1. Tokenization: Apply word tokenization on all documents in the collection, and answer the following questions:
    a. How many tokens (words and punctuation symbols) are in each document? How many tokens are in the entire collection? A token is a linguistic unit such as a word, punctuation mark, or alpha-numeric strings.

Number of tokens in each document:
Jerry Decided To Buy a Gun.txt: 302 tokens
Rentals at the Oceanside Community.txt: 376 tokens
Gasoline Prices Hit Record High.txt: 292 tokens
Cloning Pets.txt: 262 tokens
Crazy Housing Prices.txt: 390 tokens
Man Injured at Fast Food Place.txt: 170 tokens
A Festival of Books.txt: 307 tokens
Food Fight Erupted in Prison.txt: 222 tokens
Better To Be Unlucky.txt: 356 tokens
Sara Went Shopping.txt: 165 tokens
Freeway Chase Ends at Newsstand.txt: 335 tokens
Trees Are a Threat.txt: 335 tokens
A Murder-Suicide.txt: 398 tokens
Happy and Unhappy Renters.txt: 313 tokens
Pulling Out Nine Tons of Trash.txt: 293 tokens

Total number of tokens in the entire collection: 4516

    b. How many tokens that you found in a) are unique?

Number of unique tokens in the entire collection: 1475

2. Stop words removal: Remove the stop words from all documents. You can use the list of words defined by Python NLTK library. Answer the following questions:
    a. How many tokens are in each document after removing all the stop words? How many tokens are in the entire collection?

Number of tokens in each document after removing stop words:
Jerry Decided To Buy a Gun.txt: 136 tokens
Rentals at the Oceanside Community.txt: 193 tokens
Gasoline Prices Hit Record High.txt: 150 tokens
Cloning Pets.txt: 122 tokens
Crazy Housing Prices.txt: 182 tokens
Man Injured at Fast Food Place.txt: 90 tokens
A Festival of Books.txt: 165 tokens
Food Fight Erupted in Prison.txt: 116 tokens
Better To Be Unlucky.txt: 172 tokens
Sara Went Shopping.txt: 89 tokens
Freeway Chase Ends at Newsstand.txt: 159 tokens
Trees Are a Threat.txt: 169 tokens
A Murder-Suicide.txt: 187 tokens

Happy and Unhappy Renters.txt: 156 tokens
Pulling Out Nine Tons of Trash.txt: 164 tokens
Total number of tokens in the entire collection after removing stop words: 2250

b. How many tokens that you found in a) are unique?

Number of unique tokens in the entire collection after removing stop words: 1326

3. Compute TF-IDF: Compute TF-IDF for each document and output the TF-IDF matrix.

TF-IDF Matrix:

```
[[0.         0.03680481 0.         ... 0.         0.         0.        ]
 [0.         0.03178339 0.         ... 0.         0.         0.        ]
 [0.         0.         0.07142847 ... 0.         0.         0.0620236 ]
 ...
 [0.         0.         0.         ... 0.         0.         0.        ]
 [0.         0.         0.         ... 0.         0.         0.        ]
 [0.         0.         0.         ... 0.         0.         0.        ]]
```

4. Compute the Cosine Similarity: Compute the cosine similarity for each pair of two documents in the document collection, and output the similarity matrix.

Cosine Similarity Matrix:

```
[[1.         0.05333519 0.04759814 0.03603771 0.07037399 0.09409098
  0.05575282 0.0254604  0.03920851 0.0625272  0.03913447 0.02481109
  0.08674685 0.0694597  0.06397718]
 [0.05333519 1.         0.0478978  0.01826014 0.06686765 0.04439379
  0.0967798  0.04375897 0.03236334 0.01882887 0.029094   0.06307101
  0.07719559 0.10713778 0.0853535 ]
 [0.04759814 0.0478978  1.         0.02332013 0.10219899 0.05718802
  0.05070035 0.01778412 0.02783499 0.05216823 0.03262522 0.04516425
  0.09121227 0.03326993 0.04015156]
 [0.03603771 0.01826014 0.02332013 1.         0.06346931 0.04541256
  0.03839961 0.01659594 0.04854771 0.02546449 0.00264413 0.04115327
  0.04855055 0.02237809 0.03073008]
 [0.07037399 0.06686765 0.10219899 0.06346931 1.         0.04603659
  0.0829648  0.03006285 0.07253301 0.05214578 0.01510865 0.10142665
  0.12080674 0.0964729  0.05709894]
 [0.09409098 0.04439379 0.05718802 0.04541256 0.04603659 1.
  0.05001542 0.01690136 0.01499247 0.02735224 0.0231691  0.01586125
  0.05581454 0.04978417 0.03555886]
 [0.05575282 0.0967798  0.05070035 0.03839961 0.0829648  0.05001542
  1.         0.05462838 0.03753653 0.01344814 0.07505113 0.05481003
  0.09447087 0.06298408 0.06724288]
 [0.0254604  0.04375897 0.01778412 0.01659594 0.03006285 0.01690136
  0.05462838 1.         0.0137655  0.01240238 0.02872424 0.03048414
  0.07079327 0.0225104  0.04124595]
 [0.03920851 0.03236334 0.02783499 0.04854771 0.07253301 0.01499247
  0.03753653 0.0137655  1.         0.02906372 0.04287437 0.06722818
```

```
 0.03410111 0.04683304 0.04075483]
[0.0625272  0.01882887 0.05216823 0.02546449 0.05214578 0.02735224
 0.01344814 0.01240238 0.02906372 1.         0.01748669 0.0369262
 0.0478454  0.02849687 0.03295305]
[0.03913447 0.029094   0.03262522 0.00264413 0.01510865 0.0231691
 0.07505113 0.02872424 0.04287437 0.01748669 1.         0.02586369
 0.04532269 0.03832402 0.03936543]
[0.02481109 0.06307101 0.04516425 0.04115327 0.10142665 0.01586125
 0.05481003 0.03048414 0.06722818 0.0369262  0.02586369 1.
 0.04404302 0.02941161 0.05728974]
[0.08674685 0.07719559 0.09121227 0.04855055 0.12080674 0.05581454
 0.09447087 0.07079327 0.03410111 0.0478454  0.04532269 0.04404302
 1.         0.09656904 0.08144351]
[0.0694597  0.10713778 0.03326993 0.02237809 0.0964729  0.04978417
 0.06298408 0.0225104  0.04683304 0.02849687 0.03832402 0.02941161
 0.09656904 1.         0.0613346 ]
[0.06397718 0.0853535  0.04015156 0.03073008 0.05709894 0.03555886
 0.06724288 0.04124595 0.04075483 0.03295305 0.03936543 0.05728974
 0.08144351 0.0613346  1.        ]]
```