Basic Text Processing with NLTK Tutorial

Uzair Ahmad

In this tutorial, we will explore basic text processing using the Natural Language Toolkit (NLTK) in Python. We will cover searching in text, counting vocabulary, and basic text statistics.

Note: Before proceeding, make sure you have NLTK installed (`pip install nltk`) and download the necessary resources using `nltk.download('book')` .

In [1]:

```
!pip install nltk
```

```
Requirement already satisfied: nltk in c:\users\admin\anaconda3\envs\school_desktop\l
ib\site-packages (3.8.1)
Requirement already satisfied: click in c:\users\admin\anaconda3\envs\school_desktop
\lib\site-packages (from nltk) (8.0.4)
Requirement already satisfied: joblib in c:\users\admin\anaconda3\envs\school_desktop
\lib\site-packages (from nltk) (1.2.0)
Requirement already satisfied: regex>=2021.8.3 in c:\users\admin\anaconda3\envs\schoo
l_desktop\lib\site-packages (from nltk) (2022.7.9)
Requirement already satisfied: tqdm in c:\users\admin\anaconda3\envs\school_desktop\l
ib\site-packages (from nltk) (4.65.0)
Requirement already satisfied: colorama in c:\users\admin\anaconda3\envs\school_deskt
op\lib\site-packages (from click->nltk) (0.4.6)
```

In [2]:

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

```
[nltk_data] Downloading collection 'book'
[nltk_data]    |
[nltk_data]    | Downloading package abc to
[nltk_data]    |     C:\Users\Admin\AppData\Roaming\nltk_data...
[nltk_data]    |   Unzipping corpora\abc.zip.
[nltk_data]    | Downloading package brown to
[nltk_data]    |     C:\Users\Admin\AppData\Roaming\nltk_data...
[nltk_data]    |   Unzipping corpora\brown.zip.
[nltk_data]    | Downloading package chat80 to
[nltk_data]    |     C:\Users\Admin\AppData\Roaming\nltk_data...
[nltk_data]    |   Unzipping corpora\chat80.zip.
[nltk_data]    | Downloading package cmudict to
[nltk_data]    |     C:\Users\Admin\AppData\Roaming\nltk_data...
[nltk_data]    |   Unzipping corpora\cmudict.zip.
[nltk_data]    | Downloading package conll2000 to
[nltk_data]    |     C:\Users\Admin\AppData\Roaming\nltk_data...
[nltk_data]    |   Unzipping corpora\conll2000.zip.
[nltk_data]    | Downloading package conll2002 to
[nltk_data]    |     C:\Users\Admin\AppData\Roaming\nltk_data...
[nltk_data]    |   Unzipping corpora\conll2002.zip.
[nltk_data]    | Downloading package dependency_treebank to
[nltk_data]    |     C:\Users\Admin\AppData\Roaming\nltk_data...
[nltk_data]    |   Unzipping corpora\dependency_treebank.zip.
[nltk_data]    | Downloading package genesis to
[nltk_data]    |     C:\Users\Admin\AppData\Roaming\nltk_data...
[nltk_data]    |   Unzipping corpora\genesis.zip.
[nltk_data]    | Downloading package gutenberg to
[nltk_data]    |     C:\Users\Admin\AppData\Roaming\nltk_data...
[nltk_data]    |   Unzipping corpora\gutenberg.zip.
[nltk_data]    | Downloading package ieer to
[nltk_data]    |     C:\Users\Admin\AppData\Roaming\nltk_data...
[nltk_data]    |   Unzipping corpora\ieer.zip.
[nltk_data]    | Downloading package inaugural to
[nltk_data]    |     C:\Users\Admin\AppData\Roaming\nltk_data...
[nltk_data]    |   Unzipping corpora\inaugural.zip.
[nltk_data]    | Downloading package movie_reviews to
[nltk_data]    |     C:\Users\Admin\AppData\Roaming\nltk_data...
[nltk_data]    |   Unzipping corpora\movie_reviews.zip.
[nltk_data]    | Downloading package nps_chat to
[nltk_data]    |     C:\Users\Admin\AppData\Roaming\nltk_data...
[nltk_data]    |   Unzipping corpora\nps_chat.zip.
[nltk_data]    | Downloading package names to
[nltk_data]    |     C:\Users\Admin\AppData\Roaming\nltk_data...
[nltk_data]    |   Unzipping corpora\names.zip.
[nltk_data]    | Downloading package ppattach to
[nltk_data]    |     C:\Users\Admin\AppData\Roaming\nltk_data...
[nltk_data]    |   Unzipping corpora\ppattach.zip.
[nltk_data]    | Downloading package reuters to
[nltk_data]    |     C:\Users\Admin\AppData\Roaming\nltk_data...
[nltk_data]    | Downloading package senseval to
[nltk_data]    |     C:\Users\Admin\AppData\Roaming\nltk_data...
[nltk_data]    |   Unzipping corpora\senseval.zip.
[nltk_data]    | Downloading package state_union to
[nltk_data]    |     C:\Users\Admin\AppData\Roaming\nltk_data...
[nltk_data]    |   Unzipping corpora\state_union.zip.
[nltk_data]    | Downloading package stopwords to
[nltk_data]    |     C:\Users\Admin\AppData\Roaming\nltk_data...
[nltk_data]    |   Unzipping corpora\stopwords.zip.
[nltk_data]    | Downloading package swadesh to
[nltk_data]    |     C:\Users\Admin\AppData\Roaming\nltk_data...
```

```
[nltk_data]    |   Unzipping corpora\swadesh.zip.
[nltk_data]    | Downloading package timit to
[nltk_data]    |     C:\Users\Admin\AppData\Roaming\nltk_data...
[nltk_data]    |   Unzipping corpora\timit.zip.
[nltk_data]    | Downloading package treebank to
[nltk_data]    |     C:\Users\Admin\AppData\Roaming\nltk_data...
[nltk_data]    |   Unzipping corpora\treebank.zip.
[nltk_data]    | Downloading package toolbox to
[nltk_data]    |     C:\Users\Admin\AppData\Roaming\nltk_data...
[nltk_data]    |   Unzipping corpora\toolbox.zip.
[nltk_data]    | Downloading package udhr to
[nltk_data]    |     C:\Users\Admin\AppData\Roaming\nltk_data...
[nltk_data]    |   Unzipping corpora\udhr.zip.
[nltk_data]    | Downloading package udhr2 to
[nltk_data]    |     C:\Users\Admin\AppData\Roaming\nltk_data...
[nltk_data]    |   Unzipping corpora\udhr2.zip.
[nltk_data]    | Downloading package unicode_samples to
[nltk_data]    |     C:\Users\Admin\AppData\Roaming\nltk_data...
[nltk_data]    |   Unzipping corpora\unicode_samples.zip.
[nltk_data]    | Downloading package webtext to
[nltk_data]    |     C:\Users\Admin\AppData\Roaming\nltk_data...
[nltk_data]    |   Unzipping corpora\webtext.zip.
[nltk_data]    | Downloading package wordnet to
[nltk_data]    |     C:\Users\Admin\AppData\Roaming\nltk_data...
[nltk_data]    | Downloading package wordnet_ic to
[nltk_data]    |     C:\Users\Admin\AppData\Roaming\nltk_data...
[nltk_data]    |   Unzipping corpora\wordnet_ic.zip.
[nltk_data]    | Downloading package words to
[nltk_data]    |     C:\Users\Admin\AppData\Roaming\nltk_data...
[nltk_data]    |   Unzipping corpora\words.zip.
[nltk_data]    | Downloading package maxent_treebank_pos_tagger to
[nltk_data]    |     C:\Users\Admin\AppData\Roaming\nltk_data...
[nltk_data]    |   Unzipping taggers\maxent_treebank_pos_tagger.zip.
[nltk_data]    | Downloading package maxent_ne_chunker to
[nltk_data]    |     C:\Users\Admin\AppData\Roaming\nltk_data...
[nltk_data]    |   Unzipping chunkers\maxent_ne_chunker.zip.
[nltk_data]    | Downloading package universal_tagset to
[nltk_data]    |     C:\Users\Admin\AppData\Roaming\nltk_data...
[nltk_data]    |   Unzipping taggers\universal_tagset.zip.
[nltk_data]    | Downloading package punkt to
[nltk_data]    |     C:\Users\Admin\AppData\Roaming\nltk_data...
[nltk_data]    |   Unzipping tokenizers\punkt.zip.
[nltk_data]    | Downloading package book_grammars to
[nltk_data]    |     C:\Users\Admin\AppData\Roaming\nltk_data...
[nltk_data]    |   Unzipping grammars\book_grammars.zip.
[nltk_data]    | Downloading package city_database to
[nltk_data]    |     C:\Users\Admin\AppData\Roaming\nltk_data...
[nltk_data]    |   Unzipping corpora\city_database.zip.
[nltk_data]    | Downloading package tagsets to
[nltk_data]    |     C:\Users\Admin\AppData\Roaming\nltk_data...
[nltk_data]    |   Unzipping help\tagsets.zip.
[nltk_data]    | Downloading package panlex_swadesh to
[nltk_data]    |     C:\Users\Admin\AppData\Roaming\nltk_data...
[nltk_data]    | Downloading package averaged_perceptron_tagger to
[nltk_data]    |     C:\Users\Admin\AppData\Roaming\nltk_data...
[nltk_data]    |   Unzipping taggers\averaged_perceptron_tagger.zip.
[nltk_data]    |
[nltk_data]  Done downloading collection book
```

```
Out[2]: True
```

In [3]:
```python
from nltk.book import *
```

```
*** Introductory Examples for the NLTK Book ***
Loading text1, ..., text9 and sent1, ..., sent9
Type the name of the text or sentence to view it.
Type: 'texts()' or 'sents()' to list the materials.
text1: Moby Dick by Herman Melville 1851
text2: Sense and Sensibility by Jane Austen 1811
text3: The Book of Genesis
text4: Inaugural Address Corpus
text5: Chat Corpus
text6: Monty Python and the Holy Grail
text7: Wall Street Journal
text8: Personals Corpus
text9: The Man Who Was Thursday by G . K . Chesterton 1908
```

Searching in Text

Concordance, Similar, and Common Context Functions

To search for specific words and explore their context within a text, NLTK provides several useful functions.

- The `concordance` function shows every occurrence of a given word along with its context in the text.

In [4]:
```python
# Displaying the concordance of a word in text7
text7.concordance('S&P')
```

```
Displaying 25 of 35 matches:
r 's 500 stock-index futures pit once S&P 500 futures fell 20 index points -- t
 well *T*-1 . Late that afternoon the S&P 500 stock-index futures contract fell
xisting 30-minute , 12-point limit on S&P 500 stock-index futures trading -LRB-
ome the maximum one-day limit for the S&P 500 stock-index futures contract ; th
h slower growth the year before . The S&P index started *-1 sliding in price in
ket watchers anxious . Payouts on the S&P 500 stocks rose 10 % in 1988 , accord
ing a close watch on the yield on the S&P 500 . The figure is currently about 3
ock-market data . The last time 0 the S&P 500 yield dropped below 3 % *T*-1 was
971 and 1972 -- when the yield on the S&P 500 dropped below 3 % for at least tw
r instance , is forecasting growth in S&P 500 dividends of just under 5 % in 19
cally unchanged from 138 a year ago , S&P said 0 *T*-2 Wednesday . That followe
e year-earlier pace *T*-1 . While the S&P tally does n't measure the magnitude
turns of standard benchmarks like the S&P Not surprisingly , old-style money ma
 juggle portfolios so they mirror the S&P 500 . The indexers charge only a few
 who *T*-71 buy all 500 stocks in the S&P 500 often do n't even know what the c
n index arbitrage , the widget is the S&P 500 , and its price is constantly com
rtunity , someone who *T*-75 owns the S&P 500 widget in New York must sell it a
sell it and replace it with a cheaper S&P 500 widget in Chicago . If the money
78 to match or beat the return of the S&P 500 index , he is likely *-1 to remai
d 3 % *U* to the annual return of the S&P 500 . That represents a very thin ``
t money managers fail *-1 to beat the S&P 500 may contribute to the hysteria su
actual stocks that *T*-76 make up the S&P 500 is more cumbersome than * transac
in prices between the two markets for S&P 500 stocks -- usually it is large inv
. The bonds are rated *-1 single-A by S&P , according to the lead underwriter .
 Jeffrey Bowman , a vice president at S&P , which *T*-1 raised a warning flag f
```

- The `similar` function finds words in the text that appear in similar contexts to the specified word.

```
In [5]: # Finding similar words in text7
        text7.similar('growth')
```

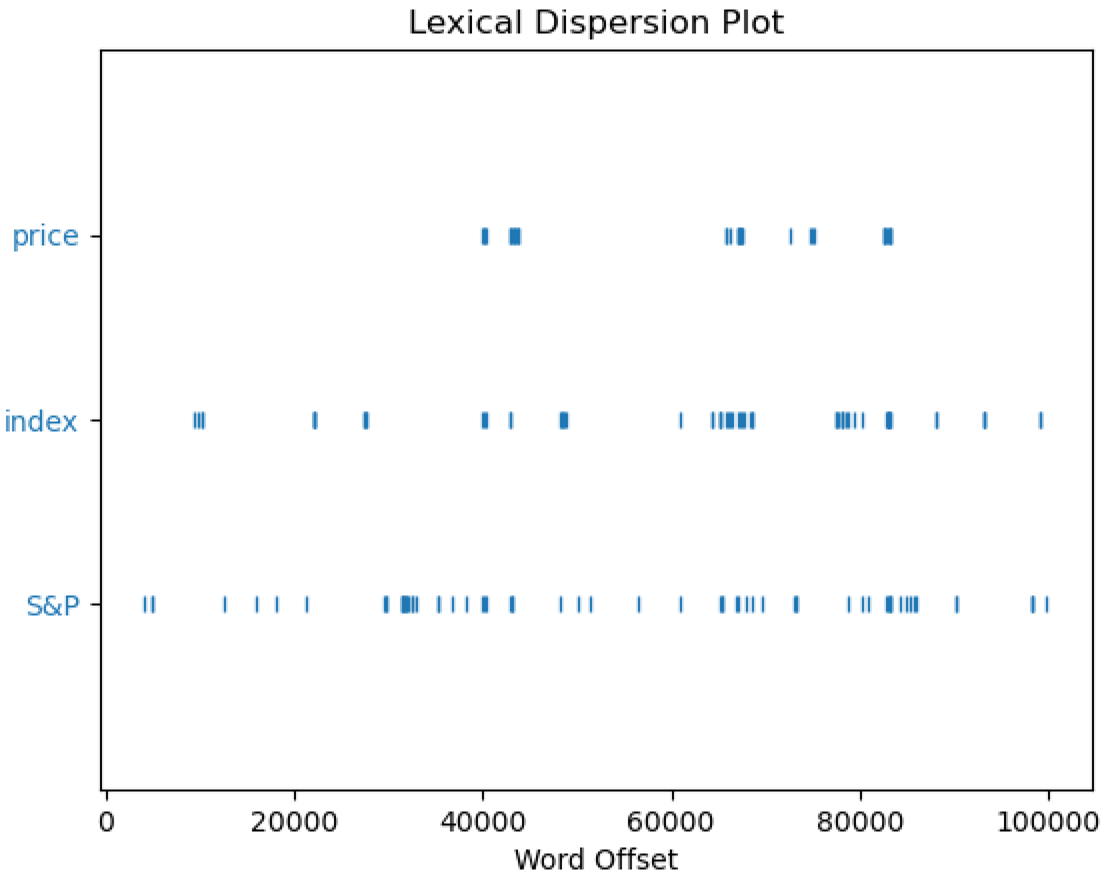
```
it institute billion government issue drop purchase stock chain order
amount case stocks strength language the board chairman fields form
```

```
In [6]: text4.similar('growth')
```

```
peace people government respect faith freedom country duty will
influence union prosperity right honor independence war support wisdom
labor conscience
```

- The `dispersion_plot` function generates a plot showing the location of words in the text.

```
In [8]: # Dispersion plot of words in a text
        text7.dispersion_plot(['S&P', 'index', 'price'])
```



Collocations

Working with collocations in NLTK involves identifying pairs or groups of words that often occur together. Collocations can provide valuable insights into the common patterns and associations within a given text. NLTK provides a module called nltk.collocations to work with collocations.

Here's a basic guide on how to work with collocations in NLTK:

In [9]:

```
from nltk import collocations
```

Create a Bigram Finder: Use BigramCollocationFinder to find bigrams (pairs of adjacent words) in the text.

In [10]:

```
bigram_finder = collocations.BigramCollocationFinder.from_words(text7)
```

Set Scoring Method: Choose a scoring method to identify collocations. One common scoring method is pmi (Pointwise Mutual Information).

In [11]:

```
bigram_finder.apply_freq_filter(5)  # Adjust the frequency filter as needed
bigram_measures = collocations.BigramAssocMeasures()
scored_bigrams = bigram_finder.score_ngrams(bigram_measures.pmi)
```

Print Top Collocations: Print the top collocations based on the chosen scoring method.

In [14]:

```
for bigram, score in scored_bigrams[:1000]:
    print(bigram, score)
```

```
('Burnham', 'Lambert') 14.29743218198336
('Merrill', 'Lynch') 14.29743218198336
('Tony', 'Lama') 14.29743218198336
('Lehman', 'Hutton') 14.034397776149568
('Hong', 'Kong') 13.81200535481312
('circuit', 'breaker') 13.81200535481312
('denying', 'wrongdoing') 13.81200535481312
('Morgan', 'Stanley') 13.589612933476673
('POP', 'Radio') 13.449435275428412
('Heritage', 'Media') 13.29743218198336
('Learning', 'Materials') 13.29743218198336
('Mississippi', 'River') 13.159928658233426
('Industrial', 'Average') 13.075039760646913
('Supreme', 'Court') 13.034397776149566
('Los', 'Angeles') 12.918920558729631
('Drexel', 'Burnham') 12.91892055872963
('Old', 'Guard') 12.882394682704518
('General', 'Motors') 12.812005354813117
('Waertsilae', 'Marine') 12.74489115895458
('Shearson', 'Lehman') 12.619360276870724
('current-carrying', 'capacity') 12.619360276870724
('Sea', 'Containers') 12.619360276870722
('loan', 'guarantees') 12.619360276870722
('fetal-tissue', 'transplants') 12.60143836887346
('San', 'Francisco') 12.371432763427139
('interstate', 'banking') 12.371432763427139
('United', 'States') 12.371432763427137
('Judge', 'Curry') 12.29743218198336
('St.', 'Louis') 12.253038062624906
('Scoring', 'High') 12.227042854091964
('Federal', 'Reserve') 12.227042854091962
('heating', 'oil') 12.227042854091962
('Dow', 'Jones') 12.223431600539582
('real', 'estate') 12.186400869594616
('Article', 'II') 12.159928658233424
('Commonwealth', 'Edison') 12.120127650175453
('lead', 'underwriter') 12.108398357593343
('Georgia', 'Gulf') 12.095798320813708
('line-item', 'veto') 12.095798320813708
('Jones', 'Industrial') 12.075039760646913
('Southeast', 'Asia') 11.975504087095999
('tentatively', 'priced') 11.918920558729631
('merchant', 'banking') 11.886005936256893
('light', 'trucks') 11.882394682704515
('Campbell', 'Soup') 11.864472774707256
('jointly', 'fined') 11.771363370315772
('unconstitutional', 'conditions') 11.712469681262206
('minimum', 'wage') 11.642080353370805
('England', 'Electric') 11.610371493643466
('alleged', 'violations') 11.503883059450787
('appropriations', 'clause') 11.490077259925755
('Dec.', '31') 11.339252357677989
('Wall', 'Street') 11.3319043992405
('closely', 'held') 11.320152258483443
('South', 'Carolina') 11.297432181983362
('South', 'Korea') 11.29743218198336
('Oct.', '13') 11.275064368954906
('Valley', 'Federal') 11.227042854091962
('Cray', 'Computer') 11.208427175924614
('test', 'scores') 11.159928658233424
```

```
('December', 'delivery') 11.13393344970048
('Reagan', 'administration') 11.127507180541048
('Greenville', 'High') 11.111565636672028
('Western', 'Union') 11.111565636672028
("'ve", 'got') 11.034397776149568
('Big', 'Board') 11.018008815610534
('Soviet', 'Union') 10.979115340648377
('Mercantile', 'Exchange') 10.975504087095999
('Cray', 'Research') 10.96400844825817
('general', 'manager') 10.93158697056579
('ended', 'Sept.') 10.918920558729631
('Justice', 'Department') 10.91892055872963
('National', 'Association') 10.905114759204599
('Random', 'House') 10.838000563346064
('Stock', 'Exchange') 10.805579085653685
('tender', 'offer') 10.80173701935929
('Transportation', 'Department') 10.797358578848717
('S&P', '500') 10.765711702539182
('First', 'Boston') 10.743401540814094
('auto', 'maker') 10.696528137393184
('Sept.', '30') 10.686259801939354
('card', 'holders') 10.674501831063182
('purchasing', 'managers') 10.660002261368067
('Mrs.', 'Ward') 10.645902064323217
('High', 'School') 10.642080353370805
('bankruptcy', 'court') 10.642080353370805
('both', 'sides') 10.642080353370805
('Mrs.', 'Yeargin') 10.63938875736454
('percentage', 'point') 10.623875757993215
('1987', 'crash') 10.619360276870724
('American', 'Express') 10.55327108641295
('White', 'House') 10.450977440236818
('manufacturing', 'sector') 10.394393911870447
('debt', 'ceiling') 10.358832726647504
('President', 'Bush') 10.35257373617582
('appropriations', 'bills') 10.31253907437357
('fourth', 'quarter') 10.311931751678474
('construction', 'spending') 10.286652343230116
('Says', '*ICH*-1') 10.26768483858931
('Commerce', 'Department') 10.242390818425397
('Exchange', 'composite') 10.20996934073302
('stock-index', 'futures') 10.201507761984825
('&', 'Poor') 10.193095522168624
('early', 'retirement') 10.193095522168624
('nine', 'months') 10.151754726787726
('mutual', 'funds') 10.138233587134108
('third', 'quarter') 10.127507180541048
('Standard', '&') 10.055591998418688
('index', 'arbitrage') 9.967283580291031
('six', 'months') 9.952445918564319
('&', 'Loan') 9.93006111633483
('work', 'force') 9.925873319371398
('we', "'ve") 9.841752698207172
('90', 'days') 9.831457717479292
('net', 'income') 9.80254502556405
('executive', 'officer') 9.779824949063968
('I', "'m") 9.773870225926348
('So', 'far') 9.744891158954584
('appropriations', 'bill') 9.73846488979515
('Bush', 'administration') 9.712469681262203
```

```
('chief', 'executive') 9.666827015706493
('pretax', 'profit') 9.645902064323217
('no', 'longer') 9.59145428030084
('third-quarter', 'net') 9.568423311645496
('operating', 'officer') 9.503883059450787
('fiscal', '1989') 9.490077259925757
('term', 'bonds') 9.473682821675087
('8', '3\\/4') 9.449435275428412
('Mrs.', 'Hills') 9.43945118685579
('federal', 'funding') 9.432361762069469
('York', 'Stock') 9.408408007168358
('dividend', 'growth') 9.39535860267262
('first', 'nine') 9.390541586374843
('vice', 'president') 9.390048441594484
('1', '3\\/4') 9.38574060011102
('junk', 'bond') 9.38286765848942
('financial', 'institutions') 9.379045947537012
('We', "'re") 9.376715431507824
('interest', 'rates') 9.335046473783137
('late', 'Tuesday') 9.330263574450731
('-LRB-', 'D.') 9.306477321586366
('shares', 'outstanding') 9.301043435535739
('Exchange', 'Commission') 9.297432181983359
('fiscal', '1990') 9.260209718027161
('New', 'Hampshire') 9.253038062624906
('New', 'Orleans') 9.253038062624906
('New', 'York') 9.253038062624906
('New', 'York-based') 9.253038062624906
('above', '50') 9.24780141425876
('operating', 'profit') 9.23086456504437
('senior', 'vice') 9.159928658233424
('other', 'hand') 9.138233587134108
('composite', 'trading') 9.098938028344278
('due', 'Nov.') 9.07092365217468
('Savings', '&') 9.055591998418688
('junk', 'bonds') 9.048661233823621
('financial', 'officer') 9.034397776149568
('Sales', 'rose') 9.025657926585408
('Containers', "'") 9.014498218711863
('Terms', 'were') 8.997308457414347
('years', 'ago') 8.985374331120058
('commercial', 'banks') 8.946934934899229
('net', 'loss') 8.946934934899227
('money', 'managers') 8.891439822307522
('Industries', 'Inc.') 8.849522433241281
('chief', 'operating') 8.831457717479292
('Nov.', '30') 8.80173701935929
('cash', 'income') 8.800778099389865
('&', 'Co') 8.778058022889782
('executive', 'vice') 8.77290553512418
('same', 'time') 8.754555082629217
('New', 'England') 8.729476106567894
('even', 'though') 8.67391644049281
('years', 'old') 8.648339343842487
('It', '*EXP*-1') 8.631549685878639
('model', 'year') 8.62840541647373
('bonds', 'due') 8.621240010088945
('suspension', ';') 8.588530885094071
('rather', 'than') 8.586938799178345
('recent', 'weeks') 8.57496615751227
```

```
('In', 'addition') 8.561209342607672
('pence', '-LRB-') 8.512928199053793
('&', 'Co.') 8.507204112596687
('we', "'re") 8.498344875909355
('so', 'far') 8.469613157366041
('money', 'manager') 8.458480415031417
('preferred', 'stock') 8.416420218200448
('common', 'shares') 8.413011867178536
('preferred', 'shares') 8.40795863945225
('how', 'much') 8.390541586374841
('Program', 'trading') 8.38642547790261
('futures', 'contracts') 8.379045947537012
('below', '50') 8.373332296342618
('well', 'below') 8.373332296342618
('chief', 'financial') 8.361972434178071
('next', 'week') 8.356873873609448
('Japanese', 'investments') 8.346030890309049
('U.S.', 'currency') 8.346030890309049
('If', 'you') 8.338074166480705
('his', 'departure') 8.321985416342397
('were', 'jointly') 8.319236552301708
('an', 'ounce') 8.315579528693618
('analysts', 'say') 8.291005912823929
('ca', "n't") 8.275064368954908
('wo', "n't") 8.275064368954904
('futures', 'contract') 8.23998190979946
('you', "'re") 8.217058765519342
('year', 'ending') 8.213367917194885
('country', 'funds') 8.199634131798248
('already', 'own') 8.19729551069791
('October', '1988') 8.196769842978162
('first', 'half') 8.18890772520519
('Among', 'other') 8.17475946315922
('program', 'trading') 8.173468636575775
('admitting', 'or') 8.171450528128643
('or', 'denying') 8.171450528128643
('net', 'cash') 8.105632680918287
('an', 'hour') 8.093187107357172
('Last', 'year') 8.08408490024992
('billion', 'yen') 8.08408490024992
('two', 'weeks') 8.058390631771365
('I', 'think') 8.05766319192694
('program', 'trades') 8.045713089377402
('program', 'traders') 8.0457130893774
('130', 'million') 8.038159694945765
('Mr.', 'Hahn') 7.998224163596081
('30', 'days') 7.977308583942747
('at', 'par') 7.968308585691794
('at', '99') 7.968308585691792
('Mr.', 'Baum') 7.898688490045167
('shares', 'closed') 7.879579667097461
('Mr.', 'Trudeau') 7.846221070151033
('last', 'month') 7.838000563346062
('year', 'earlier') 7.83003927764338
('Mr.', 'Nixon') 7.805579085653687
('Mr.', 'Reupke') 7.805579085653685
('less', 'than') 7.805579085653685
('an', 'interview') 7.801006355863862
('Mr.', 'Lane') 7.779106874292493
('Mr.', 'Spiegel') 7.779106874292493
```

```
('three', 'months') 7.736717227508883
('first', 'time') 7.735189757762287
('holding', 'company') 7.734495987592203
('compared', 'with') 7.733663903531328
('more', 'than') 7.707603926090387
('at', 'least') 7.705274179858
('managers', "'") 7.692570123824501
("n't", 'elaborate') 7.69010186823375
('will', 'receive') 7.677579034592192
('last', 'week') 7.66807556190375
('shall', 'be') 7.658200018734082
('Mr.', 'Steinberg') 7.653575992208637
('Mr.', 'McGovern') 7.653575992208635
('Group', 'Inc.') 7.640935811429864
('Mr.', 'Phelan') 7.621154514516258
('Mr.', 'Simmons') 7.609181872850181
('did', "n't") 7.606270277430278
('does', "n't") 7.56457098614989
('A', 'spokesman') 7.552598344483814
('last', 'year') 7.528869742922815
('he', 'believes') 7.510835820092552
('small', 'investors') 7.50457682962087
('five', 'years') 7.5008517315199335
('these', 'days') 7.487503316261929
('will', 'remain') 7.484933956649797
("n't", 'disclosed') 7.444989370397218
("'", 'report') 7.429535717990705
('Mr.', 'Bernstein') 7.4165367949077865
('Mr.', 'Dinkins') 7.4165367949077865
('this', 'week') 7.401901448539389
('Mr.', 'Wilder') 7.390541586374841
('owns', 'about') 7.347897248966349
('was', 'named') 7.340732123531307
('days', ';') 7.3364373420709335
('51', '%') 7.3330335497801755
('this', 'fall') 7.3181907421501595
('their', 'own') 7.312159467729915
('it', '*EXP*-2') 7.309505014283937
('recent', 'years') 7.304384942625129
('Japanese', 'companies') 7.304210714614422
('be', 'reached') 7.226089006096297
(';', '8') 7.201507761984825
('heavily', 'on') 7.197295510697909
('million', 'guilders') 7.190162788390815
('200', 'million') 7.163690577029627
('higher', 'prices') 7.163348178774575
('recent', 'months') 7.151754726787727
('do', "n't") 7.151075651679452
('be', 'able') 7.143626845904324
('disgorge', '$') 7.131520243047673
('$', '25,000') 7.131520243047671
('Japanese', 'investment') 7.111565636672028
('some', 'analysts') 7.1036604385866795
('year', 'ago') 7.102943927501235
('recent', 'days') 7.098938028344277
("n't", 'clear') 7.082419291012512
('able', '*-2') 7.080201465762691
('25,000', '*U*') 7.0802014657626895
('an', 'additional') 7.074571429189824
('next', 'year') 7.073278145293337
```

```
('Mr.', 'Courter') 7.068613491487479
('$', '15,000') 7.061130915156273
('few', 'years') 7.051404201455256
('as', 'well') 7.030645641288459
('5', '%') 7.029964482144129
('has', 'taken') 7.021573735791984
('common', 'stock') 7.020935516343004
('15,000', '*U*') 7.009812137871291
('it', '*EXP*-1') 7.002076489091687
('she', 'did') 6.983339926510103
('will', 'continue') 6.970360783820039
('earlier', 'this') 6.97026743872985
('argue', 'that') 6.962935413592943
('10', 'billion') 6.955980074502232
('Japanese', 'investors') 6.9416406352297155
('Mr.', 'Coleman') 6.931109967737543
('one', 'month') 6.922889460932577
('his', 'own') 6.893142117538526
('Moody', "'s") 6.864472774707254
('Poor', "'s") 6.864472774707254
('Rally', "'s") 6.864472774707254
('Weisfield', "'s") 6.864472774707252
("n't", 'yet') 6.860026869676064
('they', "'re") 6.8544886861346335
('had', 'been') 6.83157429418663
('unchanged', 'at') 6.830805061941858
('should', 'be') 6.802589928069258
('``', 'Cosby') 6.801462977181455
('futures', 'prices') 6.800778099389866
('three', 'years') 6.773870225926348
('billion', '*U*') 6.756003197250504
('Dealers', 'said') 6.73977702725794
('$', '5,000') 6.716482743768827
('are', 'priced') 6.713371647556595
('has', 'been') 6.7133205778016976
('million', '*U*') 6.713320004638501
('25', '%') 6.702983159530483
('suggested', 'that') 6.699901007759147
('``', 'We') 6.696907046138863
('earned', '$') 6.6720886244103745
('trading', 'yesterday') 6.670701031310573
('yen', '-LRB-') 6.669047400971078
('fiscal', 'year') 6.669047400971074
('7', '%') 6.666457283505368
('5,000', '*U*') 6.665163966483847
('$', '10,000') 6.662034959746453
('9', '%') 6.648535375508105
('from', 'continuing') 6.648391616348912
('sources', 'said') 6.646667622866456
('$', '250') 6.646093415877429
('months', ';') 6.640792807510344
('such', 'as') 6.632096264798184
('he', 'added') 6.621867132481297
('$', '130') 6.616947070217915
('50', '%') 6.616826515780767
('Cosby', "''") 6.616545261263671
('1,000', '*U*') 6.594774638592449
('ability', '*') 6.579444262631419
('an', 'estimated') 6.578613934527414
('an', 'analyst') 6.56069202653015
```

```
('be', 'built') 6.558664345183168
('%', 'term') 6.555425971116623
('information', 'about') 6.554348126433773
('two', 'years') 6.5468013173804245
('$', '300') 6.546557742326517
('11', '%') 6.538352457757684
('I', 'do') 6.534404291230958
('*U*', 'fine') 6.523808117238305
('have', 'been') 6.514582850318213
('well', 'as') 6.507083685231446
('10,000', '*U*') 6.495238965041537
('look', 'at') 6.48288175852155
('could', 'get') 6.479808924471929
('it', 'expects') 6.4766150001191924
('if', 'you') 6.4721553519284925
('12', '%') 6.4648234223357175
('can', 'do') 6.462735411112989
('for', 'instance') 6.459741181554797
('this', 'year') 6.455340707500849
('this', 'month') 6.443721624234017
('*', 'Take') 6.441940738881485
('5\\/8', '%') 6.439948753696687
(';', 'John') 6.4359730156218475
('trying', '*-2') 6.428124769182999
('15', '%') 6.426142954171658
('ban', 'on') 6.419687932034357
('may', 'be') 6.406661251738118
('you', 'do') 6.402614418675416
('$', '55') 6.394554648881467
('stock', 'market') 6.394393911870447
('1\\/2', '%') 6.385500969674311
('two', 'months') 6.372144795213025
('an', 'increase') 6.361383218306743
('for', 'example') 6.360205508003883
('20', '%') 6.344529188618004
('going', '*-2') 6.343235871596486
('she', 'says') 6.338767936650793
("n't", 'think') 6.33646491361905
('nation', "'s") 6.308079426182868
('be', 'acquired') 6.295629939349373
('valued', '*') 6.289937645436437
('talks', 'with') 6.286204926560108
('rumors', 'that') 6.284863508480303
('signs', 'that') 6.284863508480303
('will', 'become') 6.283300095480145
('people', 'who') 6.270632122639647
('refused', '*-1') 6.263825643133675
('you', 'can') 6.257416503135481
('$', '200') 6.257051125131532
('effort', '*') 6.245543526077983
('$', '70') 6.238435446964184
('In', 'October') 6.236496121293772
('dealers', 'said') 6.225203854428182
('*EXP*-1', 'is') 6.184797201312085
('shares', 'rose') 6.177661020030456
('``', 'You') 6.164033056566163
('indicated', 'that') 6.155580491535337
('ways', '0') 6.154814526536784
('be', 'paid') 6.143626845904324
('it', 'completed') 6.139580012841622
```

```
('does', 'not') 6.131118434093425
('Courter', "'s") 6.127507180541048
('stock', 'markets') 6.125905075944546
('said', '0') 6.102347106642647
('higher', 'than') 6.097759837146997
('speculation', 'that') 6.0884662956768025
('3', '%') 6.081494782784214
('will', 'be') 6.075289716141171
('those', 'who') 6.074234909836143
('closed', 'at') 6.0614179900832745
('and', 'minivans') 6.058072331705892
('and', 'vans') 6.058072331705892
('fact', 'that') 6.056044817984423
('6', '%') 6.05292563058744
('say', '0') 6.043453417589081
('3', 'million') 6.038159694945765
('He', 'also') 6.037218295211947
('believe', '0') 6.031957778751249
('close', 'at') 6.020776005585931
('comment', 'on') 6.019757325145724
('may', 'not') 6.019757325145724
('Ratners', "'s") 6.016475868152304
('3\\/4', '%') 6.011105454892814
("n't", 'want') 6.002045874548491
('quoted', '*-1') 6.000791237299879
('demand', 'for') 5.994077609205986
('$', '100') 5.979517149602621
('``', 'There') 5.971387978623767
('workers', 'at') 5.968308585691792
('aimed', '*') 5.968009550549075
('attempt', '*') 5.968009550549075
('be', 'used') 5.950981767961929
('spokesman', 'said') 5.946227904725367
('$', '4') 5.9388751651052765
('did', 'not') 5.931809625870018
('There', 'is') 5.929631038811634
('would', 'be') 5.928120810153114
('In', 'September') 5.926919129522947
('is', 'expected') 5.921069025308206
('think', '0') 5.900713245472998
('would', 'like') 5.886466052682703
('100', 'million') 5.886156601500716
('could', 'be') 5.880592440070533
('might', 'be') 5.8541202287093395
('designed', '*') 5.842478668465214
('thought', '0') 5.839312700808852
('efforts', '*') 5.83050602679914
('for', 'alleged') 5.829690791305104
('%', 'stake') 5.818460376950419
('40', '%') 5.818460376950419
('$', '3') 5.809592148160309
('appears', '*-1') 5.808146159357483
('disclosed', '*-1') 5.808146159357483
('willing', '*-1') 5.808146159357483
('account', 'for') 5.807664484975103
('want', '*-2') 5.807182971356276
('its', 'own') 5.7995360028509015
('analyst', 'at') 5.798383584249484
('traded', '*') 5.798084549106761
('It', 'also') 5.7758317422596335
```

```
('40', 'million') 5.775125289111971
('U.S.', 'trade') 5.772564028425723
('a', 'pound') 5.74437892921695
('believes', '0') 5.73977702725794
("n't", 'see') 5.739011468714695
('if', 'they') 5.735189757762287
('10', '%') 5.73099753570008
('We', 'have') 5.72938107049916
('likely', '*-1') 5.727226163973919
('Georgia-Pacific', "'s") 5.726969250957318
('Marine', "'s") 5.726969250957318
('no', 'one') 5.716985162384699
('Mr.', 'Cray') 5.7060434121027725
('issued', '*') 5.704975144715281
('*', 'making') 5.704975144715279
('produced', '*') 5.704975144715279
('stock', 'prices') 5.690595181639443
('saying', '0') 5.66938769936654
('its', 'bid') 5.6593583448026425
('help', '*-2') 5.653936711060593
('*-1', 'saying') 5.63822115791517
('allowed', '*-1') 5.63822115791517
('an', 'average') 5.637507623580982
('are', 'being') 5.63245165217303
('he', 'did') 5.627028837597077
('trading', 'companies') 5.624158445373542
('share', 'prices') 5.620205853748043
('trucks', 'and') 5.610613354734671
('Mr.', 'Bush') 5.609181872850181
('or', 'so') 5.606665909345118
('he', 'does') 5.603945224484036
('not', 'only') 5.600603717394646
('says', '0') 5.597554954605474
('known', '*') 5.589497927295346
('is', 'likely') 5.5771146240908465
('based', '*') 5.55813375638601
("''", 'says') 5.553351434776477
('down', 'from') 5.541476412432399
('expected', '*-1') 5.540665848492498
('according', 'to') 5.539875493043908
('to', 'disgorge') 5.539875493043908
('According', 'to') 5.5398754930439065
('access', 'to') 5.5398754930439065
('consented', 'to') 5.5398754930439065
('to', 'cover') 5.5398754930439065
('to', 'discuss') 5.5398754930439065
('to', 'expand') 5.5398754930439065
('could', "n't") 5.538098774788699
('based', 'on') 5.535880949538882
('``', 'I') 5.527996371250737
('he', 'says') 5.5190563268975215
('in', 'September') 5.515547947441874
('interested', 'in') 5.515547947441874
('are', 'still') 5.506920770089172
('for', 'comment') 5.504595417339059
('futures', 'market') 5.50457682962087
('can', 'be') 5.4959285898352075
('but', 'not') 5.495598103486765
('This', 'is') 5.4922257265043335
('continue', '*-1') 5.486218064470121
```

```
('failed', '*-1') 5.486218064470121
('a', 'bottle') 5.481344523383154
('more', 'like') 5.479041737513732
('sector', 'is') 5.463656574307562
('The', 'following') 5.45545906305618
('15', 'million') 5.453197194224611
('it', 'plans') 5.444434594370044
('be', 'sold') 5.443187127763235
('yield', 'from') 5.441940738881486
('authority', '*') 5.441940738881485
('PS', 'of') 5.44007317222472
('millions', 'of') 5.44007317222472
('source', 'of') 5.44007317222472
('of', 'Medicine') 5.440073172224718
('size', 'of') 5.440073172224718
('found', 'that') 5.439373457535929
('she', 'was') 5.417899984053767
('in', '1991') 5.416012273890962
('has', 'already') 5.406863891676775
('*-1', 'To') 5.398755223219782
('agreed', '*-1') 5.398755223219782
('had', 'no') 5.395057067497335
('are', 'expected') 5.391443552669234
('called', '*') 5.383047049827917
('will', 'help') 5.3805972968350595
('Taiwan', 'and') 5.380000426593256
('The', 'department') 5.378643466005348
('500', 'million') 5.375194682223338
('continues', '*-1') 5.370740847050186
('priced', '*-1') 5.370740847050186
('wanted', '*-1') 5.370740847050186
("n't", 'even') 5.357526529146877
('officials', 'said') 5.355113177022616
('$', '40') 5.35391266438412
('should', "n't") 5.349064950398681
('declined', '*-1') 5.348714540720186
('trying', '*-1') 5.348714540720186
('to', 'extend') 5.347230415101512
('but', 'I') 5.3428204086588735
('a', 'share') 5.341414262238679
('?', "''") 5.339011285734758
('a', 'lot') 5.329341429938104
('priced', '*') 5.32646352146155
('noted', 'that') 5.319079223818218
('right', '*') 5.317952021606034
('contributed', 'to') 5.317483071707461
('to', 'file') 5.317483071707461
('fined', '*-1') 5.316293063027809
('when', 'he') 5.314438607289052
('there', 'is') 5.311653480862512
('is', 'seeking') 5.303191902114316
('``', 'If') 5.301536580316096
('``', 'This') 5.301536580316096
('million', 'shares') 5.292732522031363
('and', 'chief') 5.292537585342917
('director', 'at') 5.290236680579156
('on', 'Wall') 5.280623894296808
('Japan', "'s") 5.2795102739861
('response', 'to') 5.276841087210112
('to', 'pursue') 5.276841087210112
```

```
('$', '10') 5.266449823133781
('helped', '*-1') 5.263825643133675
('scheduled', '*-1') 5.263825643133675
('which', '*T*-1') 5.259467513903889
('a', 'university') 5.258952102046706
('Columbia', "'s") 5.239981909799461
('expects', '*-1') 5.238290551026536
('``', 'What') 5.238033638009938
('8', '%') 5.233497876229263
('or', 'sell') 5.232851072792787
('a', 'joint') 5.22980575638719
('estimated', 'that') 5.225969819426737
('analysts', 'said') 5.225203854428182
('result', 'in') 5.2233671959485655
('when', 'they') 5.223290719230857
('will', 'take') 5.221899550816001
('``', 'It') 5.220616584932532
('under', 'which') 5.220616584932529
('scheduled', '*') 5.219548317545037
('yen', 'from') 5.219548317545037
('to', 'acquire') 5.217947398156545
('100', '*U*') 5.217704989512624
('The', 'Soviet') 5.207531549612593
('agreement', '*') 5.202474804186098
('named', '*-1') 5.196711447275137
('for', 'each') 5.195146261733385
('%', 'increase') 5.186192161450904
('10', 'million') 5.1730892750318755
('They', 'are') 5.171317738304733
('says', '*T*-1') 5.157369326285153
('it', 'does') 5.139580012841622
('to', 'findings') 5.124837993765064
('stake', 'in') 5.116451992032054
('``', 'She') 5.093643728674763
('``', 'When') 5.093643728674763
('want', '*-1') 5.087668687979846
('they', 'do') 5.080686323788056
('advantage', 'of') 5.077503092840011
('sale', 'of') 5.07750309284001
('drop', 'in') 5.074975356055893
('able', '*-1') 5.071180565191279
('in', 'October') 5.070237437049229
('way', '0') 5.069925628950269
('and', 'Learning') 5.058072331705892
("'s", 'proposal') 5.057117852649652
("'s", 'largest') 5.05711785264965
('an', 'American') 5.056845260293452
('to', 'ease') 5.0544486658736645
('to', 'settle') 5.0544486658736645
('He', 'said') 5.041805564707598
('number', 'of') 5.035682917145383
('$', '2') 5.031984569496759
('buy', 'or') 5.0284925742866005
('China', "'s") 5.027971506990134
('led', '*') 5.026903239602641
('``', 'So') 5.026529532816227
('lack', 'of') 5.025035672945876
('lot', 'of') 5.025035672945874
('because', 'they') 5.022471709842758
('we', 'have') 5.020935516343004
```

```
('there', 'are') 5.004420429559987
('in', 'August') 5.000974774612116
('in', '1986') 5.000974774612114
('wants', '*-1') 5.000791237299879
('they', 'say') 4.998224163596081
('but', 'he') 4.9948971052385644
('who', '*T*-3') 4.992647375339878
('plans', '*-1') 4.98371772394094
('spokesman', 'for') 4.981693884750152
('0', 'they') 4.973064089697679
('expected', '*-2') 4.9647242483427565
('may', 'have') 4.962041827289436
('new', 'company') 4.957582179098738
('to', 'consider') 4.954912992322752
('kind', 'of') 4.954646345054476
('majority', 'of') 4.954646345054476
('of', 'watches') 4.954646345054476
('The', 'average') 4.940885890226424
("'s", 'decision') 4.938473356151031
('cents', 'a') 4.937024007159344
('they', 'can') 4.935488408248121
('says', '*T*-2') 4.934093826176365
('1', 'million') 4.933823035131029
('under', 'its') 4.922392750636437
('0', 'it') 4.917187591505176
('*U*', '-RRB-') 4.916702733479813
('built', '*-1') 4.901255563748965
('were', "n't") 4.900940062942114
('a', 'fixed') 4.896382022661998
('orders', 'for') 4.894541935655072
('to', 'win') 4.887798796464216
("''", 'While') 4.879579667097461
('was', 'made') 4.877331602691063
('used', '*') 4.874900146157591
('I', 'would') 4.8738660159030704
('a', 'single') 4.869909811300808
('they', 'were') 4.868025440469381
('role', 'in') 4.86347125086218
('to', 'operate') 4.861803587931268
('to', 'produce') 4.861803587931268
('is', 'based') 4.859957511004822
('part', 'of') 4.855110671503564
('acquisition', 'of') 4.855110671503562
('know', '0') 4.839312700808852
('an', 'investment') 4.830152701523376
('to', 'buy') 4.811955038480708
('decided', '*-1') 4.808146159357483
('used', '*-1') 4.808146159357483
('to', 'begin') 4.802909898877703
('country', "'s") 4.8003424372875365
('in', 'Washington') 4.799340913442464
('Street', "'s") 4.786470262705981
('decision', '*') 4.7789757261590555
('$', '50') 4.777883288432971
('issue', 'was') 4.777795929140149
('The', 'Treasury') 4.775978963550735
('0', '*T*-3') 4.775452758861935
('Treasury', 'said') 4.774542445418618
('-RRB-', '--') 4.773870225926348
('is', 'being') 4.769759702033243
```

```
('administration', "'s") 4.76493710115634
('subject', 'to') 4.762267914380356
('to', 'meet') 4.762267914380356
('instead', 'of') 4.76200126711208
('introduction', 'of') 4.76200126711208
('Board', "'s") 4.760136114892518
('profit', 'from') 4.753884745196228
('lead', 'to') 4.751379598237618
('designed', '*-1') 4.749252470303917
('would', 'have') 4.7465357915382285
('a', 'great') 4.74437892921695
('posted', 'a') 4.74437892921695
('a', 'principal') 4.744378929216948
('a', 'line-item') 4.744378929216946
('concern', 'said') 4.73977702725794
('in', 'January') 4.737940368778322
('its', 'common') 4.733358926246417
('all', 'this') 4.731225888517853
('terms', 'of') 4.729579789419702
('today', "'s") 4.726969250957318
('in', '1985') 4.7208668554193824
('investments', 'in') 4.7208668554193824
('.', 'About') 4.716985162384699
('.', 'Although') 4.716985162384699
('.', 'Last') 4.716985162384699
('.', 'More') 4.716985162384699
('.', 'Several') 4.716985162384699
('.', 'Since') 4.716985162384699
('.', 'Under') 4.716985162384699
('Ltd', '.') 4.716985162384699
('.', 'Meanwhile') 4.716985162384697
('.', 'Moreover') 4.716985162384697
('.', 'Still') 4.716985162384697
('.', 'Terms') 4.716985162384697
('Co', '.') 4.716985162384697
('.', 'Fees') 4.716985162384695
('.', 'From') 4.716985162384695
('.', 'His') 4.716985162384695
('.', 'However') 4.716985162384695
('.', 'Instead') 4.716985162384695
('.', 'Its') 4.716985162384695
('.', 'Sales') 4.716985162384695
('.', 'Until') 4.716985162384695
('Corp', '.') 4.716985162384695
('Inc', '.') 4.716985162384695
('N.J', '.') 4.716985162384695
('-RRB-', ';') 4.716080934814581
('1', '%') 4.714123717135681
('to', 'yield') 4.70980049448622
('as', 'much') 4.708717546401099
('they', 'had') 4.708717546401097
('But', 'some') 4.706770286018095
('companies', 'are') 4.699565848031568
('they', 'are') 4.699565848031565
('which', '*T*-2') 4.697054628875517
('Campbell', "'s") 4.694547773264944
('to', 'block') 4.691878586488956
('$', '500') 4.690947651661691
('director', 'of') 4.688000685668303
('loss', 'of') 4.688000685668303
```

```
('to', 'see') 4.6818944979163355
('noted', '0') 4.680883338204373
('by', 'Congress') 4.672892578201525
('to', 'make') 4.671979029051251
('There', "'s") 4.6718276967648595
('or', 'even') 4.66895018759946
('but', 'they') 4.66671001987952
('to', 'prevent') 4.665406375127768
('head', 'of') 4.662465593561167
('began', '*-1') 4.656143065912433
('would', "n't") 4.655168078124362
('amount', 'of') 4.651577277418429
('he', 'would') 4.651473594566623
('0', '*T*-2') 4.6415209059336
('in', 'effect') 4.638404695227409
('in', 'recent') 4.638404695227408
('the', 'Mississippi') 4.637436289553383
('the', 'Philippines') 4.637436289553383
('violate', 'the') 4.637436289553383
('the', 'Cray-3') 4.637436289553381
('the', 'same') 4.637436289553381
('throughout', 'the') 4.637436289553381
('.', 'For') 4.634523002192722
('of', 'America') 4.632718250167116
('*EXP*-1', "'s") 4.627433577406405
('*ICH*-3', 'in') 4.622463151358385
('in', 'April') 4.622463151358385
('about', '$') 4.614944717306766
('He', 'has') 4.609356755575519
('decline', 'in') 4.608657351833356
('a', 'one-time') 4.606875405467012
('a', 'recession') 4.606875405467012
('a', 'similar') 4.606875405467012
('.', 'Among') 4.601507944964762
('company', 'said') 4.5877739338128904
('fine', 'and') 4.584141143373481
('.', 'Despite') 4.5794816386347605
('of', 'course') 4.565604054308579
('revenue', 'of') 4.565604054308579
('.', 'But') 4.562951533712276
('*-3', 'to') 4.561901799373906
('offer', 'for') 4.560504158489714
('.', 'In') 4.554364961565618
('that', 'we') 4.547897914314097
('.', 'Analysts') 4.547060160942385
('.', 'Each') 4.547060160942385
('$', '15') 4.546557742326517
('because', 'it') 4.542339182342031
('to', 'find') 4.5398754930439065
('.', 'On') 4.5364129167428775
('end', 'of') 4.533182576616198
('``', 'very') 4.531764841066648
('bonds', 'are') 4.531168316335847
('million', 'common') 4.527197775668386
('.', 'According') 4.5243400844423025
('.', 'Also') 4.5243400844423025
('.', 'Most') 4.5243400844423025
('form', 'of') 4.522535332416691
('added', '0') 4.517384605921491
('.', 'As') 4.510534284917274
```

```
('.', 'The') 4.5072802589570315
('increase', 'in') 4.5062100828625375
('0', 'she') 4.498525578670176
('.', 'At') 4.494592741048251
('.', 'Says') 4.494592741048251
('.', 'Then') 4.494592741048251
('.', 'Your') 4.494592741048251
('-LRB-', '$') 4.4910626297348095
("'", 'he') 4.485300727985415
('a', 'slowing') 4.481344523383154
('a', 'young') 4.481344523383154
('on', 'Tuesday') 4.481088476698501
('.', 'They') 4.480917804151172
('the', 'past') 4.479895012566903
('0', 'he') 4.47924947703472
('to', 'sell') 4.475745155624189
('who', '*T*-1') 4.4739770892392485
(':', '``') 4.4737175556987765
("'", 'He') 4.471114821742104
('It', "'s") 4.468544098376114
('a', 'package') 4.4642710100242144
('company', "'s") 4.456814805794007
('.', 'Another') 4.453950756550903
('.', 'Because') 4.453950756550903
('.', 'Indeed') 4.453950756550903
('.', 'Other') 4.453950756550903
('.', 'Profit') 4.453950756550903
('.', 'Ralston') 4.453950756550903
('.', 'Some') 4.453950756550903
('.', 'Those') 4.453950756550903
('.', 'Without') 4.453950756550903
('.', 'Yesterday') 4.453950756550903
('much', 'as') 4.4456831405673025
('the', 'highest') 4.444791211610989
('the', 'region') 4.444791211610989
('to', 'get') 4.444718260003565
("n't", 'be') 4.443187127763235
('``', 'They') 4.436112602002963
('to', 'help') 4.43553883322917
('department', "'s") 4.4315133674311475
('.', 'After') 4.427478545189711
('.', 'And') 4.427478545189711
('.', 'By') 4.427478545189711
('.', 'With') 4.427478545189711
('to', 'reduce') 4.4243982756239735
('a', 'special') 4.422450834329586
('.', 'One') 4.4121305808562745
('*-2', 'to') 4.410107618073578
('it', 'would') 4.409500804260656
('to', 'boost') 4.402371969293972
('a', 'few') 4.400424527999588
('In', 'his') 4.400373315027114
("'", 'she') 4.394152839927219
('0', 'we') 4.391853723837631
('are', "n't") 4.391443552669237
('he', 'said') 4.38614007264324
('the', 'nation') 4.385897522557418
('.', 'He') 4.385141598632252
('$', '30') 4.384286313427639
('other', 'than') 4.383346084970636
```

```
('they', 'have') 4.377079326568282
('the', 'ABA') 4.374401883719587
('Association', 'of') 4.36968384433332
('had', "n't") 4.368173773346387
('a', 'third-quarter') 4.365867305963217
('launched', 'a') 4.365867305963217
(',', 'Calif.') 4.365217430014976
(',', 'Conn.') 4.365217430014976
(',', 'Ill.') 4.365217430014976
(',', 'N.J.') 4.365217430014976
(',', 'N.Y.') 4.365217430014976
(',', 'Pa.') 4.365217430014976
(',', 'according') 4.365217430014976
(',', 'however') 4.365217430014976
('Conn.', ',') 4.365217430014976
('Ill.', ',') 4.365217430014976
('Indeed', ',') 4.365217430014976
('N.J.', ',') 4.365217430014976
('N.Y.', ',') 4.365217430014976
(',', 'Colo.') 4.365217430014974
(',', 'Fla.') 4.365217430014974
(',', 'Mass.') 4.365217430014974
(',', 'Mich.') 4.365217430014974
(',', 'N.J') 4.365217430014974
(',', 'meanwhile') 4.365217430014974
('Colo.', ',') 4.365217430014974
('Heights', ',') 4.365217430014974
('Meanwhile', ',') 4.365217430014974
('Moreover', ',') 4.365217430014974
('meanwhile', ',') 4.365217430014974
('do', '*T*-2') 4.3644792897886475
('Corp.', 'said') 4.361265404004211
('bid', 'for') 4.360205508003885
('.', 'Dealers') 4.35441508299999
('.', 'Now') 4.35441508299999
('firm', "'s") 4.349899601877496
('.', 'That') 4.34925337788421
('change', 'in') 4.348898078032423
('the', 'world') 4.347929672358397
('.', 'These') 4.345016384997738
('it', 'will') 4.345003695505323
('is', 'not') 4.343833886611662
('make', 'it') 4.342072876740367
('there', 'was') 4.334189277664532
('0', '*T*-1') 4.330354127476522
('.', 'Many') 4.329962039275452
('a', 'result') 4.329341429938104
('.', 'When') 4.324667739605935
('they', 'could') 4.320152258483443
('*-1', 'to') 4.316626364413349
('the', 'Soviets') 4.31550819466602
('the', 'size') 4.31550819466602
('is', "n't") 4.311160167456823
('Inc.', 'said') 4.309789186513125
('*U*', ';') 4.306205140651521
('a', 'gain') 4.303806337830968
('of', 'asbestos') 4.302569648474783
('type', 'of') 4.302569648474783
('.', 'Even') 4.301947663105855
('and', 'marketing') 4.292537585342915
```

```
('The', 'Chicago') 4.292228714187878
('$', '1') 4.290217989066731
('earnings', 'for') 4.282202996002612
('The', 'company') 4.281088156582676
('But', 'he') 4.27744439980685
('to', 'improve') 4.276841087210112
('the', 'Reagan') 4.274866210168675
('Calif.', ',') 4.272108025623496
('for', 'several') 4.267096103612401
('is', 'part') 4.265717196695652
('It', 'was') 4.263222756310391
('a', 'bad') 4.258952102046706
('.', 'A') 4.257553543747399
('.', 'So') 4.257553543747399
('.', 'Such') 4.257553543747399
('I', 'have') 4.245834793290028
('*', 'by') 4.245543526077984
('.', 'While') 4.243053974052284
('became', 'a') 4.241878588687767
('as', 'many') 4.240568710662691
("'s", 'construction') 4.232204559207741
('.', 'Net') 4.231558335214455
('.', 'Yet') 4.231558335214455
('*T*-1', '?') 4.227758654176551
('However', ',') 4.2277139062650395
('in', 'Tokyo') 4.2233671959485655
('the', 'main') 4.222398790274541
('the', 'Constitution') 4.2223987902745375
('In', 'New') 4.2159487438896885
('a', 'large') 4.208326028976737
('plans', '*') 4.202474804186098
('sold', '*-1') 4.200815845607872
('chairman', 'of') 4.196147589338629
('rise', 'in') 4.193619852554514
('*U*', 'each') 4.192676195021102
('across', 'the') 4.18997731258216
('Department', 'said') 4.18723600422916
('.', 'Program') 4.1864704456859165
('state', "'s") 4.186400869594616
('.', 'It') 4.181653429388142
```

Counting Vocabulary

Lexical Diversity, Tokens, and Types

To understand the richness of vocabulary in a text, we can calculate lexical diversity, tokens, and types.

Lexical diversity is the ratio of the number of unique words to the total number of words in the text.

In [15]:

```python
# Calculating lexical diversity of a text
def lexical_diversity(text):
    return len(set(text)) / len(text)
```

```
print(lexical_diversity(text1))
print(lexical_diversity(text7))
```

```
0.07406285585022564
0.12324685128531129
```

Basic Statistics of Text

Word Frequency and Cumulative Word Frequency Plot

To analyze word frequency in a text, NLTK provides the `FreqDist` class.

In [16]:

```
# Finding word frequency in a text
fdist = FreqDist(text7)
fdist
```

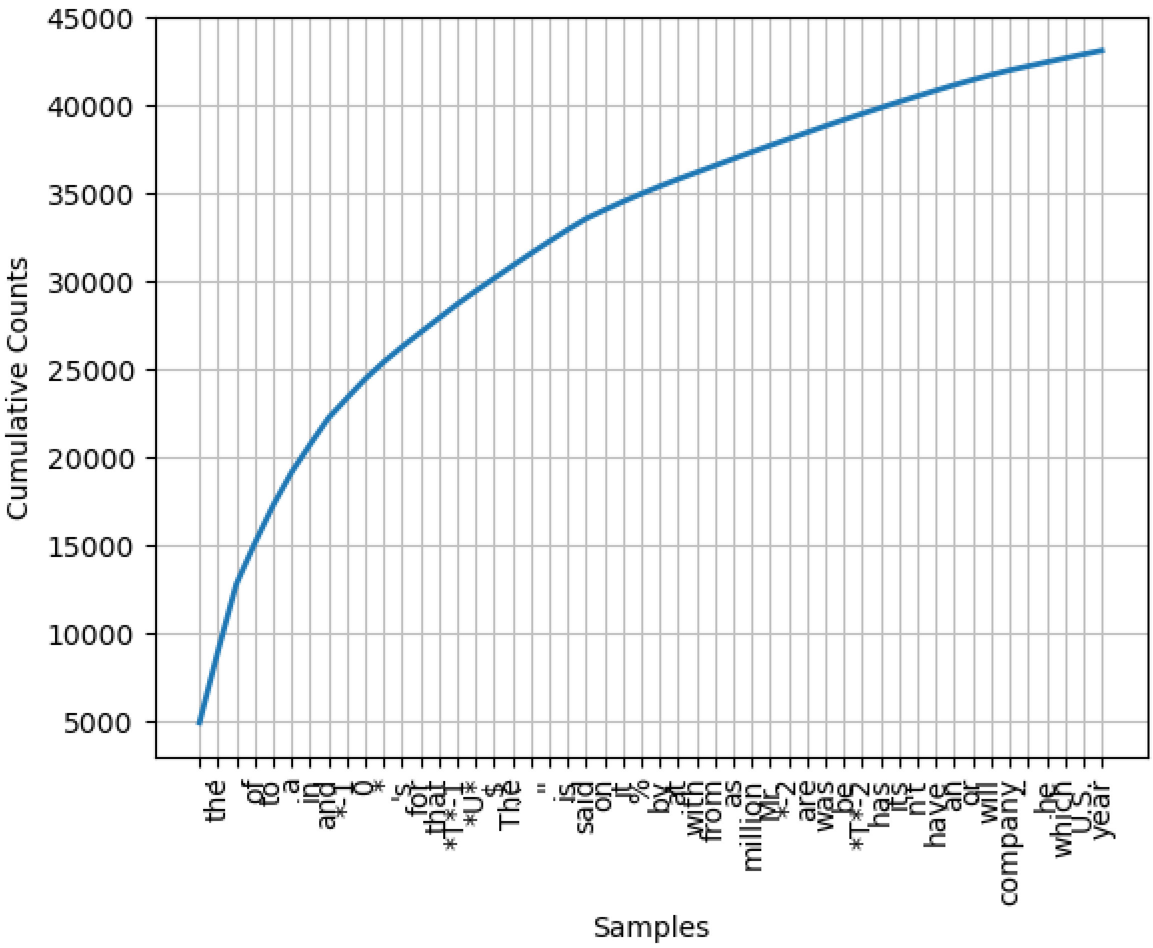
Out[16]:

```
FreqDist({',': 4885, 'the': 4045, '.': 3828, 'of': 2319, 'to': 2164, 'a': 1878, 'in':
1572, 'and': 1511, '*-1': 1123, '0': 1099, ...})
```

The `FreqDist` object contains the frequency distribution of words in the text.

In [17]:

```
# Plotting the cumulative word frequency
fdist.plot(50, cumulative=True)
```



Out[17]:

```
<Axes: xlabel='Samples', ylabel='Cumulative Counts'>
```

The `plot` function generates a cumulative word frequency plot for the top 50 words in the text.

The brown corpus

The Brown Corpus is a landmark linguistic resource and one of the earliest and most widely used corpora in natural language processing (NLP). Compiled in the early 1960s by researchers at Brown University, hence its name, the corpus represents a diverse and comprehensive collection of text samples from various genres and sources.

Containing over a million words, the Brown Corpus serves as a representative snapshot of American English, encompassing both written and spoken language. It is categorized into numerous genres such as news, fiction, science, and more, allowing researchers to explore language patterns across different domains.

The Brown Corpus has played a crucial role in advancing computational linguistics and NLP research, providing valuable insights into language structure, semantics, and usage. Linguists and language researchers commonly use it for studying language variation, developing language models, and evaluating algorithms in text processing and analysis.

Due to its historical significance and enduring relevance, the Brown Corpus remains an essential resource for anyone interested in the exploration and analysis of natural language.

In [18]:

```python
import nltk

# Download the Brown Corpus
nltk.download('brown')

# Load the Brown Corpus
from nltk.corpus import brown
```

```
[nltk_data] Downloading package brown to
[nltk_data]     C:\Users\Admin\AppData\Roaming\nltk_data...
[nltk_data]   Package brown is already up-to-date!
```

In [52]:

```python
# Accessing categories in the Brown Corpus
categories = brown.categories()
print("Categories in the Brown Corpus:", categories)

# Accessing words in a specific category (e.g., 'news')
news_words = brown.words(categories='news')
print("\nSample words from the 'news' category:", news_words[:10])
```

```
Categories in the Brown Corpus: ['adventure', 'belles_lettres', 'editorial', 'fictio
n', 'government', 'hobbies', 'humor', 'learned', 'lore', 'mystery', 'news', 'religio
n', 'reviews', 'romance', 'science_fiction']

Sample words from the 'news' category: ['The', 'Fulton', 'County', 'Grand', 'Jury',
'said', 'Friday', 'an', 'investigation', 'of']
```

```
[nltk_data] Downloading package brown to /root/nltk_data...
[nltk_data]   Package brown is already up-to-date!
```

Exercises

1. How many words are there in each text book of nltk. ? How many distinct words are there? Tabulate the results.
2. Compare the lexical diversity scores for adventure, editorial and hobbies categories of brown corpus. Which genre is more lexically diverse?
3. Produce a dispersion plot of the four main protagonists in Sense and Sensibility (text2): Elinor, Marianne, Edward, and Willoughby. What can you observe about the different roles played by the males and females in this novel? Can you identify the couples?
4. Find the collocations in Monty Python and the Holy .
5. Find all the four-letter words in the Chat Corpus (text5). With the help of a frequency distribution (FreqDist), show these words in decreasing order of frequency.
6. Define a function called vocab_size(text) that has a single parameter for the text, and which returns the vocabulary size of the text.

In [56]:

```
# 1. How many words are there in each text book of nltk. ? How many distinct words are
print(text1, f"has {len(text1)} many words and {len(set(text1))} unique words")
print(text2, f"has {len(text2)} many words and {len(set(text2))} unique words")
print(text3, f"has {len(text3)} many words and {len(set(text3))} unique words")
print(text4, f"has {len(text4)} many words and {len(set(text4))} unique words")
print(text5, f"has {len(text5)} many words and {len(set(text5))} unique words")
print(text6, f"has {len(text6)} many words and {len(set(text6))} unique words")
print(text7, f"has {len(text7)} many words and {len(set(text7))} unique words")
```

```
<Text: Moby Dick by Herman Melville 1851> has 260819 many words and 19317 unique word
s
<Text: Sense and Sensibility by Jane Austen 1811> has 141576 many words and 6833 uniq
ue words
<Text: The Book of Genesis> has 44764 many words and 2789 unique words
<Text: Inaugural Address Corpus> has 152901 many words and 10025 unique words
<Text: Chat Corpus> has 45010 many words and 6066 unique words
<Text: Monty Python and the Holy Grail> has 16967 many words and 2166 unique words
<Text: Wall Street Journal> has 100676 many words and 12408 unique words
```

In [62]:

```
# 2. Compare the lexical diversity scores for adventure, editorial and hobbies categor
adventure = brown.words(categories='adventure')
editorial = brown.words(categories='editorial')
hobbies = brown.words(categories='hobbies')

def lexical_diversity(text):
    return len(set(text)) / len(text)

print(lexical_diversity(adventure))
print(lexical_diversity(editorial))
print(lexical_diversity(hobbies))

# therefore editorials are the most lexical diverse category
```

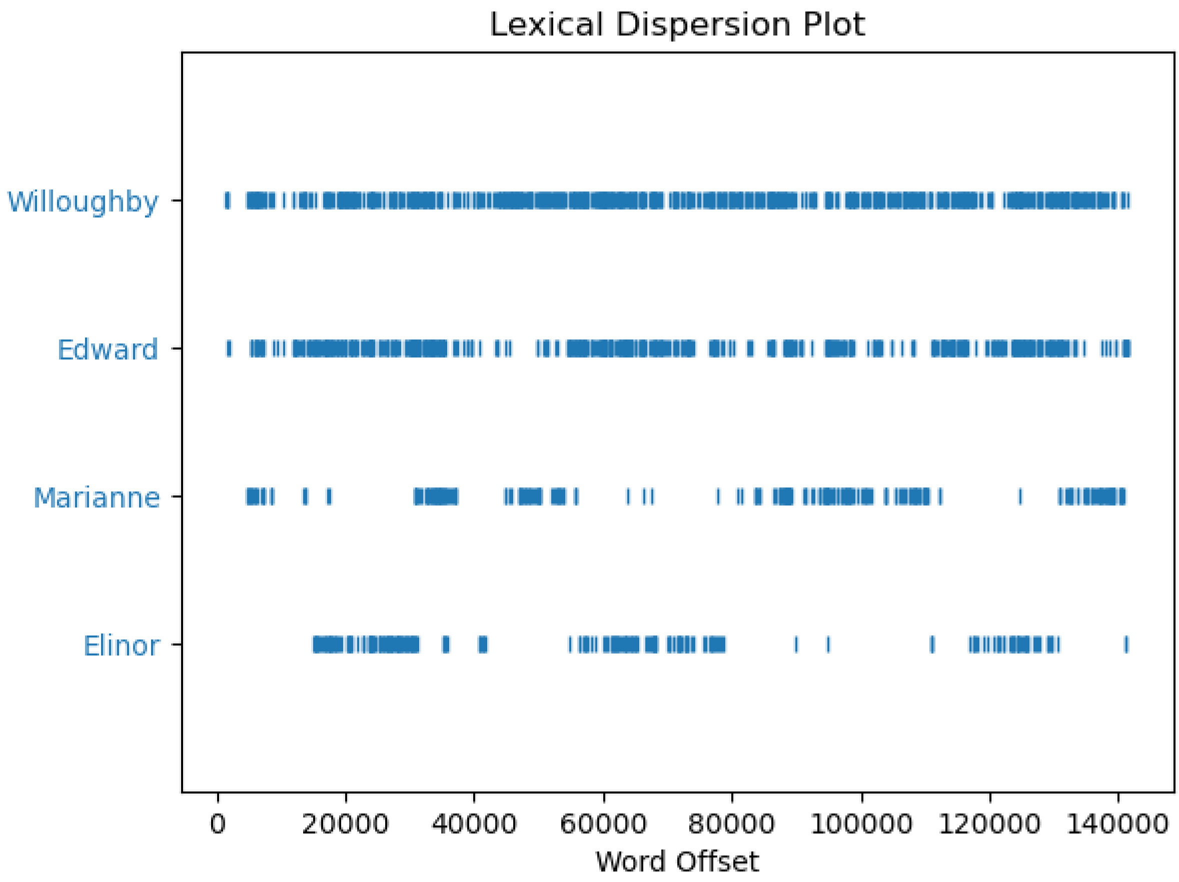
```
0.1279743878169075
0.16054152327770924
0.14493897625842492
```

In [64]:

```
# 3. Produce a dispersion plot of the four main protagonists in Sense and Sensibility

text2.dispersion_plot(['Elinor', 'Marianne', 'Edward', 'Willoughby'])

# we can observe that the males in this book are more involved that the females. As fc
```



In [66]:

```
# 4. Find the collocations in Monty Python and the Holy .

bigram_finder = collocations.BigramCollocationFinder.from_words(text6)

bigram_finder.apply_freq_filter(5)  # Adjust the frequency filter as needed
bigram_measures = collocations.BigramAssocMeasures()
scored_bigrams = bigram_finder.score_ngrams(bigram_measures.pmi)

for bigram, score in scored_bigrams[:1000]:
    print(bigram, score)
```

```
('ALL', 'HEADS') 11.728515783677448
('dramatic', 'chord') 11.243088956507204
('DEAD', 'PERSON') 11.05044387856481
('Round', 'Table') 11.05044387856481
('dona', 'eis') 11.050443878564808
('eis', 'requiem') 11.050443878564808
('angels', 'sing') 10.980054550673412
('Iesu', 'domine') 10.728515783677448
('Pie', 'Iesu') 10.728515783677448
('rewr', '!]') 10.728515783677448
('OLD', 'MAN') 10.710593875680186
('CARTOON', 'CHARACTER') 10.050443878564808
('OF', 'NI') 9.962981037314467
('saw', 'saw') 9.782408791672648
('clap', 'clap') 9.558590782235134
('Heh', 'heh') 9.465481377843652
('Who', 'Say') 9.406587688790083
('All', 'right') 9.395092049952254
('Three', 'questions') 9.243088956507204
('heh', 'heh') 9.20244697200986
('make', 'sure') 9.20244697200986
('FRENCH', 'GUARDS') 9.192462883437237
('hee', 'hee') 8.971492537169986
('Bring', 'out') 8.783657337869906
('Holy', 'Grail') 8.6259460500369
('Shut', 'up') 8.617484471288703
('squeak', 'squeak') 8.582409401753251
('KNIGHTS', 'OF') 8.560882593743123
('mumble', 'mumble') 8.558590782235134
('music', 'stops') 8.555032962821945
('Run', 'away') 8.497661783592804
('MIDDLE', 'HEAD') 8.295556376401342
('LEFT', 'HEAD') 8.29555637640134
('RIGHT', 'HEAD') 8.29555637640134
('Sir', 'Robin') 8.272001648422444
('clop', 'clop') 8.180079158981403
('Ha', 'ha') 8.155626115256865
('my', 'liege') 8.150439668541528
('King', 'Arthur') 8.12563137495903
('have', 'seen') 8.100131002665162
('ve', 'got') 8.073163955064894
('Burn', 'her') 8.05044387856481
('Knights', 'Who') 8.044017609405374
('BLACK', 'KNIGHT') 7.962981037314471
('GREEN', 'KNIGHT') 7.962981037314469
('Sir', 'Galahad') 7.890572541786421
('FRENCH', 'GUARD') 7.856859851652798
('have', 'been') 7.837096596831369
('Come', 'on') 7.79541530874608
('A', 'witch') 7.78209682388229
('ha', 'ha') 7.725240651644585
('quest', 'for') 7.627538135952626
('-', 'MASTER') 7.591012259927512
('CART', '-') 7.591012259927512
('Sir', 'Knight') 7.475535042507575
('Hold', 'it') 7.308976892163663
('Arthur', 'music') 7.192462883437237
('Sir', 'Launcelot') 7.16557958928629
('t', 'leave') 7.131580641290213
('brave', 'Sir') 7.06989624123098
```

```
('HEAD', 'KNIGHT') 7.066074530278572
('OFFICER', '#') 7.0617591917926426
('SOLDIER', '#') 7.0617591917926426
('VILLAGER', '#') 7.0617591917926426
('#', '1') 7.042650368844937
('#', '2') 7.0269937736319665
('get', 'him') 6.962981037314469
('your', 'dead') 6.947156070152786
('i', '--') 6.840990512935859
('j', '--') 6.840990512935859
('#', '3') 6.79872478595885
('Bridge', 'of') 6.746663130387706
('of', 'Death') 6.746663130387706
('going', 'to') 6.669014771928785
('t', 'want') 6.668180120449971
('does', 'it') 6.630904987051025
('said', 'it') 6.630904987051025
('Stop', 'that') 6.465481377843652
('she', 'is') 6.448054306085467
('your', 'quest') 6.443113564815199
('GUEST', '#') 6.383687286680006
('Thank', 'you') 6.378018536593315
('Did', 'you') 6.378018536593313
('GUARD', '#') 6.333838737229446
('your', 'name') 6.319124847539744
('it', 'again') 6.3089768921636615
('m', 'not') 6.2931296390068
('s', 'going') 6.258815829586322
('boom', 'boom') 6.236662687347772
('out', 'your') 6.236662687347772
('Are', 'you') 6.11498413075952
('to', 'tell') 6.1029112984589435
('I', 'feel') 6.056090441705949
('Who', 'are') 6.028076065536355
('I', 'am') 5.9889762458474145
('We', 'are') 5.902545183452496
('eh', '?') 5.871530093895242
('the', 'Round') 5.826442204366705
('bonk', ']') 5.765041659702561
('clang', ']') 5.765041659702561
('clunk', ']') 5.765041659702561
('howl', ']') 5.765041659702561
('roar', ']') 5.765041659702561
('chanting', ']') 5.765041659702559
('chord', ']') 5.765041659702559
('pause', ']') 5.765041659702559
('stops', ']') 5.765041659702559
('Knights', 'of') 5.746663130387706
('[', 'bonk') 5.733031264799941
('[', 'clang') 5.733031264799941
('[', 'clunk') 5.733031264799941
('[', 'howl') 5.733031264799941
('[', 'roar') 5.733031264799941
('[', 'angels') 5.733031264799939
('[', 'chanting') 5.733031264799939
('[', 'dramatic') 5.733031264799939
('[', 'pause') 5.733031264799939
('and', 'make') 5.710593875680184
('It', 'is') 5.65267202569394
('[', 'singing') 5.636815949540635
('come', 'and') 5.611058202129268
('on', 'with') 5.569855608330949
('the', 'Holy') 5.563407798532911
('What', 'is') 5.548083124416742
('the', 'Britons') 5.536935587171719
('the', 'room') 5.536935587171719
('If', 'you') 5.530021630038364
('singing', ']') 5.511285067456777
('?', 'OLD') 5.508960014510533
('seek', 'the') 5.5045141094793415
('a', 'bit') 5.495855026887172
(']', 'Bring') 5.417118356282254
('boom', ']') 5.402471580317851
('is', "'.") 5.396524005445388
('a', 'shrubbery') 5.380377809467234
('re', 'not') 5.3709637790593625
('thud', ']') 5.350004160423715
("'", 'll') 5.332767455498413
("'", 'm') 5.332767455498413
("'", 're') 5.332767455498413
("'", 've') 5.332767455498413
('didn', "'") 5.332767455498413
('doesn', "'") 5.332767455498413
('don', "'") 5.332767455498413
("'", 't') 5.332767455498411
('isn', "'") 5.332767455498411
('[', 'boom') 5.328641009720604
("'", 's') 5.322499120044585
('[', 'thud') 5.317993765521095
('sing', ']') 5.279614832532319
('I', 'don') 5.262541319173376
('--', 'if') 5.2560280122147045
(']', 'Pie') 5.250468486872801
('and', 'get') 5.236662687347771
('We', 'have') 5.222987750450693
('of', 'Camelot') 5.216148413688927
('King', 'of') 5.16170062966655
('uh', '--') 5.140550794794768
("'", 'ni') 5.110375034161963
('and', 'his') 5.099159163597836
('is', 'your') 5.09370473350573
("'", 'd') 5.0697330496646185
('Say', "'") 5.0697330496646185
('if', '--') 5.0633829342723065
('to', 'be') 5.039216623141554
('You', 'are') 5.004229323581985
('music', ']') 4.994523505825329
('name', 'of') 4.9811283840247285
('I', 'think') 4.978087929704678
('I', 'didn') 4.918586917956016
('I', 'mean') 4.918586917956016
('with', 'it') 4.86840430077768
('name', '?') 4.854456580536299
('let', "'") 4.847340628328171
('I', 'can') 4.756530159847042
('a', 'witch') 4.745833279895519
('got', 'a') 4.740967524723704
('[', 'rewr') 4.733031264799941
('?', 'VILLAGER') 4.709258664996364
('In', 'the') 4.6889386806167686
```

```
('find', 'the') 4.6889386806167686
('[', 'music') 4.68240519172997
('That', "'") 4.680690758918718
('They', "'") 4.6546955503857745
('do', 'you') 4.612483790230337
(',', 'please') 4.536716282612373
(',', 'sir') 4.536716282612373
(',', 'lad') 4.536716282612371
('Um', ',') 4.536716282612371
('from', 'the') 4.5045141094793415
('into', 'the') 4.50451410947934
('s', 'got') 4.47793311888991
('s', 'what') 4.47793311888991
('the', 'Knights') 4.463872124981995
('?', 'TIM') 4.450066325456966
('you', 'must') 4.439419081257455
('Well', ',') 4.425684970223628
('What', '?') 4.4220519492873684
('Christ', '!') 4.404785446156099
('Hic', '!') 4.404785446156099
('Ow', '!') 4.404785446156099
('Shh', '!') 4.404785446156097
(',', 'yes') 4.399212758862436
('think', 'I') 4.39312542898352
('domine', ',') 4.3847131891673214
('requiem', '.') 4.378018536593315
('.', 'ALL') 4.378018536593313
('if', 'you') 4.378018536593313
('you', 'get') 4.378018536593313
('s', 'not') 4.366572009942205
('He', "'") 4.332767455498413
('in', 'this') 4.3264985752220255
('the', 'Grail') 4.32394186383752
('Hello', '.') 4.3214350082269455
(',', 'eh') 4.314323861275923
(',', 'yeah') 4.314323861275923
('?', 'FRENCH') 4.306330847995515
('you', 'do') 4.2905556953429755
('out', 'of') 4.287231511750408
('right', '?') 4.286567593174086
('I', 'was') 4.278482863042397
('you', 'go') 4.240515012843378
('Uh', ',') 4.237156000753464
('...', 'BEDEVERE') 4.2370634916400824
('I', 'have') 4.235060582751268
(',', 'uh') 4.22859398725004
('he', 'is') 4.218186764186875
('It', "'") 4.203484438553447
('[', 'squeak') 4.19697836455973
('.', 'DEAD') 4.185373458650917
('Concorde', '!') 4.18239302481965
('clap', ']') 4.180079158981405
('mumble', ']') 4.180079158981405
('clop', ']') 4.180079158981403
('witch', '!') 4.149528390914023
('[', 'King') 4.1480687640787846
('[', 'clap') 4.148068764078783
('[', 'clop') 4.148068764078783
('[', 'mumble') 4.148068764078783
('.', 'Hello') 4.137010437089518
('are', 'you') 4.137010437089518
('!', 'GUESTS') 4.124677526963362
('swamp', '.') 4.11498413075952
('?', 'SOLDIER') 4.093922515231689
('the', 'name') 4.060907458003728
(',', 'but') 4.060278238669385
('Oh', ',') 4.059137317264813
(',', 'let') 4.051289455442131
('!', 'Hold') 4.042215366771389
('this', 'is') 4.024842875360923
('um', ',') 4.022143109782611
('Ni', '!') 4.020121595920774
(']', '[') 4.002217897615328
(']', 'NARRATOR') 3.9995069133395837
('...', 'FATHER') 3.9978758277606534
('!', 'This') 3.9897479468772534
('.', 'BRIDGEKEEPER') 3.962981037314469
('!', 'Shh') 3.9573264691848777
('for', 'the') 3.951973086450563
('Look', ',') 3.9517537818912167
('dead', '!') 3.9453538275188027
('at', 'you') 3.945059129317208
('!', 'Burn') 3.9283474022131113
('what', '?') 3.9239975137893772
('uh', ',') 3.9237394057216193
('!', 'Go') 3.919358618985857
('Aaaaugh', '!') 3.919358618985857
('spanking', '!') 3.919358618985857
('squeak', ']') 3.90706066457499
('Sorry', '.') 3.8925917094230726
('lad', '.') 3.8925917094230726
('Ohh', '!') 3.8902122733263393
('again', '!') 3.8902122733263393
(',', 'O') 3.8846395860326766
('at', 'the') 3.8789096242608387
('.', 'Just') 3.863445363763553
('but', 'I') 3.863445363763553
(',', 'dona') 3.8586443774997345
('well', ',') 3.8586443774997345
('ROBIN', ':') 3.8252364416213087
(':', 'Hic') 3.825236441621307
('BEDEVERE', ':') 3.825236441621307
('BRIDGEKEEPER', ':') 3.825236441621307
('CROWD', ':') 3.825236441621307
('CUSTOMER', ':') 3.825236441621307
('DENNIS', ':') 3.825236441621307
('FATHER', ':') 3.825236441621307
('GALAHAD', ':') 3.825236441621307
('GUARDS', ':') 3.825236441621307
('GUESTS', ':') 3.825236441621307
('HEADS', ':') 3.825236441621307
('HERBERT', ':') 3.825236441621307
('INSPECTOR', ':') 3.825236441621307
('LAUNCELOT', ':') 3.825236441621307
('MAN', ':') 3.825236441621307
('MASTER', ':') 3.825236441621307
('MINSTREL', ':') 3.825236441621307
('NARRATOR', ':') 3.825236441621307
('PERSON', ':') 3.825236441621307
('PIGLET', ':') 3.825236441621307
```

```
('RANDOM', ':') 3.825236441621307
('WOMAN', ':') 3.825236441621307
('3', ':') 3.825236441621305
(':', 'Aaaaugh') 3.825236441621305
('CHARACTER', ':') 3.825236441621305
('CONCORDE', ':') 3.825236441621305
('CRONE', ':') 3.825236441621305
('DINGO', ':') 3.825236441621305
('GIRLS', ':') 3.825236441621305
('GOD', ':') 3.825236441621305
('MONKS', ':') 3.825236441621305
('NI', ':') 3.825236441621305
('ZOOT', ':') 3.825236441621305
('?', 'LAUNCELOT') 3.8094691257629893
('of', 'the') 3.8080636750518497
('1', ':') 3.8061276186735995
(',', 'um') 3.7997506884461636
('in', 'the') 3.792494872443367
('ARTHUR', ':') 3.77300648284875
(':', 'Well') 3.7707886575989296
(']', 'BLACK') 3.765041659702561
('2', ':') 3.754847113729909
('.', 'OFFICER') 3.7259418400136184
('yes', '.') 3.7259418400136184
('TIM', ':') 3.725700768070391
('?', 'GALAHAD') 3.707864082924612
(']', 'He') 3.7009113222828454
('.', 'MIDDLE') 3.6999466314806764
('you', 'can') 3.6999466314806764
(':', 'Shut') 3.6997055595374473
('MAYNARD', ':') 3.6997055595374473
(':', 'Yes') 3.6808465322861323
('?', 'BLACK') 3.6788850159528472
('--', 'FATHER') 3.671065511493545
('!', 'DINGO') 3.6678198519898935
('Please', '!') 3.6678198519898917
(',', 'no') 3.662247164696229
('.', 'LEFT') 3.6410529424271054
(':', 'Then') 3.632591363678909
('?', 'ARTHUR') 3.630638571098787
('KNIGHT', ':') 3.6204166497328067
(',', 'brave') 3.619178442804346
(':', 'Ah') 3.6028440202848575
(':', 'Oooh') 3.6028440202848575
('Grail', '?') 3.5914221747025064
(':', 'What') 3.5842283421175125
('is', 'a') 3.5752894943815754
('with', 'a') 3.5698556083309487
(':', 'Agh') 3.5622020357875126
(':', 'Are') 3.5622020357875126
(':', 'Three') 3.5622020357875126
('!', 'Thank') 3.5567885396011487
('away', '!') 3.5491753554912737
('sir', '.') 3.547943538035627
('No', ',') 3.536716282612373
(',', 'when') 3.536716282612371
('So', ',') 3.536716282612371
(':', 'Uh') 3.525676159762398
('!', 'MINSTREL') 3.5117006500726102
('room', '.') 3.5035494186771707
```

```
(':', 'Get') 3.503308346733945
(':', 'How') 3.503308346733945
(']', 'FRENCH') 3.4920231652961444
('!', 'Ha') 3.487247606348072
('.', 'HERBERT') 3.484933740509824
('they', "'") 3.4847705489434624
('have', 'to') 3.4745265174466606
('to', 'have') 3.4745265174466606
('!', 'Run') 3.4718996420146357
('Right', '.') 3.4304859564874484
(':', 'Look') 3.410198942342463
(':', 'My') 3.4101989423424612
(':', 'Where') 3.4101989423424612
('!', 'A') 3.404785446156099
('!', 'Ohh') 3.404785446156099
('!', 'Stop') 3.404785446156099
('!', 'rewr') 3.404785446156099
('her', '!') 3.404785446156099
('rewr', '!') 3.404785446156099
('up', '!') 3.404785446156099
('!', 'CONCORDE') 3.4047854461560974
('tell', '!') 3.4047854461560974
(':', 'Oh') 3.401428732714731
('?', 'ROBIN') 3.3934827970905985
(':', 'Right') 3.3922770343452004
(':', 'No') 3.3846638502353237
('we', "'") 3.378571145111538
('.', 'CARTOON') 3.3780185365933146
('.', 'All') 3.378018536593313
('.', 'So') 3.378018536593313
('Camelot', '.') 3.378018536593313
('have', 'a') 3.3528970730451277
('.', 'ROBIN') 3.340543831174651
('can', "'") 3.332767455498413
('We', "'") 3.332767455498411
('Yes', ',') 3.3143238612759234
('You', "'") 3.308920713544044
(':', 'The') 3.2846680602586034
('now', ',') 3.2736818767785785
('liege', '!') 3.2672819224061644
('on', 'a') 3.2631942700968963
('.', 'CART') 3.2625413191733763
('is', 'that') 3.252134096110211
('Britons', '.') 3.240515012843378
('leave', '.') 3.240515012843378
(':', 'Now') 3.240273940900149
('Launcelot', '!') 3.220360875018672
('?', 'GUARD') 3.2139989672234393
('Grail', '.') 3.1974462909514916
('?', 'HEAD') 3.18703191962317
('!', 'DENNIS') 3.1823930248196515
('Arthur', ',') 3.1741462032276626
('me', '?') 3.1553230598958333
('KNIGHTS', ':') 3.13934503204937
('.', 'Come') 3.1370104370895184
('!', 'CROWD') 3.124677526963362
('.', 'GIRLS') 3.1149841307595203
('Right', '!') 3.1093295626299273
(':', 'But') 3.0882708474551013
(':', 'There') 3.0882708474551013
```

```
(':', 'Please') 3.0882708474550995
('.', 'It') 3.0560904417059493
('!', 'CUSTOMER') 3.02627382290237
('.', 'GALAHAD') 3.024381581978611
('?', 'BEDEVERE') 3.011182084223753
('Who', "'") 3.010839360611051
('I', "'") 3.0108393606110475
('.', 'CUSTOMER') 2.9995069133395837
('you', 'are') 2.9995069133395837
('.', 'We') 2.994689897041807
(':', 'If') 2.9772395350663565
('.', 'GUARD') 2.9720261769174776
('you', 'have') 2.9720261769174776
('!', 'TIM') 2.9573264691848777
(',', 'what') 2.9517537818912167
(',', 'who') 2.951753781891213
('!', 'VILLAGER') 2.9376594357288006
(':', 'And') 2.93626775401005
('.', 'That') 2.9185869179560164
('he', "'") 2.9065027007963167
('Yes', '.') 2.892591709423069
('.', 'LAUNCELOT') 2.884978525313196
(']', 'KNIGHTS') 2.8775163889609736
('him', '.') 2.87551819606413
('see', '.') 2.87551819606413
('.', 'KNIGHTS') 2.8690048891054545
('!', 'FRENCH') 2.86873254591589
('...', '[') 2.8503882154381
('.', 'BEDEVERE') 2.8395986218091878
('it', "'") 2.8392279825408515
(':', 'All') 2.825236441621305
(':', 'Not') 2.825236441621305
('!', 'GALAHAD') 2.819822945434943
('!', 'There') 2.819822945434943
('off', '!') 2.8198229454349413
(':', 'You') 2.801389699666938
('.', 'ARTHUR') 2.793056035872155
(',', 'King') 2.7818287804489046
('brave', ',') 2.7711815362493954
(',', 'if') 2.7591087039488187
('!', 'Ni') 2.7570871900869776
(',', 'we') 2.75244497366781
(':', 'Ni') 2.7300792085809675
('HEAD', ':') 2.714205129232562
('ha', '!') 2.71290774151843
('?', 'FATHER') 2.7016050924529296
(',', 'they') 2.6887193760574224
('not', 'a') 2.6885001048295685
('!', 'FATHER') 2.6754330360997685
('!', 'KNIGHTS') 2.6547636991644463
('.', 'FATHER') 2.648666126536982
('that', "'") 2.640889750860744
('!', 'NARRATOR') 2.6392506997931218
(':', 'Stop') 2.632591363678909
('!', 'SOLDIER') 2.627177867492545
('.', 'CROWD') 2.6124837902303373
(',', 'my') 2.6107168640561493
(':', 'Run') 2.6028440202848593
(':', 'So') 2.6028440202848575
('.', 'SOLDIER') 2.6004109579297605
```

```
('Burn', '!') 2.5747104475984113
(':', 'A') 2.5736976746253433
('.', 'DENNIS') 2.5706636145357074
('right', '.') 2.5706636145357074
('it', '!') 2.5702090553634704
('!', 'ARTHUR') 2.568284178438976
('that', '?') 2.549601999007878
('!', 'LAUNCELOT') 2.5491753554912737
('one', '.') 2.547943538035627
('I', '--') 2.5470767942180927
(',', 'so') 2.536716282612371
('.', 'You') 2.534744040280767
('...', 'ARTHUR') 2.523944639428244
('no', '.') 2.5035494186771707
(':', 'It') 2.5033083467339434
('s', 'it') 2.4913536346522314
('GUARD', ':') 2.489633409836868
('Robin', '!') 2.4787860275998757
(']', 'ARTHUR') 2.4748224245425376
(':', 'Come') 2.446724818367578
('!', '[') 2.4449248370093137
('on', '!') 2.435159095199616
('.', 'VILLAGER') 2.4083921856368313
('Look', '!') 2.4047854461560974
(',', 'I') 2.400343840881085
('!', 'BEDEVERE') 2.3809387042017303
('.', 'BLACK') 2.3780185365933146
(':', 'We') 2.3777774646500873
(',', 'and') 2.366791281170059
(':', 'That') 2.3658048229840105
("'", 'Ni') 2.3631411045419313
('-', 'a') 2.3583515031372375
('!', 'Who') 2.34589175710253
('No', '.') 2.3000160245920416
('!', 'ROBIN') 2.2893082287361644
('--', 'a') 2.286401661258223
('go', '.') 2.240515012843378
(':', 'He') 2.208565081172811
('it', '?') 2.2004524353854915
('.', 'He') 2.1763846754236624
('s', 'a') 2.1636585965459805
(']', 'FATHER') 2.109689831090007
('!', 'HEAD') 2.1093295626299273
(']', 'LAUNCELOT') 2.10207664698013
('is', 'the') 2.0985217498035063
('here', '.') 2.0979106174005775
('no', ',') 2.0772846639750746
(':', 'Ha') 2.0597016952583296
(',', 'you') 2.03421594208319
('here', ',') 2.034215942083188
('--', 'I') 2.016562077519314
(',', 'Sir') 2.0062015659135923
(':', 'Burn') 1.995161443063619
(':', 'Who') 1.9887351739041854
('.', 'Right') 1.9450591293172081
('away', '.') 1.9374459452073314
(',', 'he') 1.9178064499678769
('!', '...') 1.9144598195730183
('Grail', '!') 1.9022851056269143
(']', '...') 1.8823986103407204
```

```
('!', 'He') 1.881223490099087
('!', 'It') 1.860464929932288
(':', 'I') 1.8532508177909008
('on', '.') 1.823429684915677
('!', 'We') 1.819822945434943
('that', '!') 1.819822945434943
(',', 'it') 1.7952492962112245
('.', 'TIM') 1.7930560358721568
('!', 'The') 1.7267135410434609
(',', 'Arthur') 1.6887193760574224
('me', '.') 1.6618115025939026
('.', 'I') 1.631592613178043
('a', '--') 1.6083297561455847
(',', 'that') 1.6038304784709077
('.', 'No') 1.5895226417870258
(':', 'Sir') 1.5842283421175125
(',', 'not') 1.5773582671097177
(':', '[') 1.573913018314208
('!', 'And') 1.5682841784389776
('and', 'the') 1.5569815293734752
('of', 'a') 1.5140023735974317
('.', 'Oh') 1.5035494186771707
(':', 'Launcelot') 1.5033083467339452
('No', '!') 1.4787860275998757
(',', 'And') 1.4778225935588019
('you', "'") 1.4676970355845214
('to', 'the') 1.4638721249819948
('.', 'HEAD') 1.4304859564874484
('--', 'ARTHUR') 1.349137416606185
('me', ',') 1.3350824214427206
('!', 'No') 1.326782934154826
('.', 'And') 1.3191248475397437
('.', '[') 1.308533436272029
('Oh', '.') 1.2970985412097455
(',', 'this') 1.2390357339716864
("'", 'is') 1.1898095016563683
('--', '[') 1.1085403998921457
('!', 'You') 1.0590106093143685
('you', '?') 1.0064596739813503
(',', '...') 0.9760013281378939
('it', '.') 0.8064765516344785
('you', '.') 0.7930560358721586
('that', '.') 0.793056035872155
('Oh', '!') 0.6234257326314392
('!', 'I') 0.4978948505475813
('t', '.') 0.43316009078577444
('you', '!') 0.4327998223256948
(',', 'in') 0.43237962279763664
('you', ',') 0.32372255927817406
(':', '...') 0.11251839370178018
(',', 'a') -0.6959444741779031
(',', 'the') -0.87993046952813
('I', ',') -1.1357090593591224
(',', "'") -1.8590320455666607
(':', "'") -2.085085059387488
```

In [71]:

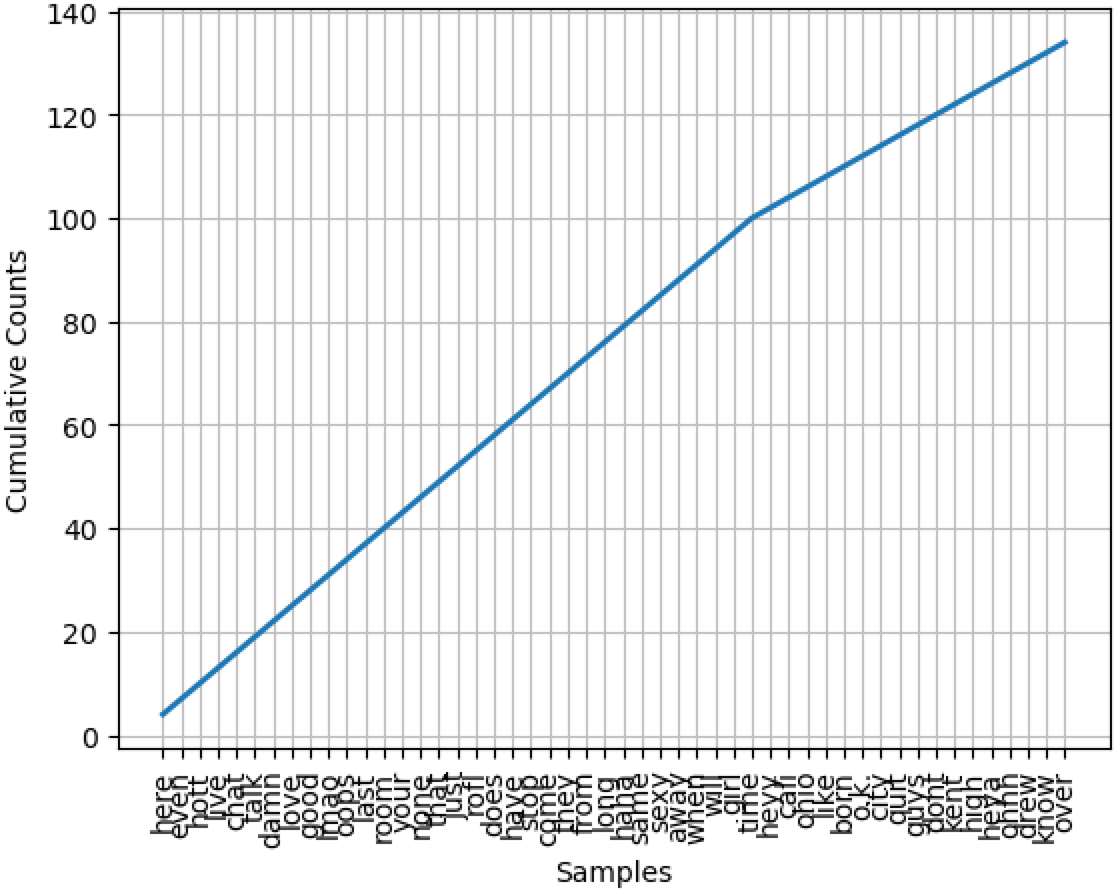
```
# 5. Find all the four-letter words in the Chat Corpus (text5). With the help of a fre

four_letter_words = [word.lower() for word in set(text5) if len(word) == 4]
```

```
fdist = FreqDist(four_letter_words)
fdist

fdist.plot(50, cumulative=True)

print("Four-Letter Words in Chat Corpus with Frequency:")
for word, frequency in fdist.most_common():
    print(f"{word}: {frequency}")
```



```
Four-Letter Words in Chat Corpus with Frequency:
here: 4
even: 3
hott: 3
live: 3
chat: 3
talk: 3
damn: 3
love: 3
good: 3
lmao: 3
oops: 3
last: 3
room: 3
your: 3
none: 3
that: 3
just: 3
rofl: 3
does: 3
have: 3
stop: 3
come: 3
they: 3
from: 3
long: 3
haha: 3
same: 3
sexy: 3
away: 3
when: 3
will: 3
girl: 3
time: 3
heyy: 2
cali: 2
ohio: 2
like: 2
born: 2
o.k.: 2
city: 2
quit: 2
guys: 2
dont: 2
kent: 2
high: 2
heya: 2
ohhh: 2
drew: 2
know: 2
over: 2
yoko: 2
name: 2
gosh: 2
hill: 2
what: 2
ummm: 2
bone: 2
huge: 2
care: 2
```

```
nice: 2
rock: 2
rule: 2
dude: 2
down: 2
hugs: 2
mine: 2
awww: 2
food: 2
more: 2
elle: 2
seen: 2
hawt: 2
then: 2
turn: 2
show: 2
caps: 2
halo: 2
drop: 2
john: 2
hold: 2
slip: 2
song: 2
male: 2
kids: 2
lies: 2
this: 2
yall: 2
only: 2
some: 2
tell: 2
came: 2
poor: 2
cool: 2
mono: 2
home: 2
whoa: 2
hail: 2
tiff: 2
kiss: 2
pour: 2
take: 2
holy: 2
days: 2
part: 2
rang: 2
back: 2
road: 2
rush: 2
hard: 2
hiya: 2
life: 2
phil: 2
hand: 2
size: 2
well: 2
else: 2
fine: 2
kewl: 2
yeah: 2
```

```
ahhh: 2
sure: 2
pm's: 2
type: 2
nope: 2
dang: 2
okay: 2
loud: 2
wind: 2
mary: 2
face: 2
been: 2
seee: 2
teck: 2
were: 2
kold: 2
tisk: 2
with: 2
kool: 2
king: 2
west: 2
kick: 2
eyes: 2
woot: 2
lets: 2
nick: 2
swim: 2
mode: 2
late: 2
help: 2
lord: 2
very: 2
cute: 2
look: 2
need: 2
pmsl: 2
evil: 2
deep: 2
baby: 1
tenn: 1
akon: 1
dogs: 1
shop: 1
thah: 1
perv: 1
judy: 1
cars: 1
dumb: 1
poop: 1
book: 1
u175: 1
roof: 1
hide: 1
1200: 1
hump: 1
body: 1
ciao: 1
wide: 1
howz: 1
pick: 1
```

```
            gray: 1
            gold: 1
            wash: 1
            hola: 1
            note: 1
            u542: 1
            golf: 1
            adds: 1
            give: 1
            free: 1
            draw: 1
            <~~~: 1
            line: 1
            sing: 1
            scum: 1
            main: 1
            jump: 1
            1cos: 1
            hair: 1
            offa: 1
            side: 1
            burp: 1
            wild: 1
            joke: 1
            seem: 1
            both: 1
            ??!!: 1
            bomb: 1
            whou: 1
            span: 1
            hazy: 1
            barn: 1
            whew: 1
            ebay: 1
            paid: 1
            beer: 1
            pain: 1
            pass: 1
            lapd: 1
            dint: 1
            luvs: 1
            u117: 1
            1996: 1
            kong: 1
            ,,,,: 1
            tape: 1
            whud: 1
            u219: 1
            haaa: 1
            than: 1
            u114: 1
            goin: 1
            wrek: 1
            u120: 1
            waaa: 1
            grrr: 1
            heat: 1
            typo: 1
            shup: 1
            1900: 1
```

```
muah: 1
mmmm: 1
frst: 1
matt: 1
wean: 1
must: 1
cock: 1
gret: 1
laid: 1
term: 1
u115: 1
nova: 1
numb: 1
till: 1
:o *: 1
shot: 1
boot: 1
cyas: 1
u106: 1
bust: 1
dark: 1
pics: 1
sets: 1
oooh: 1
pull: 1
fast: 1
lime: 1
page: 1
jeff: 1
dyed: 1
itch: 1
hgey: 1
nads: 1
u143: 1
gone: 1
lots: 1
brwn: 1
asks: 1
slap: 1
dojn: 1
cams: 1
u169: 1
39.3: 1
u138: 1
yell: 1
bare: 1
nawt: 1
cuss: 1
hour: 1
????: 1
wood: 1
<333: 1
bred: 1
sayn: 1
2006: 1
!...: 1
blah: 1
boyz: 1
wubs: 1
!!!!: 1
```

```
                u101: 1
                woof: 1
                fool: 1
                scar: 1
                lool: 1
                mris: 1
                lala: 1
                sell: 1
                near: 1
                2day: 1
                much: 1
                save: 1
                u111: 1
                fart: 1
                bite: 1
                flaw: 1
                tend: 1
                many: 1
                sori: 1
                hows: 1
                jane: 1
                nods: 1
                limp: 1
                lick: 1
                ntmn: 1
                98.5: 1
                wine: 1
                keys: 1
                blue: 1
                plus: 1
                u139: 1
                ussy: 1
                guyz: 1
                ohwa: 1
                nose: 1
                goof: 1
                gays: 1
                ghet: 1
                comp: 1
                grew: 1
                argh: 1
                vbox: 1
                gooo: 1
                u103: 1
                ogan: 1
                yeee: 1
                u197: 1
                wazz: 1
                base: 1
                gees: 1
                u134: 1
                pm'n: 1
                brat: 1
                uses: 1
                cops: 1
                pink: 1
                dotn: 1
                grlz: 1
                pope: 1
                dood: 1
```

```
1299: 1
100%: 1
yes.: 1
easy: 1
rubs: 1
u156: 1
febe: 1
tall: 1
pfft: 1
####: 1
bein: 1
cast: 1
sips: 1
meat: 1
mkay: 1
wooo: 1
sand: 1
past: 1
ex's: 1
98.6: 1
join: 1
wana: 1
peek: 1
cure: 1
true: 1
fire: 1
felt: 1
3:45: 1
anti: 1
abou: 1
nada: 1
tail: 1
trip: 1
docs: 1
gift: 1
knee: 1
meds: 1
18st: 1
weed: 1
land: 1
clay: 1
cant: 1
wire: 1
u146: 1
yout: 1
butt: 1
prep: 1
serg: 1
orgy: 1
u118: 1
mass: 1
hiii: 1
dead: 1
game: 1
u148: 1
beat: 1
yeas: 1
smax: 1
ok'd: 1
best: 1
```

```
clue: 1
hick: 1
alot: 1
u989: 1
cost: 1
givs: 1
aint: 1
2:55: 1
vamp: 1
ques: 1
mame: 1
paul: 1
seat: 1
u144: 1
call: 1
jack: 1
rent: 1
ally: 1
u112: 1
town: 1
sock: 1
junk: 1
<3's: 1
blew: 1
u190: 1
hope: 1
dirt: 1
tits: 1
rich: 1
tooo: 1
u819: 1
boom: 1
flow: 1
lion: 1
tjhe: 1
eeek: 1
them: 1
wats: 1
otay: 1
10th: 1
done: 1
addy: 1
able: 1
babe: 1
spot: 1
....: 1
werd: 1
fear: 1
exit: 1
real: 1
lose: 1
.. .: 1
maps: 1
ribs: 1
grin: 1
u181: 1
beam: 1
feel: 1
1985: 1
yada: 1
```

```
salt: 1
xmas: 1
menu: 1
humm: 1
amen: 1
vent: 1
wont: 1
bong: 1
pork: 1
lyin: 1
wore: 1
ahah: 1
porn: 1
vvil: 1
yw's: 1
poof: 1
hits: 1
york: 1
clap: 1
firs: 1
2pac: 1
fuck: 1
64.8: 1
sexs: 1
peel: 1
cash: 1
akdt: 1
samn: 1
lady: 1
wher: 1
door: 1
form: 1
u108: 1
ruff: 1
u149: 1
tyvm: 1
pasa: 1
jess: 1
week: 1
meep: 1
ssri: 1
ther: 1
made: 1
4.20: 1
u136: 1
quiz: 1
chop: 1
caan: 1
sooo: 1
heys: 1
u122: 1
)))): 1
doin: 1
hogs: 1
gals: 1
used: 1
brad: 1
dies: 1
okey: 1
mauh: 1
```

```
>:->: 1
u132: 1
gags: 1
gear: 1
whos: 1
play: 1
skin: 1
u102: 1
6:38: 1
dear: 1
howl: 1
rest: 1
walk: 1
!???: 1
sat.: 1
tere: 1
lol.: 1
make: 1
deop: 1
owww: 1
lame: 1
drug: 1
fair: 1
ewww: 1
area: 1
sign: 1
slam: 1
liam: 1
word: 1
lost: 1
orta: 1
u126: 1
case: 1
eeww: 1
eats: 1
nawp: 1
loss: 1
miss: 1
twit: 1
lake: 1
7:45: 1
bear: 1
<<<<: 1
nooo: 1
poot: 1
foot: 1
mofo: 1
buff: 1
pair: 1
nude: 1
ouch: 1
crap: 1
puff: 1
tart: 1
hiom: 1
sore: 1
poem: 1
sort: 1
cepn: 1
ears: 1
```

```
    team: 1
    crop: 1
    slow: 1
    sent: 1
    mind: 1
    u141: 1
    u158: 1
    whoo: 1
    wait: 1
    coem: 1
    6:41: 1
    surf: 1
    suck: 1
    ages: 1
    tick: 1
    also: 1
    safe: 1
    u520: 1
    u110: 1
    poll: 1
    moms: 1
    nana: 1
    bull: 1
    told: 1
    rape: 1
    legs: 1
    yesh: 1
    toop: 1
    chit: 1
    said: 1
    deal: 1
    piff: 1
    meet: 1
    ssid: 1
    hmph: 1
    duet: 1
    u105: 1
    dork: 1
    u988: 1
    lube: 1
    scuk: 1
    u109: 1
    bacl: 1
    wear: 1
    sick: 1
    hear: 1
    icky: 1
    1980: 1
    hurr: 1
    took: 1
    raed: 1
    6:53: 1
    most: 1
    idea: 1
    ugly: 1
    noth: 1
    outs: 1
    u172: 1
    guts: 1
    ding: 1
```

```
hank: 1
pwns: 1
lazy: 1
date: 1
kind: 1
zone: 1
bike: 1
find: 1
such: 1
http: 1
hint: 1
card: 1
hell: 1
worl: 1
knew: 1
aime: 1
u170: 1
pies: 1
glad: 1
dawn: 1
sext: 1
babi: 1
reub: 1
hmmm: 1
ever: 1
iowa: 1
ways: 1
geez: 1
u164: 1
prob: 1
ruth: 1
into: 1
yawn: 1
1930: 1
disc: 1
prof: 1
benz: 1
yess: 1
t he: 1
rose: 1
cmon: 1
sexi: 1
kill: 1
1.98: 1
?!?!: 1
u129: 1
fock: 1
jeep: 1
acid: 1
fade: 1
ahem: 1
mahn: 1
typr: 1
news: 1
evah: 1
fawk: 1
tips: 1
sang: 1
bell: 1
it's: 1
```

```
xbox: 1
wyte: 1
imma: 1
outa: 1
guns: 1
lust: 1
mami: 1
dust: 1
fits: 1
list: 1
boys: 1
wife: 1
u133: 1
u153: 1
chip: 1
four: 1
twin: 1
shit: 1
feet: 1
hurt: 1
z-ro: 1
u130: 1
feat: 1
u154: 1
shes: 1
warm: 1
uyes: 1
cold: 1
asss: 1
spin: 1
eric: 1
temp: 1
anal: 1
shut: 1
ring: 1
soda: 1
fake: 1
thot: 1
bird: 1
mite: 1
u113: 1
soul: 1
hong: 1
arms: 1
tide: 1
post: 1
((((: 1
hooo: 1
lisa: 1
akst: 1
; ..: 1
left: 1
mark: 1
neck: 1
whys: 1
corn: 1
puts: 1
busy: 1
enuf: 1
toss: 1
```

```
chik: 1
grrl: 1
u104: 1
sigh: 1
plow: 1
lawl: 1
park: 1
urls: 1
pool: 1
idnt: 1
hook: 1
u116: 1
u147: 1
u145: 1
"...: 1
aunt: 1
mess: 1
lead: 1
tock: 1
dman: 1
perk: 1
cell: 1
luck: 1
bugs: 1
rats: 1
bloe: 1
nuff: 1
!!!.: 1
move: 1
mena: 1
send: 1
tthe: 1
kept: 1
ctrl: 1
herd: 1
toes: 1
whip: 1
goes: 1
half: 1
kmph: 1
ball: 1
nerd: 1
rick: 1
6:51: 1
u142: 1
thnx: 1
mins: 1
soup: 1
sean: 1
rain: 1
u123: 1
9.53: 1
tlak: 1
each: 1
boss: 1
bowl: 1
star: 1
plan: 1
self: 1
u168: 1
```

```
gimp: 1
dump: 1
vega: 1
says: 1
u155: 1
heal: 1
u137: 1
1.99: 1
mang: 1
crib: 1
test: 1
mama: 1
u100: 1
soft: 1
went: 1
once: 1
sits: 1
work: 1
gets: 1
pigs: 1
doll: 1
tina: 1
mean: 1
deaf: 1
bied: 1
jude: 1
wrap: 1
este: 1
fort: 1
cums: 1
boed: 1
pimp: 1
daft: 1
u128: 1
calm: 1
dawg: 1
u107: 1
tune: 1
hate: 1
choc: 1
wuts: 1
bois: 1
coat: 1
any1: 1
cook: 1
open: 1
allo: 1
isnt: 1
wish: 1
giva: 1
jerk: 1
newp: 1
east: 1
army: 1
<---: 1
vote: 1
seth: 1
ones: 1
thru: 1
mike: 1
```

```
u163: 1
grea: 1
text: 1
yoll: 1
want: 1
wack: 1
bend: 1
byes: 1
inch: 1
stay: 1
joey: 1
spat: 1
u121: 1
ltns: 1
woah: 1
fish: 1
ride: 1
haze: 1
opps: 1
blow: 1
u150: 1
n9ne: 1
ladz: 1
keep: 1
tory: 1
band: 1
u165: 1
lung: 1
u820: 1
45.5: 1
.op.: 1
head: 1
moon: 1
root: 1
sum1: 1
soon: 1
site: 1
toke: 1
hots: 1
hang: 1
9:10: 1
ooer: 1
hall: 1
elev: 1
pure: 1
yard: 1
push: 1
next: 1
roll: 1
kina: 1
runs: 1
read: 1
troy: 1
ello: 1
five: 1
wins: 1
pray: 1
jush: 1
thje: 1
club: 1
```

```
gawd: 1
nite: 1
u196: 1
full: 1
rn's: 1
pine: 1
fall: 1
u119: 1
hehe: 1
hero: 1
jail: 1
spit: 1
wall: 1
4:03: 1
dick: 1
heck: 1
uhhh: 1
bout: 1
lois: 1
waht: 1
puke: 1
out.: 1
sink: 1
year: 1
caca: 1
heee: 1
eggs: 1
ltnc: 1
snow: 1
syck: 1
died: 1
```

In [73]:

```
# 6. Define a function called vocab_size(text) that has a single parameter for the tex

def vocab_size(text):
    return len(set(text))

vocab_size(text1)
```

Out[73]:

```
19317
```

In [ ]: