")
def lexical_diversity(DahmerTokens):
    return len(set(DahmerTokens)) / len(DahmerTokens)
lexical_diversity(DahmerTokens)

print("The lexical diversity in the 'Dahmer Reddit Post' is:"␣
 ↪,lexical_diversity(DahmerTokens))
print()
print()
```

```
print("Question 3")

# Question 3
fdist3 = FreqDist(DahmerTokens)
fdist3.most_common(20)
#excluding punctuation and '2y' because it's not a proper word
print("Top 10 most frequent words and their counts in the 'Dahmer Reddit Post'␣
 ↪are:" ,fdist3.most_common(20))

print("This excludes any punctuation and the word'2y' becasue it is not a␣
 ↪proper word.")
print()
print()

print("Question 4")

# Question 4
Dahmerlong = [word for word in DahmerTokens if len(word) >= 10]

longDist = FreqDist(Dahmerlong)

print("All words  in the 'Dahmer Reddit Post' that are at least 10 characters␣
 ↪long and their counts inclide the following:" ,longDist.most_common())
print()
print()

print("Question 5")

# 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("The longest sentence in the 'Dahmer Reddit Post' is:" ,longest_sentence)
print("word count:" ,word_count)
print()
print()


print("Question 6")

#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("A stemmed version of the longest sentence is:" ,stemmed_word)
```

```
showing info https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/index.xml
Question 1
The number of words in the 'Dahmer Reddit Post' is: 9356


Question 2
The lexical diversity in the 'Dahmer Reddit Post' is: 0.1974134245404019


Question 3
Top 10 most frequent words and their counts in the 'Dahmer Reddit Post' are:
[('.', 324), ('¥', 263), (',', 262), ('Share', 250), ('the', 246), ('to', 223),
('of', 186), ('a', 164), ('and', 155), ('I', 136), ('ago', 133), ('2y', 131),
('Downvote', 125), ('Upvote', 124), ('that', 115), ('he', 101), ('is', 98),
('was', 93), ('in', 93), ('for', 65)]
This excludes any punctuation and the word'2y' becasue it is not a proper word.


Question 4
All words  in the 'Dahmer Reddit Post' that are at least 10 characters long and
their counts inclide the following: [('Progressive', 12), ('Libertarian', 11),
('conditions', 7), ('Independent', 6), ('rehabilitate', 6), ('understand', 5),
('u/almightywhacko', 5), ('almightywhacko', 5), ('consequences', 5),
('government', 5), ('Sinthasomphone', 5), ('problematic', 4), ('u/tie-
dyed_dolphin', 4), ('tie-dyed_dolphin', 4), ('confinement', 4), ('committing',
4), ('spidersinterweb', 3), ('Christopher', 3), ('rehabilitated', 3),
('u/Indrigotheir', 3), ('Indrigotheir', 3), ('effectively', 3), ('individual',
3), ('justification', 3), ('inevitably', 3), ('reasonably', 3),
('realistically', 3), ('department', 3), ('automatically', 2), ('recognized',
2), ('acceptable', 2), ('u/carlse20', 2), ('absolutely', 2), ('Hot_Dog_Cobbler',
2), ('particular', 2), ('wrongfully', 2), ('intentional', 2), ('definitely', 2),
('premeditated', 2), ('accidental', 2), ('u/Barbados_slim12', 2),
('Barbados_slim12', 2), ('retribution', 2), ('reasonable', 2), ('sanctioned',
2), ('predisposition', 2), ('difference', 2), ('rehabilitation', 2),
('rehabilitative', 2), ('u/Kerplonk', 2), ('incarcerated', 2), ('incarceration',
2), ('individuals', 2), ('afflictions', 2), ('epicgrilledchees', 2),
('areyouseriousdotard', 2), ('exceptions', 2), ('crossroads', 2), ('Democratic',
2), ('segregated', 2), ('meticulously', 2), ('ultimately', 2), ('civilization',
2), ('increasingly', 2), ('completely', 2), ('compromised', 2), ('potentially',
2), ('likelihood', 2), ('reprehensible', 2), ('compulsive', 2),
('dysregulation', 2), ('unsupervised', 2), ('paramedics', 2),
('u/AutoModerator', 1), ('AutoModerator', 1), ('originally', 1), ('moderators',
1), ('UpvoteVote', 1), ('NicklAAAAs', 1), ('alcoholism', 1), ('needlessly', 1),
('disappointed', 1), ('PM_ME_YOUR_DARKNESS', 1), ('conversation', 1),
('pathological', 1), ('questionable', 1), ('perpetrated', 1), ('disgusting', 1),
```

('tysontysontyson1', 1), ('collapsingrebel', 1), ('cannibalistic', 1),
('commitment', 1), ('compulsion', 1), ('crazy-bisquit', 1), ('intentionally',
1), ('imprisoned', 1), ('confessing', 1), ('u/imveganbro', 1), ('imveganbro',
1), ('u/ImInOverMyHead95', 1), ('ImInOverMyHead95', 1), ('everywhere', 1),
('sympathetic', 1), ('perspective', 1), ('lyssthebitchcalore', 1),
('disturbances', 1), ('remorseful', 1), ('encountered', 1),
('//youtu.be/HP5SQDdcgDk', 1), ('childhoods', 1), ('sociopathic', 1),
('preferential', 1), ('benefitted', 1), ('carelessly', 1), ('regardless', 1),
('restitution', 1), ('manslaughter', 1), ('personally', 1), ('alternative', 1),
('essentially', 1), ('cosmicnitwit', 1), ('u/Kitchen_Agency4375', 1),
('Kitchen_Agency4375', 1), ('dehumanising', 1), ('punishments', 1),
('justifying', 1), ('prisoners/people', 1), ('legislature', 1), ('republicans',
1), ('unconstitutional', 1), ('inadequacy', 1), ('inaccuracy', 1),
('identified', 1), ('preferable', 1), ('administration', 1), ('biological', 1),
('biological/chemical', 1), ('incredibly', 1), ('SoÉ.thereÕs', 1),
('defensively', 1), ('considered', 1), ('permanently', 1), ('propensity', 1),
('beforehand', 1), ('proportional', 1), ('dryadbride', 1), ('Late-Bar-8498', 1),
('interesting', 1), ('Regardless', 1), ('possibility', 1), ('re-offending', 1),
('discourage', 1), ('clarification', 1), ('eventually', 1), ('circulation', 1),
('ÒdisappearÓ', 1), ('psychopathy', 1), ('circulationÓ', 1), ('insinuating', 1),
('euphemisms', 1), ('BlueCollarBeagle', 1), ('Mississippi', 1),
('u/ExplorersxMuse', 1), ('ExplorersxMuse', 1), ('parameters', 1),
('conceivable', 1), ('whatsoever', 1), ('everything', 1), ('miscarriage', 1),
('comsumption', 1), ('institutions', 1), ('population', 1), ('preserving', 1),
('displaying', 1), ('accountable', 1), ('LoopyMercutio', 1), ('Unfortunately',
1), ('grammanarchy', 1), ('exonerated', 1), ('technicality', 1), ('exceptional',
1), ('RandomGrasspass', 1), ('u/ill-independent', 1), ('ill-independent', 1),
('euthanasia', 1), ('interpersonal', 1), ('implication', 1), ('coexisting', 1),
('imprisonment', 1), ('u/Stormlight1984', 1), ('Stormlight1984', 1),
('immediately', 1), ('Understanding', 1), ('laws/rules/policies', 1),
('existentially', 1), ('individualÕs', 1), ('fuckedupness', 1), ('illegalize-
it', 1), ('girlfriend_pregnant', 1), ('u/Scalage89', 1), ('retaliation', 1),
('prevention', 1), ('noticeable', 1), ('u/Heyoteyo', 1), ('Conservative', 1),
('rehabilitatableÉ', 1), ('participant', 1), ('fundamental', 1), ('perpetuity',
1), ('healthcare', 1), ('thoroughly', 1), ('determined', 1), ('viscerally', 1),
('u/TarnishedVictory', 1), ('TarnishedVictory', 1), ('musicalpants999', 1),
('Complicated', 1), ('surprising', 1), ('u/onikaizoku11', 1), ('onikaizoku11',
1), ('documentaries', 1), ('scientific', 1), ('individually', 1), ('unassisted',
1), ('body-specifically', 1), ('brain-should', 1), ('occurrences', 1),
('proclivities', 1), ('atrocities', 1), ('vulnerable', 1), ('betterment', 1),
('u/OpenMindTulsaBill', 1), ('OpenMindTulsaBill', 1), ('AskALiberal', 1),
('devastated', 1), ('philosophies', 1), ('introspection', 1), ('bobbyjames1986',
1), ('TheTarkovskyParadigm', 1), ('trilobright', 1), ('distinction', 1),
('misleading', 1), ('criminally', 1), ('themselves', 1), ('necessarily', 1),
('facilitated', 1), ('u/SwagLord5002', 1), ('SwagLord5002', 1), ('predicament',
1), ('considering', 1), ('reoriented', 1), ('rehabilitating', 1), ('reoffended',
1), ('conviction', 1), ('eye-for-an-eye', 1), ('executioner', 1), ('clinically',
1), ('developmentally', 1), ('singlehandedly', 1), ('authoritarian', 1),

('fearmongering', 1), ('especially', 1), ('admittedly', 1), ('nonexistent', 1), ('reoffending', 1), ('negligible', 1), ('Personally', 1), ('metaphorical', 1), ('therapeutic', 1), ('impulsive/compulsive', 1), ('redirecting', 1), ('subsequent', 1), ('combination', 1), ('obsessive-compulsive/impulsive', 1), ('internally', 1), ('fantasizing', 1), ('complicated', 1), ('frequently', 1), ('characterized', 1), ('disproportionate', 1), ('percentage', 1), ('psychotherapy', 1), ('previously', 1), ('conspiracies', 1), ('virulently', 1), ('Additionally', 1), ('supposedly', 1), ('coincidence', 1), ('universally', 1), ('not_a_flying_toy_', 1), ('Homegrownscientist', 1), ('homophobic', 1), ('24_Elsinore', 1), ('equivalent', 1), ('hydrochloric', 1), ('authorities', 1), ('threatening', 1), ('semi-naked', 1), ('masterminds', 1), ('QuixoticMarten', 1), ('deliberate', 1), ('adjustments', 1), ('enforcement', 1), ('intimidated', 1), ('BenMullen2', 1), ('prescribed', 1), ('failedtalkshowhost', 1), ('danceswithronin', 1), ('u/Tron_1981', 1), ('demographic', 1), ('Just_a_reddit_duck', 1), ('NimishApte', 1)]

Question 5
The longest sentence in the 'Dahmer Reddit Post' is: On the other hand, I've come to believe that the prison system itself is so innately flawed compared to other countries, that we would never be in the predicament of even considering having to use the death penalty if we reoriented our prisons more towards rehabilitating criminals rather than punishing them, which is not to say that I think that the crimes that many criminals commit are acceptable by any metric, but more that the system inevitably fucks over people who could otherwise reasonably be released back into civilized society (namely those with drug abuse charges) and ultimately puts them in a position where the state is forced to divide between whether or not they should be given a life sentence or a death sentence when realistically, if they had focused on trying to rehabilitate the person instead of punishing them, they may have never reoffended or committed crimes more grievous than their original conviction.
word count: 165

Question 6
A stemmed version of the longest sentence is: ['on', 'the', 'other', 'hand', ',', 'i', 've', 'come', 'to', 'believ', 'that', 'the', 'prison', 'system', 'itself', 'is', 'so', 'innat', 'flaw', 'compar', 'to', 'other', 'countri', ',', 'that', 'we', 'would', 'never', 'be', 'in', 'the', 'predica', 'of', 'even', 'consid', 'have', 'to', 'use', 'the', 'death', 'penalti', 'if', 'we', 'reorient', 'our', 'prison', 'more', 'toward', 'rehabilit', 'crimin', 'rather', 'than', 'punish', 'them', ',', 'which', 'is', 'not', 'to', 'say', 'that', 'i', 'think', 'that', 'the', 'crime', 'that', 'mani', 'crimin', 'commit', 'are', 'accept', 'by', 'ani', 'metric', ',', 'but', 'more', 'that', 'the', 'system', 'inevit', 'fuck', 'over', 'peopl', 'who', 'could', 'otherwis', 'reason', 'be', 'releas', 'back', 'into', 'civil', 'societi', '(', 'name', 'those', 'with', 'drug', 'abus', 'charg', ')', 'and', 'ultim', 'put', 'them', 'in', 'a', 'posit', 'where', 'the', 'state', 'is', 'forc', 'to', 'divid', 'between', 'whether',

```
'or', 'not', 'they', 'should', 'be', 'given', 'a', 'life', 'sentenc', 'or', 'a',
'death', 'sentenc', 'when', 'realist', ',', 'if', 'they', 'had', 'focus', 'on',
'tri', 'to', 'rehabilit', 'the', 'person', 'instead', 'of', 'punish', 'them',
',', 'they', 'may', 'have', 'never', 'reoffend', 'or', 'commit', 'crime',
'more', 'grievous', 'than', 'their', 'origin', 'convict', '.']
```

[ ]: