Assignment # 1:	NLTK
Course:	CAP 6776-003 Information Retrieval
Professor:	Dr. Dingding Wang
Student Number:	Z23596812
Student Name:	Renee Raven
Term:	Fall 2022

For this assignment I've added also uploaded:

The code file: cap6776_a1_renee_raven.py

The output file: nltk_output.txt

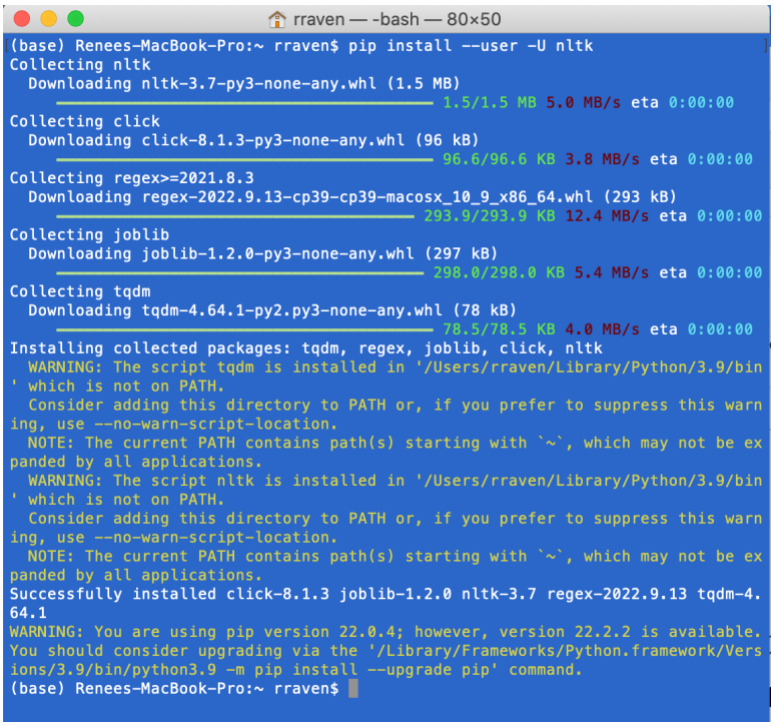
Thank you

Step 1: Install Python and NLTK

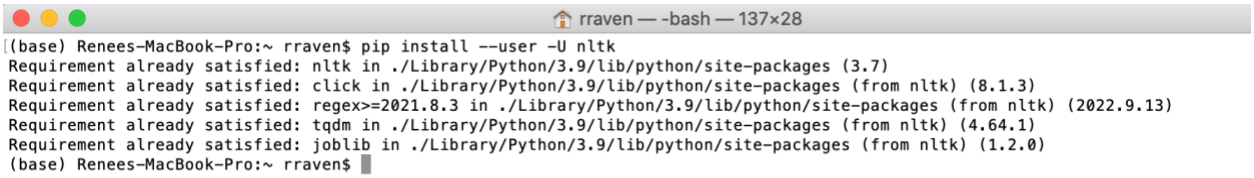
Python is already installed on my computer.

```
(base) Renees-MacBook-Pro:~ rraven$ /usr/local/bin/python3
Python 3.10.7 (v3.10.7:6cc6b13308, Sep  5 2022, 14:02:52) [Clang 13.0.0 (clang-1
300.0.29.30)] on darwin
Type "help", "copyright", "credits" or "license" for more information.
>>>
```

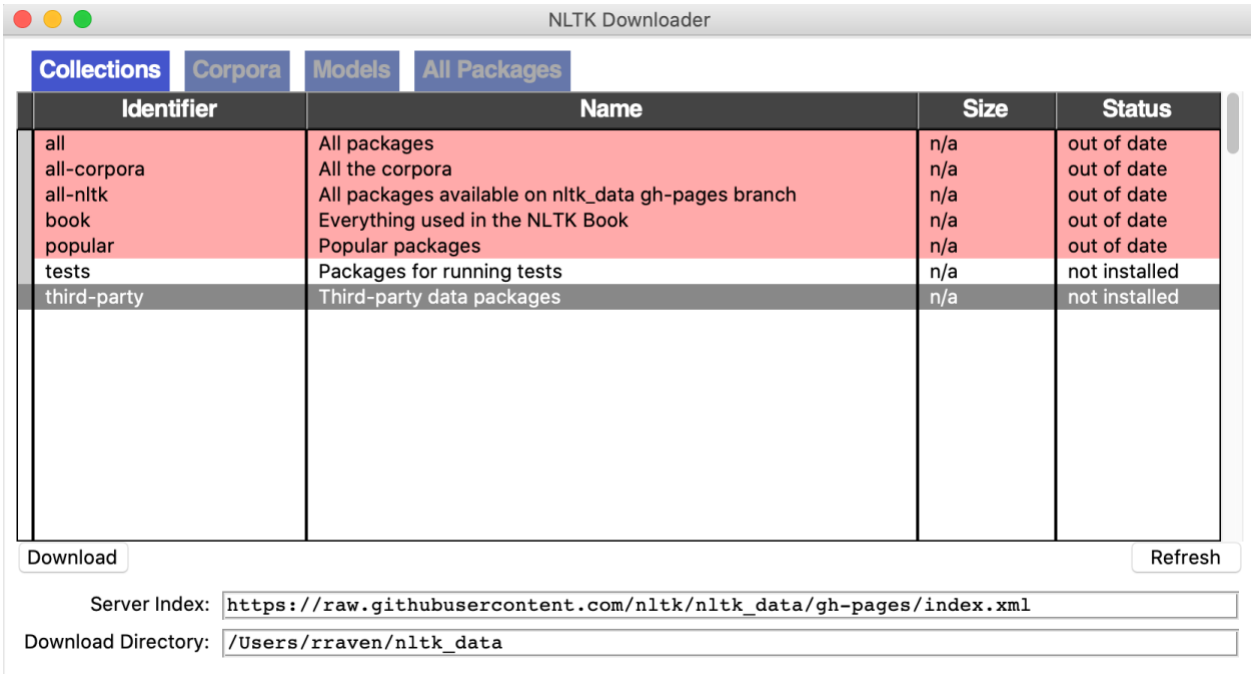
Installed nltk



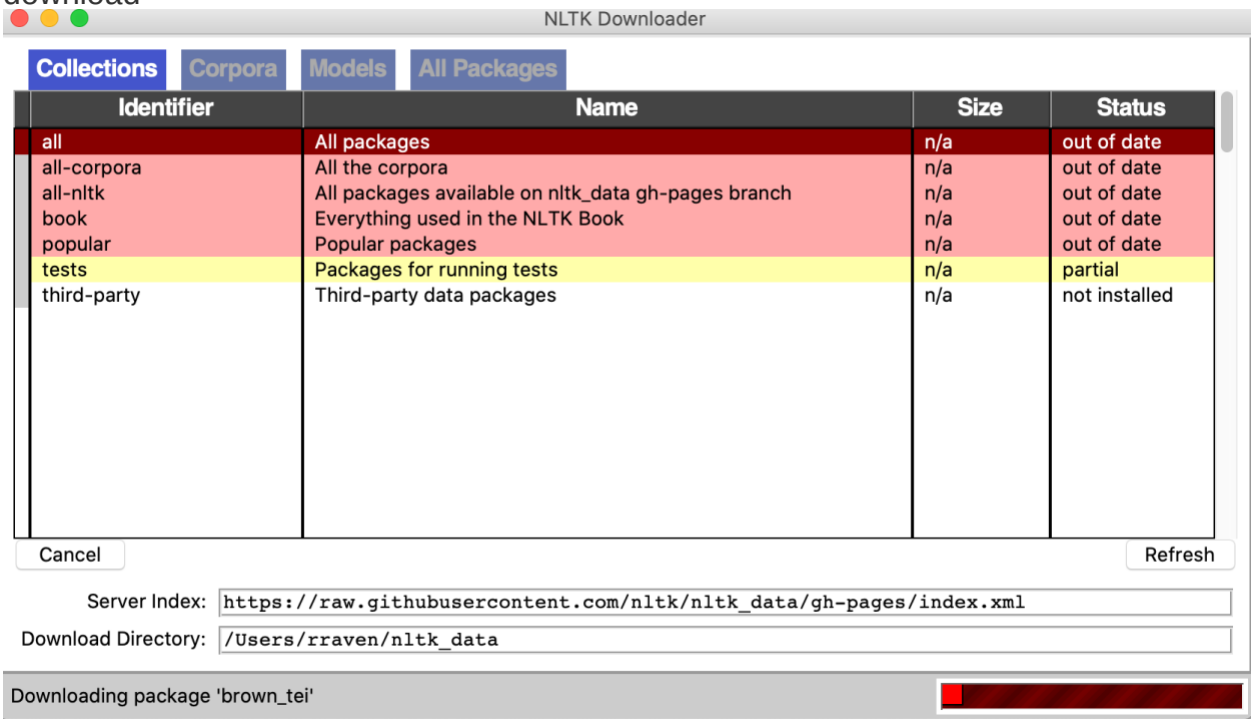
Verified installation of nltk



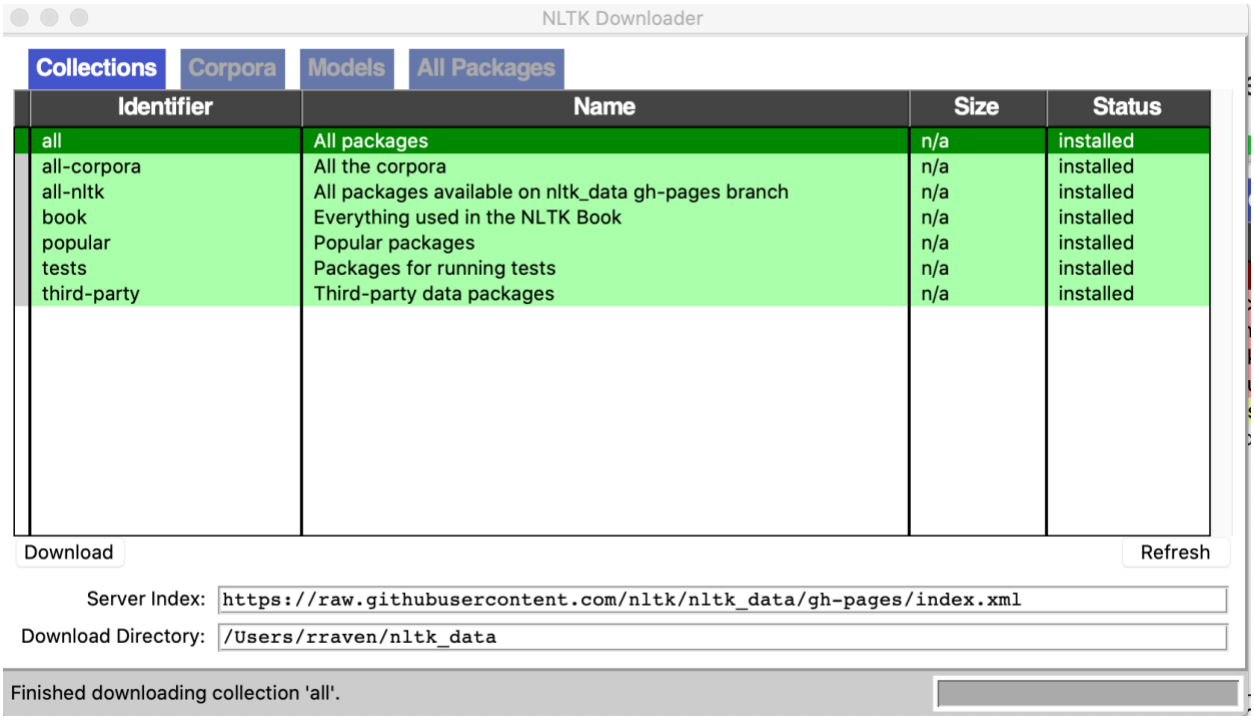
Used nltk.download() to open downloader



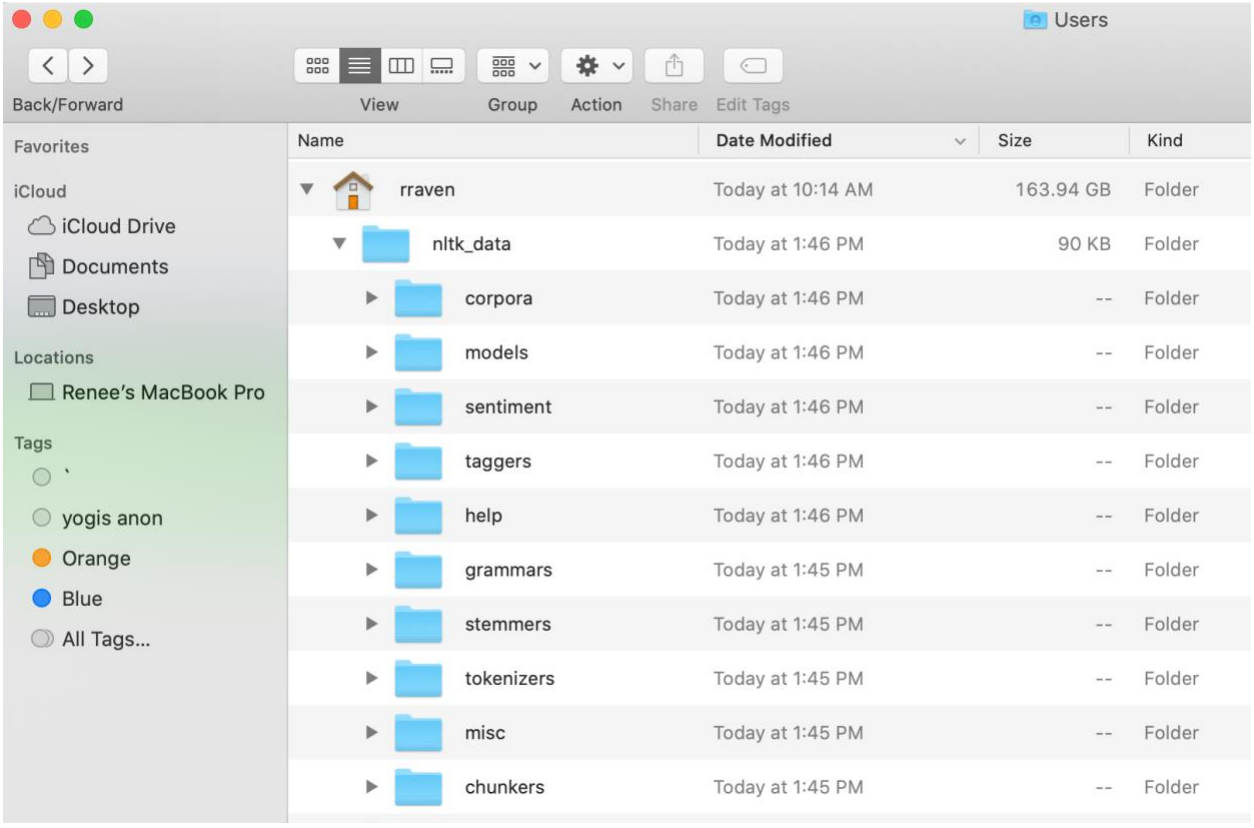
Clicked all packages and then download



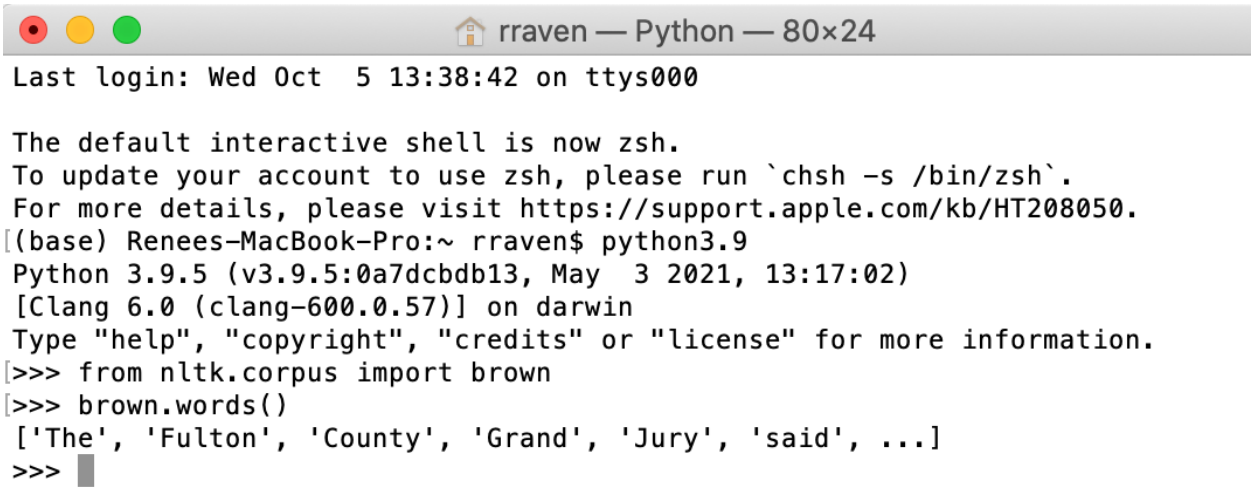
Downloaded all packages for nltk



Verification of installation



Verified nltk is working



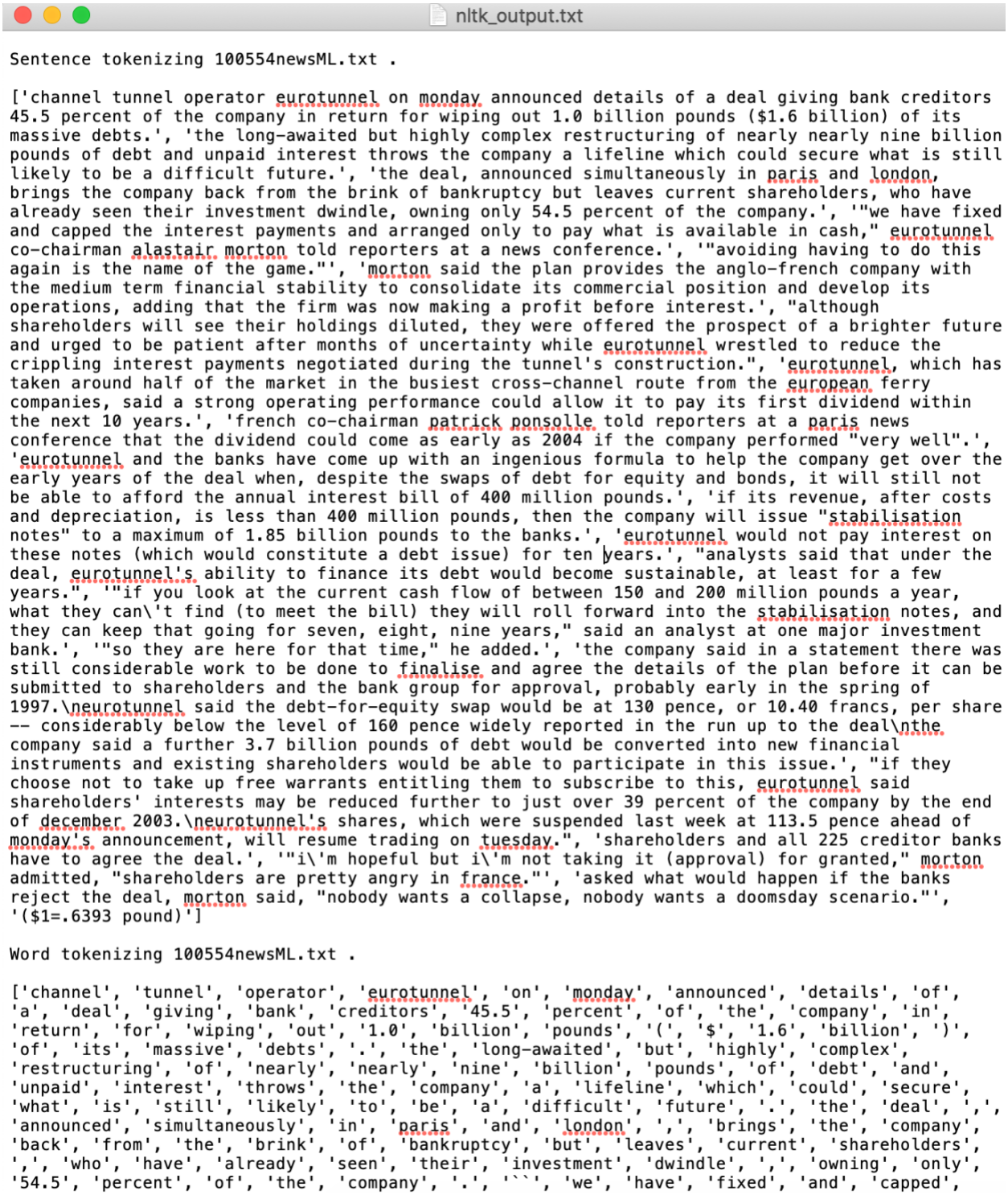
Step 2: Tokenize the documents into words remove stop words, and conduct stemming

Code used:

Get Started # Renee Raven Untitled-1 7

```python
# Renee Raven
# CAP6776 Assignment 1 nltk
# Code adapted from Shuwen Wang's (excellent) video

# import modules and libraries
import string
import os
from nltk.tokenize import sent_tokenize, word_tokenize
from nltk.corpus import stopwords
from nltk.stem.porter import PorterStemmer
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.feature_extraction.text import TfidfTransformer
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.metrics.pairwise import cosine_similarity

# save path to files for convenience
data_path = '/Users/rraven/info_retriv'

# initialize
token_dict = {}
all_stemmed_words = []
ps = PorterStemmer()
stop_words = set(stopwords.words("english")) | set(string.punctuation)
output_file = open('/Users/rraven/Desktop/nltk_output.txt','w')
text_lines = "\n"

for subdir, dirs, files in os.walk(data_path):
    for file in files:
        file_path = subdir + os.path.sep + file
        file_contents = open(file_path,'r')
        if '.txt' in file_path:
            text = file_contents.read()
            lowered = text.lower()
            token_dict[file] = lowered
            file_contents.close()
num_docs = len(token_dict)

doc_names = []
for file_name in token_dict.keys():
    doc_names.append(file_name)

for file in token_dict.keys():
    # tokenize
    words = word_tokenize(token_dict[file])
    print("number of words", len(words)) # just to make sure that are different files
    output_file.write((text_lines + "Sentence tokenizing" + ' ' + file + " .\n" + text_lines))
    output_file.write(str(sent_tokenize(token_dict[file])) + "\n")

    output_file.write(text_lines + "Word tokenizing" + ' ' + file + " .\n" + text_lines)
    output_file.write(str(words) + "\n")

    # remove stop words
    no_stop_words = []
    for w in words:
        if w not in stop_words:
            no_stop_words.append(w)
    output_file.write(text_lines + "Stop words removed from" + ' ' + file + " .\n" + text_lines)
    output_file.write(str(no_stop_words) + "\n")

    # stemming
    stemmed_words = []
    for w in words:
        if w not in stop_words:
            stemmed_words.append(ps.stem(w))
    output_file.write(text_lines + "Stemming" + ' ' + file + " .\n" + text_lines)
    output_file.write(str(stemmed_words) + "\n")

    all_stemmed_words.append(stemmed_words)

```

Screenshot of output of code that performs sentence tokenization, word tokenization, removal of stop words and stemming for 3 files:



Copied output of code that performs sentence tokenization, word tokenization, removal of stop words and stemming for 3 files:

Sentence tokenizing 100554newsML.txt .

['channel tunnel operator eurotunnel on monday announced details of a deal giving bank creditors 45.5 percent of the company in return for wiping out 1.0 billion pounds ($1.6 billion) of its massive debts.', 'the long-awaited but highly complex restructuring of nearly nearly nine billion pounds of debt and unpaid interest throws the company a lifeline which could secure what is still likely to be a difficult future.', 'the deal, announced simultaneously in paris and london, brings the company back from the brink of bankruptcy but leaves current shareholders, who have already seen their investment dwindle, owning only 54.5 percent of the company.', '"we have fixed and capped the interest payments and arranged only to pay what is available in cash," eurotunnel co-chairman alastair morton told reporters at a news conference.', '"avoiding having to do this again is the name of the game."', 'morton said the plan provides the anglo-french company with the medium term financial stability to consolidate its commercial position and develop its operations, adding that the firm was now making a profit before interest.', "although shareholders will see their holdings diluted, they were offered the prospect of a brighter future and urged to be patient after months of uncertainty while eurotunnel wrestled to reduce the crippling interest payments negotiated during the tunnel's construction.", 'eurotunnel, which has taken around half of the market in the busiest cross-channel route from the european ferry companies, said a strong operating performance could allow it to pay its first dividend within the next 10 years.', 'french co-chairman patrick ponsolle told reporters at a paris news conference that the dividend could come as early as 2004 if the company performed "very well".', 'eurotunnel and the banks have come up with an ingenious formula to help the company get over the early years of the deal when, despite the swaps of debt for equity and bonds, it will still not be able to afford the annual interest bill of 400 million pounds.', 'if its revenue, after costs and depreciation, is less than 400 million pounds, then the company will issue "stabilisation notes" to a maximum of 1.85 billion pounds to the banks.', 'eurotunnel would not pay interest on these notes (which would constitute a debt issue) for ten years.', "analysts said that under the deal, eurotunnel's ability to finance its debt would become sustainable, at least for a few years.", '"if you look at the current cash flow of between 150 and 200 million pounds a year, what they can\'t find (to meet the bill) they will roll forward into the stabilisation notes, and they can keep that going for seven, eight, nine years," said an analyst at one major investment bank.', '"so they are here for that time," he added.', 'the company said in a statement there was still considerable work to be done to finalise and agree the details of the plan before it can be submitted to shareholders and the bank group for approval, probably early in the spring of 1997.\neurotunnel said the debt-for-equity swap would be at 130 pence, or 10.40 francs, per share -- considerably below the level of 160 pence widely reported in the run up to the deal\nthe company said a further 3.7 billion pounds of debt would be converted into new financial instruments and existing shareholders would be able to participate in this issue.', "if they choose not to take up free warrants entitling them to subscribe to this, eurotunnel said shareholders' interests may be reduced further to just over 39 percent of the company by the end of december 2003.\neurotunnel's shares, which were suspended last week at 113.5 pence ahead of monday's announcement, will resume trading on tuesday.", 'shareholders and all 225 creditor banks have to agree the deal.', '"i\'m hopeful but i\'m not taking it (approval) for granted," morton admitted, "shareholders are pretty angry in france."', 'asked what would

happen if the banks reject the deal, morton said, "nobody wants a collapse, nobody wants a doomsday scenario."', '($1=.6393 pound)']

Word tokenizing 100554newsML.txt .

['channel', 'tunnel', 'operator', 'eurotunnel', 'on', 'monday', 'announced', 'details', 'of', 'a', 'deal', 'giving', 'bank', 'creditors', '45.5', 'percent', 'of', 'the', 'company', 'in', 'return', 'for', 'wiping', 'out', '1.0', 'billion', 'pounds', '(', '$', '1.6', 'billion', ')', 'of', 'its', 'massive', 'debts', '.', 'the', 'long-awaited', 'but', 'highly', 'complex', 'restructuring', 'of', 'nearly', 'nearly', 'nine', 'billion', 'pounds', 'of', 'debt', 'and', 'unpaid', 'interest', 'throws', 'the', 'company', 'a', 'lifeline', 'which', 'could', 'secure', 'what', 'is', 'still', 'likely', 'to', 'be', 'a', 'difficult', 'future', '.', 'the', 'deal', ',', 'announced', 'simultaneously', 'in', 'paris', 'and', 'london', ',', 'brings', 'the', 'company', 'back', 'from', 'the', 'brink', 'of', 'bankruptcy', 'but', 'leaves', 'current', 'shareholders', ',', 'who', 'have', 'already', 'seen', 'their', 'investment', 'dwindle', ',', 'owning', 'only', '54.5', 'percent', 'of', 'the', 'company', '.', '``', 'we', 'have', 'fixed', 'and', 'capped', 'the', 'interest', 'payments', 'and', 'arranged', 'only', 'to', 'pay', 'what', 'is', 'available', 'in', 'cash', ',', '"', 'eurotunnel', 'co-chairman', 'alastair', 'morton', 'told', 'reporters', 'at', 'a', 'news', 'conference', '.', '``', 'avoiding', 'having', 'to', 'do', 'this', 'again', 'is', 'the', 'name', 'of', 'the', 'game', '.', '"', 'morton', 'said', 'the', 'plan', 'provides', 'the', 'anglo-french', 'company', 'with', 'the', 'medium', 'term', 'financial', 'stability', 'to', 'consolidate', 'its', 'commercial', 'position', 'and', 'develop', 'its', 'operations', ',', 'adding', 'that', 'the', 'firm', 'was', 'now', 'making', 'a', 'profit', 'before', 'interest', '.', 'although', 'shareholders', 'will', 'see', 'their', 'holdings', 'diluted', ',', 'they', 'were', 'offered', 'the', 'prospect', 'of', 'a', 'brighter', 'future', 'and', 'urged', 'to', 'be', 'patient', 'after', 'months', 'of', 'uncertainty', 'while', 'eurotunnel', 'wrestled', 'to', 'reduce', 'the', 'crippling', 'interest', 'payments', 'negotiated', 'during', 'the', 'tunnel', "'s", 'construction', '.', 'eurotunnel', ',', 'which', 'has', 'taken', 'around', 'half', 'of', 'the', 'market', 'in', 'the', 'busiest', 'cross-channel', 'route', 'from', 'the', 'european', 'ferry', 'companies', ',', 'said', 'a', 'strong', 'operating', 'performance', 'could', 'allow', 'it', 'to', 'pay', 'its', 'first', 'dividend', 'within', 'the', 'next', '10', 'years', '.', 'french', 'co-chairman', 'patrick', 'ponsolle', 'told', 'reporters', 'at', 'a', 'paris', 'news', 'conference', 'that', 'the', 'dividend', 'could', 'come', 'as', 'early', 'as', '2004', 'if', 'the', 'company', 'performed', '``', 'very', 'well', '"', '.', 'eurotunnel', 'and', 'the', 'banks', 'have', 'come', 'up', 'with', 'an', 'ingenious', 'formula', 'to', 'help', 'the', 'company', 'get', 'over', 'the', 'early', 'years', 'of', 'the', 'deal', 'when', ',', 'despite', 'the', 'swaps', 'of', 'debt', 'for', 'equity', 'and', 'bonds', ',', 'it', 'will', 'still', 'not', 'be', 'able', 'to', 'afford', 'the', 'annual', 'interest', 'bill', 'of', '400', 'million', 'pounds', '.', 'if', 'its', 'revenue', ',', 'after', 'costs', 'and', 'depreciation', ',', 'is', 'less', 'than', '400', 'million', 'pounds', ',', 'then', 'the', 'company', 'will', 'issue', '``', 'stabilisation', 'notes', '"', 'to', 'a', 'maximum', 'of', '1.85', 'billion', 'pounds', 'to', 'the', 'banks', '.', 'eurotunnel', 'would', 'not', 'pay', 'interest', 'on', 'these', 'notes', '(', 'which', 'would', 'constitute', 'a', 'debt', 'issue', ')', 'for', 'ten', 'years', '.', 'analysts', 'said', 'that', 'under', 'the', 'deal', ',', 'eurotunnel', "'s", 'ability', 'to', 'finance', 'its', 'debt', 'would', 'become', 'sustainable', ',', 'at', 'least', 'for', 'a', 'few', 'years', '.', '``', 'if', 'you', 'look', 'at', 'the', 'current', 'cash', 'flow', 'of', 'between', '150', 'and', '200', 'million', 'pounds', 'a', 'year', ',', 'what', 'they', 'ca', "n't", 'find', '(', 'to', 'meet', 'the', 'bill', ')', 'they', 'will', 'roll', 'forward', 'into', 'the', 'stabilisation', 'notes', ',', 'and', 'they', 'can', 'keep', 'that', 'going', 'for', 'seven', ',', 'eight', ',', 'nine', 'years', ',', '"', 'said', 'an', 'analyst', 'at', 'one', 'major', 'investment', 'bank', '.', '``', 'so', 'they', 'are', 'here', 'for', 'that', 'time', ',', '"', 'he', 'added', '.', 'the', 'company', 'said', 'in', 'a', 'statement', 'there', 'was', 'still', 'considerable', 'work', 'to', 'be', 'done', 'to', 'finalise', 'and', 'agree', 'the', 'details', 'of', 'the', 'plan', 'before', 'it', 'can', 'be', 'submitted', 'to', 'shareholders', 'and', 'the', 'bank', 'group', 'for', 'approval', ',', 'probably', 'early', 'in', 'the', 'spring', 'of', '1997.', 'eurotunnel', 'said', 'the', 'debt-for-equity', 'swap', 'would', 'be', 'at', '130', 'pence', ',', 'or', '10.40', 'francs', ',', 'per', 'share', '--', 'considerably', 'below', 'the', 'level', 'of', '160', 'pence', 'widely',

'reported', 'in', 'the', 'run', 'up', 'to', 'the', 'deal', 'the', 'company', 'said', 'a', 'further', '3.7', 'billion', 'pounds', 'of', 'debt', 'would', 'be', 'converted', 'into', 'new', 'financial', 'instruments', 'and', 'existing', 'shareholders', 'would', 'be', 'able', 'to', 'participate', 'in', 'this', 'issue', '.', 'if', 'they', 'choose', 'not', 'to', 'take', 'up', 'free', 'warrants', 'entitling', 'them', 'to', 'subscribe', 'to', 'this', ',', 'eurotunnel', 'said', 'shareholders', '"', 'interests', 'may', 'be', 'reduced', 'further', 'to', 'just', 'over', '39', 'percent', 'of', 'the', 'company', 'by', 'the', 'end', 'of', 'december', '2003.', 'eurotunnel', "'s", 'shares', ',', 'which', 'were', 'suspended', 'last', 'week', 'at', '113.5', 'pence', 'ahead', 'of', 'monday', "'s", 'announcement', ',', 'will', 'resume', 'trading', 'on', 'tuesday', '.', 'shareholders', 'and', 'all', '225', 'creditor', 'banks', 'have', 'to', 'agree', 'the', 'deal', '.', '``', 'i', "'m", 'hopeful', 'but', 'i', "'m", 'not', 'taking', 'it', '(', 'approval', ')', 'for', 'granted', ',', '"', 'morton', 'admitted', ',', '``', 'shareholders', 'are', 'pretty', 'angry', 'in', 'france', '.', '"', 'asked', 'what', 'would', 'happen', 'if', 'the', 'banks', 'reject', 'the', 'deal', ',', 'morton', 'said', ',', '``', 'nobody', 'wants', 'a', 'collapse', ',', 'nobody', 'wants', 'a', 'doomsday', 'scenario', '.', '"', '(', '$', '1=.6393', 'pound', ')']

Stop words removed from 100554newsML.txt .

['channel', 'tunnel', 'operator', 'eurotunnel', 'monday', 'announced', 'details', 'deal', 'giving', 'bank', 'creditors', '45.5', 'percent', 'company', 'return', 'wiping', '1.0', 'billion', 'pounds', '1.6', 'billion', 'massive', 'debts', 'long-awaited', 'highly', 'complex', 'restructuring', 'nearly', 'nearly', 'nine', 'billion', 'pounds', 'debt', 'unpaid', 'interest', 'throws', 'company', 'lifeline', 'could', 'secure', 'still', 'likely', 'difficult', 'future', 'deal', 'announced', 'simultaneously', 'paris', 'london', 'brings', 'company', 'back', 'brink', 'bankruptcy', 'leaves', 'current', 'shareholders', 'already', 'seen', 'investment', 'dwindle', 'owning', '54.5', 'percent', 'company', '``', 'fixed', 'capped', 'interest', 'payments', 'arranged', 'pay', 'available', 'cash', '"', 'eurotunnel', 'co-chairman', 'alastair', 'morton', 'told', 'reporters', 'news', 'conference', '``', 'avoiding', 'name', 'game', '"', 'morton', 'said', 'plan', 'provides', 'anglo-french', 'company', 'medium', 'term', 'financial', 'stability', 'consolidate', 'commercial', 'position', 'develop', 'operations', 'adding', 'firm', 'making', 'profit', 'interest', 'although', 'shareholders', 'see', 'holdings', 'diluted', 'offered', 'prospect', 'brighter', 'future', 'urged', 'patient', 'months', 'uncertainty', 'eurotunnel', 'wrestled', 'reduce', 'crippling', 'interest', 'payments', 'negotiated', 'tunnel', "'s", 'construction', 'eurotunnel', 'taken', 'around', 'half', 'market', 'busiest', 'cross-channel', 'route', 'european', 'ferry', 'companies', 'said', 'strong', 'operating', 'performance', 'could', 'allow', 'pay', 'first', 'dividend', 'within', 'next', '10', 'years', 'french', 'co-chairman', 'patrick', 'ponsolle', 'told', 'reporters', 'paris', 'news', 'conference', 'dividend', 'could', 'come', 'early', '2004', 'company', 'performed', '``', 'well', '"', 'eurotunnel', 'banks', 'come', 'ingenious', 'formula', 'help', 'company', 'get', 'early', 'years', 'deal', 'despite', 'swaps', 'debt', 'equity', 'bonds', 'still', 'able', 'afford', 'annual', 'interest', 'bill', '400', 'million', 'pounds', 'revenue', 'costs', 'depreciation', 'less', '400', 'million', 'pounds', 'company', 'issue', '``', 'stabilisation', 'notes', '"', 'maximum', '1.85', 'billion', 'pounds', 'banks', 'eurotunnel', 'would', 'pay', 'interest', 'notes', 'would', 'constitute', 'debt', 'issue', 'ten', 'years', 'analysts', 'said', 'deal', 'eurotunnel', "'s", 'ability', 'finance', 'debt', 'would', 'become', 'sustainable', 'least', 'years', '``', 'look', 'current', 'cash', 'flow', '150', '200', 'million', 'pounds', 'year', 'ca', "n't", 'find', 'meet', 'bill', 'roll', 'forward', 'stabilisation', 'notes', 'keep', 'going', 'seven', 'eight', 'nine', 'years', '"', 'said', 'analyst', 'one', 'major', 'investment', 'bank', '``', 'time', '"', 'added', 'company', 'said', 'statement', 'still', 'considerable', 'work', 'done', 'finalise', 'agree', 'details', 'plan', 'submitted', 'shareholders', 'bank', 'group', 'approval', 'probably', 'early', 'spring', '1997.', 'eurotunnel', 'said', 'debt-for-equity', 'swap', 'would', '130', 'pence', '10.40', 'francs', 'per', 'share', '--', 'considerably', 'level', '160', 'pence', 'widely', 'reported', 'run', 'deal', 'company', 'said', '3.7', 'billion', 'pounds', 'debt', 'would', 'converted', 'new', 'financial', 'instruments', 'existing', 'shareholders', 'would', 'able', 'participate', 'issue', 'choose', 'take', 'free', 'warrants', 'entitling', 'subscribe', 'eurotunnel', 'said', 'shareholders', 'interests', 'may', 'reduced', '39', 'percent', 'company', 'end', 'december', '2003.', 'eurotunnel', "'s", 'shares', 'suspended',

'last', 'week', '113.5', 'pence', 'ahead', 'monday', "'s", 'announcement', 'resume', 'trading', 'tuesday', 'shareholders', '225', 'creditor', 'banks', 'agree', 'deal', '``', "'m", 'hopeful', "'m", 'taking', 'approval', 'granted', '"', 'morton', 'admitted', '``', 'shareholders', 'pretty', 'angry', 'france', '"', 'asked', 'would', 'happen', 'banks', 'reject', 'deal', 'morton', 'said', '``', 'nobody', 'wants', 'collapse', 'nobody', 'wants', 'doomsday', 'scenario', '"', '1=.6393', 'pound']

Stemming 100554newsML.txt .

['channel', 'tunnel', 'oper', 'eurotunnel', 'monday', 'announc', 'detail', 'deal', 'give', 'bank', 'creditor', '45.5', 'percent', 'compani', 'return', 'wipe', '1.0', 'billion', 'pound', '1.6', 'billion', 'massiv', 'debt', 'long-await', 'highli', 'complex', 'restructur', 'nearli', 'nearli', 'nine', 'billion', 'pound', 'debt', 'unpaid', 'interest', 'throw', 'compani', 'lifelin', 'could', 'secur', 'still', 'like', 'difficult', 'futur', 'deal', 'announc', 'simultan', 'pari', 'london', 'bring', 'compani', 'back', 'brink', 'bankruptci', 'leav', 'current', 'sharehold', 'alreadi', 'seen', 'invest', 'dwindl', 'own', '54.5', 'percent', 'compani', '``', 'fix', 'cap', 'interest', 'payment', 'arrang', 'pay', 'avail', 'cash', '"', 'eurotunnel', 'co-chairman', 'alastair', 'morton', 'told', 'report', 'news', 'confer', '``', 'avoid', 'name', 'game', '"', 'morton', 'said', 'plan', 'provid', 'anglo-french', 'compani', 'medium', 'term', 'financi', 'stabil', 'consolid', 'commerci', 'posit', 'develop', 'oper', 'ad', 'firm', 'make', 'profit', 'interest', 'although', 'sharehold', 'see', 'hold', 'dilut', 'offer', 'prospect', 'brighter', 'futur', 'urg', 'patient', 'month', 'uncertainti', 'eurotunnel', 'wrestl', 'reduc', 'crippl', 'interest', 'payment', 'negoti', 'tunnel', "'s", 'construct', 'eurotunnel', 'taken', 'around', 'half', 'market', 'busiest', 'cross-channel', 'rout', 'european', 'ferri', 'compani', 'said', 'strong', 'oper', 'perform', 'could', 'allow', 'pay', 'first', 'dividend', 'within', 'next', '10', 'year', 'french', 'co-chairman', 'patrick', 'ponsol', 'told', 'report', 'pari', 'news', 'confer', 'dividend', 'could', 'come', 'earli', '2004', 'compani', 'perform', '``', 'well', '"', 'eurotunnel', 'bank', 'come', 'ingeni', 'formula', 'help', 'compani', 'get', 'earli', 'year', 'deal', 'despit', 'swap', 'debt', 'equiti', 'bond', 'still', 'abl', 'afford', 'annual', 'interest', 'bill', '400', 'million', 'pound', 'revenu', 'cost', 'depreci', 'less', '400', 'million', 'pound', 'compani', 'issu', '``', 'stabilis', 'note', '"', 'maximum', '1.85', 'billion', 'pound', 'bank', 'eurotunnel', 'would', 'pay', 'interest', 'note', 'would', 'constitut', 'debt', 'issu', 'ten', 'year', 'analyst', 'said', 'deal', 'eurotunnel', "'s", 'abil', 'financ', 'debt', 'would', 'becom', 'sustain', 'least', 'year', '``', 'look', 'current', 'cash', 'flow', '150', '200', 'million', 'pound', 'year', 'ca', "n't", 'find', 'meet', 'bill', 'roll', 'forward', 'stabilis', 'note', 'keep', 'go', 'seven', 'eight', 'nine', 'year', '"', 'said', 'analyst', 'one', 'major', 'invest', 'bank', '``', 'time', '"', 'ad', 'compani', 'said', 'statement', 'still', 'consider', 'work', 'done', 'finalis', 'agre', 'detail', 'plan', 'submit', 'sharehold', 'bank', 'group', 'approv', 'probabl', 'earli', 'spring', '1997.', 'eurotunnel', 'said', 'debt-for-equ', 'swap', 'would', '130', 'penc', '10.40', 'franc', 'per', 'share', '--', 'consider', 'level', '160', 'penc', 'wide', 'report', 'run', 'deal', 'compani', 'said', '3.7', 'billion', 'pound', 'debt', 'would', 'convert', 'new', 'financi', 'instrument', 'exist', 'sharehold', 'would', 'abl', 'particip', 'issu', 'choos', 'take', 'free', 'warrant', 'entitl', 'subscrib', 'eurotunnel', 'said', 'sharehold', 'interest', 'may', 'reduc', '39', 'percent', 'compani', 'end', 'decemb', '2003.', 'eurotunnel', "'s", 'share', 'suspend', 'last', 'week', '113.5', 'penc', 'ahead', 'monday', "'s", 'announc', 'resum', 'trade', 'tuesday', 'sharehold', '225', 'creditor', 'bank', 'agre', 'deal', '``', "'m", 'hope', "'m", 'take', 'approv', 'grant', '"', 'morton', 'admit', '``', 'sharehold', 'pretti', 'angri', 'franc', '"', 'ask', 'would', 'happen', 'bank', 'reject', 'deal', 'morton', 'said', '``', 'nobodi', 'want', 'collaps', 'nobodi', 'want', 'doomsday', 'scenario', '"', '1=.6393', 'pound']

Sentence tokenizing 100593newsML.txt .

['anglo-french channel tunnel operator eurotunnel monday announced a deal giving its creditor banks 45.5 percent of the company in return for wiping out one billion pounds ($1.56 billion) of its debt.', 'the long-awaited restructuring brings to an end months of wrangling between eurotunnel and the 225 banks to which it owes nearly nine billion pounds ($14.1 billion).', 'the deal, announced simultaneously in paris and london, brings the company back from the brink of insolvency but leaves shareholders owning only 54.5 percent of the company.', '"the restructuring plan provides eurotunnel with the medium-term financial stability to allow it to consolidate its substantial commercial achievements to date and to develop its operations," eurotunnel co-chairman alastair morton said.', 'the firm was now making a profit before interest, he added.', "although shareholders will see their interests diluted, they were offered the prospect of a brighter future after months of uncertainty while eurotunnel wrestled to reduce crippling interest payments negotiated during the tunnel's construction.", 'eurotunnel, which has taken around half the cross-channel market from the european ferry companies, said a strong operating performance could allow it to pay its first dividend within the next 10 years.', "french co-chairman patrick ponsolle said shareholders would have to be patient before they could reap the benefits of the company's success.", 'he called the debt restructuring plan "an acceptable compromise" for holders of eurotunnel shares.', "the company said there was still considerable work to be done to finalise and agree on the details of the plan before it can be submitted to shareholders and the full 225 bank syndicate for approval, probably early in 1997.\nmonday's announcement followed two weeks of highly secretive negotiations between eurotunnel and its six leading banks.", 'this was extended to the 24 "instructing banks" at a meeting late last week in london.', 'eurotunnel said the debt-for-equity swap would be at 130 pence, or 10.40 francs, per share.', 'that is considerably below the level of around 160 pence widely reported before announcement of the deal, and will reduce outstanding debt of 8.7 billion pounds ($13.6 billion) by 1.0 billion ($1.56 billion).', 'the company said a further 3.7 billion pounds ($5.8 billion) of debt would be converted into new financial instruments and existing shareholders would be able to participate in this issue.', "if they choose not to take up free warrants entitling them to subscribe to this, eurotunnel said shareholders' interests may be reduced further to just over 39 percent of the company by the end of december 2003.\neurotunnel's shares, which were suspended last week at 113.5 pence ahead of monday's announcement, should resume trading on tuesday, the company said."]

Word tokenizing 100593newsML.txt .

['anglo-french', 'channel', 'tunnel', 'operator', 'eurotunnel', 'monday', 'announced', 'a', 'deal', 'giving', 'its', 'creditor', 'banks', '45.5', 'percent', 'of', 'the', 'company', 'in', 'return', 'for', 'wiping', 'out', 'one', 'billion', 'pounds', '(', '$', '1.56', 'billion', ')', 'of', 'its', 'debt', '.', 'the', 'long-awaited', 'restructuring', 'brings', 'to', 'an', 'end', 'months', 'of', 'wrangling', 'between', 'eurotunnel', 'and', 'the', '225', 'banks', 'to', 'which', 'it', 'owes', 'nearly', 'nine', 'billion', 'pounds', '(', '$', '14.1', 'billion', ')', '.', 'the', 'deal', ',', 'announced', 'simultaneously', 'in', 'paris', 'and', 'london', ',', 'brings', 'the', 'company', 'back', 'from', 'the', 'brink', 'of', 'insolvency', 'but', 'leaves', 'shareholders', 'owning', 'only', '54.5', 'percent', 'of', 'the', 'company', '.', '``', 'the', 'restructuring', 'plan', 'provides', 'eurotunnel', 'with', 'the', 'medium-term', 'financial', 'stability', 'to', 'allow', 'it', 'to', 'consolidate', 'its', 'substantial', 'commercial', 'achievements', 'to', 'date', 'and', 'to', 'develop', 'its', 'operations', ',', '"', 'eurotunnel', 'co-chairman', 'alastair', 'morton', 'said', '.', 'the', 'firm', 'was', 'now', 'making', 'a', 'profit', 'before', 'interest', ',', 'he', 'added', '.', 'although', 'shareholders', 'will', 'see', 'their', 'interests', 'diluted', ',', 'they', 'were', 'offered', 'the', 'prospect', 'of', 'a', 'brighter', 'future', 'after', 'months', 'of', 'uncertainty', 'while', 'eurotunnel', 'wrestled', 'to', 'reduce', 'crippling', 'interest', 'payments', 'negotiated', 'during', 'the', 'tunnel', "'s", 'construction', '.', 'eurotunnel', ',', 'which', 'has', 'taken', 'around', 'half', 'the', 'cross-

channel', 'market', 'from', 'the', 'european', 'ferry', 'companies', ',', 'said', 'a', 'strong', 'operating', 'performance', 'could', 'allow', 'it', 'to', 'pay', 'its', 'first', 'dividend', 'within', 'the', 'next', '10', 'years', '.', 'french', 'co-chairman', 'patrick', 'ponsolle', 'said', 'shareholders', 'would', 'have', 'to', 'be', 'patient', 'before', 'they', 'could', 'reap', 'the', 'benefits', 'of', 'the', 'company', "'s", 'success', '.', 'he', 'called', 'the', 'debt', 'restructuring', 'plan', '``', 'an', 'acceptable', 'compromise', '"', 'for', 'holders', 'of', 'eurotunnel', 'shares', '.', 'the', 'company', 'said', 'there', 'was', 'still', 'considerable', 'work', 'to', 'be', 'done', 'to', 'finalise', 'and', 'agree', 'on', 'the', 'details', 'of', 'the', 'plan', 'before', 'it', 'can', 'be', 'submitted', 'to', 'shareholders', 'and', 'the', 'full', '225', 'bank', 'syndicate', 'for', 'approval', ',', 'probably', 'early', 'in', '1997.', 'monday', "'s", 'announcement', 'followed', 'two', 'weeks', 'of', 'highly', 'secretive', 'negotiations', 'between', 'eurotunnel', 'and', 'its', 'six', 'leading', 'banks', '.', 'this', 'was', 'extended', 'to', 'the', '24', '``', 'instructing', 'banks', '"', 'at', 'a', 'meeting', 'late', 'last', 'week', 'in', 'london', '.', 'eurotunnel', 'said', 'the', 'debt-for-equity', 'swap', 'would', 'be', 'at', '130', 'pence', ',', 'or', '10.40', 'francs', ',', 'per', 'share', '.', 'that', 'is', 'considerably', 'below', 'the', 'level', 'of', 'around', '160', 'pence', 'widely', 'reported', 'before', 'announcement', 'of', 'the', 'deal', ',', 'and', 'will', 'reduce', 'outstanding', 'debt', 'of', '8.7', 'billion', 'pounds', '(', '$', '13.6', 'billion', ')', 'by', '1.0', 'billion', '(', '$', '1.56', 'billion', ')', '.', 'the', 'company', 'said', 'a', 'further', '3.7', 'billion', 'pounds', '(', '$', '5.8', 'billion', ')', 'of', 'debt', 'would', 'be', 'converted', 'into', 'new', 'financial', 'instruments', 'and', 'existing', 'shareholders', 'would', 'be', 'able', 'to', 'participate', 'in', 'this', 'issue', '.', 'if', 'they', 'choose', 'not', 'to', 'take', 'up', 'free', 'warrants', 'entitling', 'them', 'to', 'subscribe', 'to', 'this', ',', 'eurotunnel', 'said', 'shareholders', '"', 'interests', 'may', 'be', 'reduced', 'further', 'to', 'just', 'over', '39', 'percent', 'of', 'the', 'company', 'by', 'the', 'end', 'of', 'december', '2003.', 'eurotunnel', "'s", 'shares', ',', 'which', 'were', 'suspended', 'last', 'week', 'at', '113.5', 'pence', 'ahead', 'of', 'monday', "'s", 'announcement', ',', 'should', 'resume', 'trading', 'on', 'tuesday', ',', 'the', 'company', 'said', '.']

Stop words removed from 100593newsML.txt .

['anglo-french', 'channel', 'tunnel', 'operator', 'eurotunnel', 'monday', 'announced', 'deal', 'giving', 'creditor', 'banks', '45.5', 'percent', 'company', 'return', 'wiping', 'one', 'billion', 'pounds', '1.56', 'billion', 'debt', 'long-awaited', 'restructuring', 'brings', 'end', 'months', 'wrangling', 'eurotunnel', '225', 'banks', 'owes', 'nearly', 'nine', 'billion', 'pounds', '14.1', 'billion', 'deal', 'announced', 'simultaneously', 'paris', 'london', 'brings', 'company', 'back', 'brink', 'insolvency', 'leaves', 'shareholders', 'owning', '54.5', 'percent', 'company', '``', 'restructuring', 'plan', 'provides', 'eurotunnel', 'medium-term', 'financial', 'stability', 'allow', 'consolidate', 'substantial', 'commercial', 'achievements', 'date', 'develop', 'operations', '"', 'eurotunnel', 'co-chairman', 'alastair', 'morton', 'said', 'firm', 'making', 'profit', 'interest', 'added', 'although', 'shareholders', 'see', 'interests', 'diluted', 'offered', 'prospect', 'brighter', 'future', 'months', 'uncertainty', 'eurotunnel', 'wrestled', 'reduce', 'crippling', 'interest', 'payments', 'negotiated', 'tunnel', "'s", 'construction', 'eurotunnel', 'taken', 'around', 'half', 'cross-channel', 'market', 'european', 'ferry', 'companies', 'said', 'strong', 'operating', 'performance', 'could', 'allow', 'pay', 'first', 'dividend', 'within', 'next', '10', 'years', 'french', 'co-chairman', 'patrick', 'ponsolle', 'said', 'shareholders', 'would', 'patient', 'could', 'reap', 'benefits', 'company', "'s", 'success', 'called', 'debt', 'restructuring', 'plan', '``', 'acceptable', 'compromise', '"', 'holders', 'eurotunnel', 'shares', 'company', 'said', 'still', 'considerable', 'work', 'done', 'finalise', 'agree', 'details', 'plan', 'submitted', 'shareholders', 'full', '225', 'bank', 'syndicate', 'approval', 'probably', 'early', '1997.', 'monday', "'s", 'announcement', 'followed', 'two', 'weeks', 'highly', 'secretive', 'negotiations', 'eurotunnel', 'six', 'leading', 'banks', 'extended', '24', '``', 'instructing', 'banks', '"', 'meeting', 'late', 'last', 'week', 'london', 'eurotunnel', 'said', 'debt-for-equity', 'swap', 'would', '130', 'pence', '10.40', 'francs', 'per', 'share', 'considerably', 'level', 'around', '160', 'pence', 'widely', 'reported', 'announcement', 'deal', 'reduce', 'outstanding', 'debt', '8.7', 'billion', 'pounds', '13.6', 'billion', '1.0', 'billion',

'1.56', 'billion', 'company', 'said', '3.7', 'billion', 'pounds', '5.8', 'billion', 'debt', 'would', 'converted', 'new', 'financial', 'instruments', 'existing', 'shareholders', 'would', 'able', 'participate', 'issue', 'choose', 'take', 'free', 'warrants', 'entitling', 'subscribe', 'eurotunnel', 'said', 'shareholders', 'interests', 'may', 'reduced', '39', 'percent', 'company', 'end', 'december', '2003.', 'eurotunnel', "'s", 'shares', 'suspended', 'last', 'week', '113.5', 'pence', 'ahead', 'monday', "'s", 'announcement', 'resume', 'trading', 'tuesday', 'company', 'said']

Stemming 100593newsML.txt .

['anglo-french', 'channel', 'tunnel', 'oper', 'eurotunnel', 'monday', 'announc', 'deal', 'give', 'creditor', 'bank', '45.5', 'percent', 'compani', 'return', 'wipe', 'one', 'billion', 'pound', '1.56', 'billion', 'debt', 'long-await', 'restructur', 'bring', 'end', 'month', 'wrangl', 'eurotunnel', '225', 'bank', 'owe', 'nearli', 'nine', 'billion', 'pound', '14.1', 'billion', 'deal', 'announc', 'simultan', 'pari', 'london', 'bring', 'compani', 'back', 'brink', 'insolv', 'leav', 'sharehold', 'own', '54.5', 'percent', 'compani', '``', 'restructur', 'plan', 'provid', 'eurotunnel', 'medium-term', 'financi', 'stabil', 'allow', 'consolid', 'substanti', 'commerci', 'achiev', 'date', 'develop', 'oper', '"', 'eurotunnel', 'co-chairman', 'alastair', 'morton', 'said', 'firm', 'make', 'profit', 'interest', 'ad', 'although', 'sharehold', 'see', 'interest', 'dilut', 'offer', 'prospect', 'brighter', 'futur', 'month', 'uncertainti', 'eurotunnel', 'wrestl', 'reduc', 'crippl', 'interest', 'payment', 'negoti', 'tunnel', "'s", 'construct', 'eurotunnel', 'taken', 'around', 'half', 'cross-channel', 'market', 'european', 'ferri', 'compani', 'said', 'strong', 'oper', 'perform', 'could', 'allow', 'pay', 'first', 'dividend', 'within', 'next', '10', 'year', 'french', 'co-chairman', 'patrick', 'ponsol', 'said', 'sharehold', 'would', 'patient', 'could', 'reap', 'benefit', 'compani', "'s", 'success', 'call', 'debt', 'restructur', 'plan', '``', 'accept', 'compromis', '"', 'holder', 'eurotunnel', 'share', 'compani', 'said', 'still', 'consider', 'work', 'done', 'finalis', 'agre', 'detail', 'plan', 'submit', 'sharehold', 'full', '225', 'bank', 'syndic', 'approv', 'probabl', 'earli', '1997.', 'monday', "'s", 'announc', 'follow', 'two', 'week', 'highli', 'secret', 'negoti', 'eurotunnel', 'six', 'lead', 'bank', 'extend', '24', '``', 'instruct', 'bank', '"', 'meet', 'late', 'last', 'week', 'london', 'eurotunnel', 'said', 'debt-for-equ', 'swap', 'would', '130', 'penc', '10.40', 'franc', 'per', 'share', 'consider', 'level', 'around', '160', 'penc', 'wide', 'report', 'announc', 'deal', 'reduc', 'outstand', 'debt', '8.7', 'billion', 'pound', '13.6', 'billion', '1.0', 'billion', '1.56', 'billion', 'compani', 'said', '3.7', 'billion', 'pound', '5.8', 'billion', 'debt', 'would', 'convert', 'new', 'financi', 'instrument', 'exist', 'sharehold', 'would', 'abl', 'particip', 'issu', 'choos', 'take', 'free', 'warrant', 'entitl', 'subscrib', 'eurotunnel', 'said', 'sharehold', 'interest', 'may', 'reduc', '39', 'percent', 'compani', 'end', 'decemb', '2003.', 'eurotunnel', "'s", 'share', 'suspend', 'last', 'week', '113.5', 'penc', 'ahead', 'monday', "'s", 'announc', 'resum', 'trade', 'tuesday', 'compani', 'said']

Sentence tokenizing 100618newsML.txt .

['anglo-french channel tunnel operator eurotunnel on monday announced a deal giving creditor banks 45.5 percent of the company in return for wiping out one billion pounds ($1.56 billion) of its debt mountain.', 'the long-awaited restructuring brings to an end months of wrangling between eurotunnel and the 225 banks to which it owes nearly nine billion pounds.', 'the deal, announced simultaneously in paris and london, brings the company back from the brink of insolvency but leaves shareholders owning only 54.5 percent of the company.', '"the restructuring plan provides eurotunnel with the medium term financial stability to allow it to consolidate its substantial commercial achievements to date and to develop its operations," eurotunnel co-chairman alastair morton said.', 'the firm was now making a profit before interest, he added.', "although shareholders will see their interests diluted, they were offered the prospect of a brighter future after months of uncertainty while eurotunnel wrestled to reduce crippling interest payments

negotiated during the tunnel's construction.", 'eurotunnel, which has taken around half the cross-channel market from the european ferry companies, said a strong operating performance could allow it to pay its first dividend within the next 10 years.', "french co-chairman patrick ponsolle said shareholders would have to be patient before they could reap the benefits of the company's success.", 'he called the debt restructuring plan "an acceptable compromise" for holders of eurotunnel shares.', "the company said in a statement there was still considerable work to be done to finalise and agree the details of the plan before it can be submitted to shareholders and the full 225 bank syndicate for approval, probably early in 1997.\nmonday's announcement followed two weeks of highly secretive negotiations between eurotunnel and its six leading banks.", 'this was extended to the 24 "instructing banks" at a meeting late last week in london.', 'eurotunnel said the debt-for-equity swap would be at 130 pence, or 10.40 francs, per share.', 'that is considerably below the level of around 160 pence widely reported in the run up to the deal, and will reduce outstanding debt of 8.7 billion pounds by 1.0 billion.', 'the company said a further 3.7 billion pounds of debt would be converted into new financial instruments and existing shareholders would be able to participate in this issue.', "if they choose not to take up free warrants entitling them to subscribe to this, eurotunnel said shareholders' interests may be reduced further to just over 39 percent of the company by the end of december 2003.\neurotunnel's shares, which were suspended last week at 113.5 pence ahead of monday's announcement, should resume trading on tuesday, the company said.", '($1=.6393 pound)']

Word tokenizing 100618newsML.txt .

['anglo-french', 'channel', 'tunnel', 'operator', 'eurotunnel', 'on', 'monday', 'announced', 'a', 'deal', 'giving', 'creditor', 'banks', '45.5', 'percent', 'of', 'the', 'company', 'in', 'return', 'for', 'wiping', 'out', 'one', 'billion', 'pounds', '(', '$', '1.56', 'billion', ')', 'of', 'its', 'debt', 'mountain', '.', 'the', 'long-awaited', 'restructuring', 'brings', 'to', 'an', 'end', 'months', 'of', 'wrangling', 'between', 'eurotunnel', 'and', 'the', '225', 'banks', 'to', 'which', 'it', 'owes', 'nearly', 'nine', 'billion', 'pounds', '.', 'the', 'deal', ',', 'announced', 'simultaneously', 'in', 'paris', 'and', 'london', ',', 'brings', 'the', 'company', 'back', 'from', 'the', 'brink', 'of', 'insolvency', 'but', 'leaves', 'shareholders', 'owning', 'only', '54.5', 'percent', 'of', 'the', 'company', '.', '``', 'the', 'restructuring', 'plan', 'provides', 'eurotunnel', 'with', 'the', 'medium', 'term', 'financial', 'stability', 'to', 'allow', 'it', 'to', 'consolidate', 'its', 'substantial', 'commercial', 'achievements', 'to', 'date', 'and', 'to', 'develop', 'its', 'operations', ',', '"', 'eurotunnel', 'co-chairman', 'alastair', 'morton', 'said', '.', 'the', 'firm', 'was', 'now', 'making', 'a', 'profit', 'before', 'interest', ',', 'he', 'added', '.', 'although', 'shareholders', 'will', 'see', 'their', 'interests', 'diluted', ',', 'they', 'were', 'offered', 'the', 'prospect', 'of', 'a', 'brighter', 'future', 'after', 'months', 'of', 'uncertainty', 'while', 'eurotunnel', 'wrestled', 'to', 'reduce', 'crippling', 'interest', 'payments', 'negotiated', 'during', 'the', 'tunnel', "'s", 'construction', '.', 'eurotunnel', ',', 'which', 'has', 'taken', 'around', 'half', 'the', 'cross-channel', 'market', 'from', 'the', 'european', 'ferry', 'companies', ',', 'said', 'a', 'strong', 'operating', 'performance', 'could', 'allow', 'it', 'to', 'pay', 'its', 'first', 'dividend', 'within', 'the', 'next', '10', 'years', '.', 'french', 'co-chairman', 'patrick', 'ponsolle', 'said', 'shareholders', 'would', 'have', 'to', 'be', 'patient', 'before', 'they', 'could', 'reap', 'the', 'benefits', 'of', 'the', 'company', "'s", 'success', '.', 'he', 'called', 'the', 'debt', 'restructuring', 'plan', '``', 'an', 'acceptable', 'compromise', '"', 'for', 'holders', 'of', 'eurotunnel', 'shares', '.', 'the', 'company', 'said', 'in', 'a', 'statement', 'there', 'was', 'still', 'considerable', 'work', 'to', 'be', 'done', 'to', 'finalise', 'and', 'agree', 'the', 'details', 'of', 'the', 'plan', 'before', 'it', 'can', 'be', 'submitted', 'to', 'shareholders', 'and', 'the', 'full', '225', 'bank', 'syndicate', 'for', 'approval', ',', 'probably', 'early', 'in', '1997.', 'monday', "'s", 'announcement', 'followed', 'two', 'weeks', 'of', 'highly', 'secretive', 'negotiations', 'between', 'eurotunnel', 'and', 'its', 'six', 'leading', 'banks', '.', 'this', 'was', 'extended', 'to', 'the', '24', '``', 'instructing', 'banks', '"', 'at', 'a', 'meeting', 'late', 'last', 'week', 'in', 'london',

'.', 'eurotunnel', 'said', 'the', 'debt-for-equity', 'swap', 'would', 'be', 'at', '130', 'pence', ',', 'or', '10.40', 'francs', ',', 'per', 'share', '.', 'that', 'is', 'considerably', 'below', 'the', 'level', 'of', 'around', '160', 'pence', 'widely', 'reported', 'in', 'the', 'run', 'up', 'to', 'the', 'deal', ',', 'and', 'will', 'reduce', 'outstanding', 'debt', 'of', '8.7', 'billion', 'pounds', 'by', '1.0', 'billion', '.', 'the', 'company', 'said', 'a', 'further', '3.7', 'billion', 'pounds', 'of', 'debt', 'would', 'be', 'converted', 'into', 'new', 'financial', 'instruments', 'and', 'existing', 'shareholders', 'would', 'be', 'able', 'to', 'participate', 'in', 'this', 'issue', '.', 'if', 'they', 'choose', 'not', 'to', 'take', 'up', 'free', 'warrants', 'entitling', 'them', 'to', 'subscribe', 'to', 'this', ',', 'eurotunnel', 'said', 'shareholders', '"', 'interests', 'may', 'be', 'reduced', 'further', 'to', 'just', 'over', '39', 'percent', 'of', 'the', 'company', 'by', 'the', 'end', 'of', 'december', '2003.', 'eurotunnel', "'s", 'shares', ',', 'which', 'were', 'suspended', 'last', 'week', 'at', '113.5', 'pence', 'ahead', 'of', 'monday', "'s", 'announcement', ',', 'should', 'resume', 'trading', 'on', 'tuesday', ',', 'the', 'company', 'said', '.', '(', '$', '1=.6393', 'pound', ')']

Stop words removed from 100618newsML.txt .

['anglo-french', 'channel', 'tunnel', 'operator', 'eurotunnel', 'monday', 'announced', 'deal', 'giving', 'creditor', 'banks', '45.5', 'percent', 'company', 'return', 'wiping', 'one', 'billion', 'pounds', '1.56', 'billion', 'debt', 'mountain', 'long-awaited', 'restructuring', 'brings', 'end', 'months', 'wrangling', 'eurotunnel', '225', 'banks', 'owes', 'nearly', 'nine', 'billion', 'pounds', 'deal', 'announced', 'simultaneously', 'paris', 'london', 'brings', 'company', 'back', 'brink', 'insolvency', 'leaves', 'shareholders', 'owning', '54.5', 'percent', 'company', '``', 'restructuring', 'plan', 'provides', 'eurotunnel', 'medium', 'term', 'financial', 'stability', 'allow', 'consolidate', 'substantial', 'commercial', 'achievements', 'date', 'develop', 'operations', '"', 'eurotunnel', 'co-chairman', 'alastair', 'morton', 'said', 'firm', 'making', 'profit', 'interest', 'added', 'although', 'shareholders', 'see', 'interests', 'diluted', 'offered', 'prospect', 'brighter', 'future', 'months', 'uncertainty', 'eurotunnel', 'wrestled', 'reduce', 'crippling', 'interest', 'payments', 'negotiated', 'tunnel', "'s", 'construction', 'eurotunnel', 'taken', 'around', 'half', 'cross-channel', 'market', 'european', 'ferry', 'companies', 'said', 'strong', 'operating', 'performance', 'could', 'allow', 'pay', 'first', 'dividend', 'within', 'next', '10', 'years', 'french', 'co-chairman', 'patrick', 'ponsolle', 'said', 'shareholders', 'would', 'patient', 'could', 'reap', 'benefits', 'company', "'s", 'success', 'called', 'debt', 'restructuring', 'plan', '``', 'acceptable', 'compromise', '"', 'holders', 'eurotunnel', 'shares', 'company', 'said', 'statement', 'still', 'considerable', 'work', 'done', 'finalise', 'agree', 'details', 'plan', 'submitted', 'shareholders', 'full', '225', 'bank', 'syndicate', 'approval', 'probably', 'early', '1997.', 'monday', "'s", 'announcement', 'followed', 'two', 'weeks', 'highly', 'secretive', 'negotiations', 'eurotunnel', 'six', 'leading', 'banks', 'extended', '24', '``', 'instructing', 'banks', '"', 'meeting', 'late', 'last', 'week', 'london', 'eurotunnel', 'said', 'debt-for-equity', 'swap', 'would', '130', 'pence', '10.40', 'francs', 'per', 'share', 'considerably', 'level', 'around', '160', 'pence', 'widely', 'reported', 'run', 'deal', 'reduce', 'outstanding', 'debt', '8.7', 'billion', 'pounds', '1.0', 'billion', 'company', 'said', '3.7', 'billion', 'pounds', 'debt', 'would', 'converted', 'new', 'financial', 'instruments', 'existing', 'shareholders', 'would', 'able', 'participate', 'issue', 'choose', 'take', 'free', 'warrants', 'entitling', 'subscribe', 'eurotunnel', 'said', 'shareholders', 'interests', 'may', 'reduced', '39', 'percent', 'company', 'end', 'december', '2003.', 'eurotunnel', "'s", 'shares', 'suspended', 'last', 'week', '113.5', 'pence', 'ahead', 'monday', "'s", 'announcement', 'resume', 'trading', 'tuesday', 'company', 'said', '1=.6393', 'pound']

Stemming 100618newsML.txt .

['anglo-french', 'channel', 'tunnel', 'oper', 'eurotunnel', 'monday', 'announc', 'deal', 'give', 'creditor', 'bank', '45.5', 'percent', 'compani', 'return', 'wipe', 'one', 'billion', 'pound', '1.56', 'billion', 'debt', 'mountain', 'long-await', 'restructur', 'bring', 'end', 'month', 'wrangl', 'eurotunnel', '225', 'bank', 'owe', 'nearli', 'nine', 'billion', 'pound', 'deal', 'announc', 'simultan', 'pari', 'london', 'bring', 'compani', 'back', 'brink', 'insolv', 'leav', 'sharehold', 'own', '54.5', 'percent', 'compani', '```', 'restructur', 'plan', 'provid', 'eurotunnel', 'medium', 'term', 'financi', 'stabil', 'allow', 'consolid', 'substanti', 'commerci', 'achiev', 'date', 'develop', 'oper', '"""', 'eurotunnel', 'co-chairman', 'alastair', 'morton', 'said', 'firm', 'make', 'profit', 'interest', 'ad', 'although', 'sharehold', 'see', 'interest', 'dilut', 'offer', 'prospect', 'brighter', 'futur', 'month', 'uncertainti', 'eurotunnel', 'wrestl', 'reduc', 'crippl', 'interest', 'payment', 'negoti', 'tunnel', "'s", 'construct', 'eurotunnel', 'taken', 'around', 'half', 'cross-channel', 'market', 'european', 'ferri', 'compani', 'said', 'strong', 'oper', 'perform', 'could', 'allow', 'pay', 'first', 'dividend', 'within', 'next', '10', 'year', 'french', 'co-chairman', 'patrick', 'ponsol', 'said', 'sharehold', 'would', 'patient', 'could', 'reap', 'benefit', 'compani', "'s", 'success', 'call', 'debt', 'restructur', 'plan', '```', 'accept', 'compromis', '"""', 'holder', 'eurotunnel', 'share', 'compani', 'said', 'statement', 'still', 'consider', 'work', 'done', 'finalis', 'agre', 'detail', 'plan', 'submit', 'sharehold', 'full', '225', 'bank', 'syndic', 'approv', 'probabl', 'earli', '1997.', 'monday', "'s", 'announc', 'follow', 'two', 'week', 'highli', 'secret', 'negoti', 'eurotunnel', 'six', 'lead', 'bank', 'extend', '24', '```', 'instruct', 'bank', '"""', 'meet', 'late', 'last', 'week', 'london', 'eurotunnel', 'said', 'debt-for-equ', 'swap', 'would', '130', 'penc', '10.40', 'franc', 'per', 'share', 'consider', 'level', 'around', '160', 'penc', 'wide', 'report', 'run', 'deal', 'reduc', 'outstand', 'debt', '8.7', 'billion', 'pound', '1.0', 'billion', 'compani', 'said', '3.7', 'billion', 'pound', 'debt', 'would', 'convert', 'new', 'financi', 'instrument', 'exist', 'sharehold', 'would', 'abl', 'particip', 'issu', 'choos', 'take', 'free', 'warrant', 'entitl', 'subscrib', 'eurotunnel', 'said', 'sharehold', 'interest', 'may', 'reduc', '39', 'percent', 'compani', 'end', 'decemb', '2003.', 'eurotunnel', "'s", 'share', 'suspend', 'last', 'week', '113.5', 'penc', 'ahead', 'monday', "'s", 'announc', 'resum', 'trade', 'tuesday', 'compani', 'said', '1=.6393', 'pound']

Step 3: Calculate tf-idf for each word in each document and generate document-word matrix (each element in the matrix is the tf-idf score for a word in a document)

Code used:

```
# tf-idf

vectorizer = TfidfVectorizer()
X = vectorizer.fit_transform(token_dict.values())
doc_matrix = X.toarray()
set_vocab = vectorizer.get_feature_names_out()
print(X.shape) # verify size

output_file.write(text_lines + "TF-IDF document-word matrix.\n" + text_lines)
output_file.write("%-15s%-20s%-20s%-20s\n" % ("Words", str(doc_names[0]), str(doc_names[1]), str(doc_names[2])))

for i in range(0,len(set_vocab)):
    output_file.write("%-15s%-20s%-20s%-20s\n" % (str(set_vocab[i]), str(doc_matrix[0][i]), str(doc_matrix[2][i]), str(doc_matrix[0][i
```

Screenshot of output of code that performs sentence tokenization, word tokenization, removal of stop words and stemming for 3 files:

nltk_output.txt

TF-IDF document-word matrix.

Words	100554newsML.txt	100593newsML.txt	100618newsML.txt
10	0.026060359604361007	0.03876927945294868	0.026060359604361007
113	0.013030179802180504	0.01938463972647434	0.013030179802180504
13	0.0	0.0	0.0
130	0.013030179802180504	0.01938463972647434	0.013030179802180504
14	0.0	0.0	0.0
150	0.022062012194251066	0.0	0.022062012194251066
160	0.013030179802180504	0.01938463972647434	0.013030179802180504
1997	0.013030179802180504	0.01938463972647434	0.013030179802180504
200	0.022062012194251066	0.0	0.022062012194251066
2003	0.013030179802180504	0.01938463972647434	0.013030179802180504
2004	0.022062012194251066	0.0	0.022062012194251066
225	0.013030179802180504	0.03876927945294868	0.013030179802180504
24	0.0	0.0249612530567176	0.0
39	0.013030179802180504	0.01938463972647434	0.013030179802180504
40	0.013030179802180504	0.01938463972647434	0.013030179802180504
400	0.04412402438850213	0.0	0.04412402438850213
45	0.013030179802180504	0.01938463972647434	0.013030179802180504
54	0.013030179802180504	0.01938463972647434	0.013030179802180504
56	0.0	0.0249612530567176	0.0
6393	0.016778728932091126	0.0249612530567176	0.016778728932091126
85	0.022062012194251066	0.0	0.022062012194251066
ability	0.022062012194251066	0.0	0.022062012194251066
able	0.026060359604361007	0.01938463972647434	0.026060359604361007
acceptable	0.0	0.0249612530567176	0.0
achievements	0.0	0.0249612530567176	0.0
added	0.013030179802180504	0.01938463972647434	0.013030179802180504
adding	0.022062012194251066	0.0	0.022062012194251066
admitted	0.022062012194251066	0.0	0.022062012194251066
afford	0.022062012194251066	0.0	0.022062012194251066
after	0.026060359604361007	0.01938463972647434	0.026060359604361007
again	0.022062012194251066	0.0	0.022062012194251066
agree	0.026060359604361007	0.01938463972647434	0.026060359604361007
ahead	0.013030179802180504	0.01938463972647434	0.013030179802180504
alastair	0.013030179802180504	0.01938463972647434	0.013030179802180504
all	0.022062012194251066	0.0	0.022062012194251066
allow	0.013030179802180504	0.03876927945294868	0.013030179802180504
already	0.022062012194251066	0.0	0.022062012194251066
although	0.013030179802180504	0.01938463972647434	0.013030179802180504
an	0.026060359604361007	0.03876927945294868	0.026060359604361007
analyst	0.022062012194251066	0.0	0.022062012194251066
analysts	0.022062012194251066	0.0	0.022062012194251066
and	0.19545269703270754	0.15507711781179473	0.19545269703270754
anglo	0.013030179802180504	0.01938463972647434	0.013030179802180504
angry	0.022062012194251066	0.0	0.022062012194251066
announced	0.026060359604361007	0.03876927945294868	0.026060359604361007
announcement	0.013030179802180504	0.03876927945294868	0.013030179802180504
annual	0.022062012194251066	0.0	0.022062012194251066
approval	0.026060359604361007	0.01938463972647434	0.026060359604361007
are	0.04412402438850213	0.0	0.04412402438850213
around	0.013030179802180504	0.03876927945294868	0.013030179802180504
arranged	0.022062012194251066	0.0	0.022062012194251066
as	0.04412402438850213	0.0	0.04412402438850213
asked	0.022062012194251066	0.0	0.022062012194251066
at	0.09121125861526352	0.05815391917942302	0.09121125861526352
available	0.022062012194251066	0.0	0.022062012194251066
avoiding	0.022062012194251066	0.0	0.022062012194251066
awaited	0.013030179802180504	0.01938463972647434	0.013030179802180504
back	0.013030179802180504	0.01938463972647434	0.013030179802180504

Output of code for calculating TF-IDF:

TF-IDF document-word matrix.

Words	100554newsML.txt	100593newsML.txt	100618newsML.txt
10	0.026060359604361007	0.03876927945294868	0.026060359604361007
113	0.013030179802180504	0.01938463972647434	0.013030179802180504
13	0.0	0.0	0.0
130	0.013030179802180504	0.01938463972647434	0.013030179802180504
14	0.0	0.0	0.0
150	0.022062012194251066	0.0	0.022062012194251066
160	0.013030179802180504	0.01938463972647434	0.013030179802180504
1997	0.013030179802180504	0.01938463972647434	0.013030179802180504
200	0.022062012194251066	0.0	0.022062012194251066
2003	0.013030179802180504	0.01938463972647434	0.013030179802180504
2004	0.022062012194251066	0.0	0.022062012194251066
225	0.013030179802180504	0.03876927945294868	0.013030179802180504
24	0.0	0.024961253056717 6	0.0
39	0.013030179802180504	0.01938463972647434	0.013030179802180504
40	0.013030179802180504	0.01938463972647434	0.013030179802180504
400	0.04412402438850213	0.0	0.04412402438850213
45	0.013030179802180504	0.01938463972647434	0.013030179802180504
54	0.013030179802180504	0.01938463972647434	0.013030179802180504
56	0.0	0.024961253056717 6	0.0
6393	0.016778728932091126	0.024961253056717 6	0.016778728932091126
85	0.022062012194251066	0.0	0.022062012194251066
ability	0.022062012194251066	0.0	0.022062012194251066
able	0.026060359604361007	0.01938463972647434	0.026060359604361007
acceptable	0.0	0.024961253056717 6	0.0
achievements	0.0	0.024961253056717 6	0.0
added	0.013030179802180504	0.01938463972647434	0.013030179802180504
adding	0.022062012194251066	0.0	0.022062012194251066
admitted	0.022062012194251066	0.0	0.022062012194251066
afford	0.022062012194251066	0.0	0.022062012194251066
after	0.026060359604361007	0.01938463972647434	0.026060359604361007
again	0.022062012194251066	0.0	0.022062012194251066
agree	0.026060359604361007	0.01938463972647434	0.026060359604361007
ahead	0.013030179802180504	0.01938463972647434	0.013030179802180504
alastair	0.013030179802180504	0.01938463972647434	0.013030179802180504
all	0.022062012194251066	0.0	0.022062012194251066
allow	0.013030179802180504	0.03876927945294868	0.013030179802180504
already	0.022062012194251066	0.0	0.022062012194251066
although	0.013030179802180504	0.01938463972647434	0.013030179802180504
an	0.026060359604361007	0.03876927945294868	0.026060359604361007
analyst	0.022062012194251066	0.0	0.022062012194251066
analysts	0.022062012194251066	0.0	0.022062012194251066
and	0.19545269703270754	0.15507711781179473	0.19545269703270754
anglo	0.013030179802180504	0.01938463972647434	0.013030179802180504
angry	0.022062012194251066	0.0	0.022062012194251066
announced	0.026060359604361007	0.03876927945294868	0.026060359604361007
announcement	0.013030179802180504	0.03876927945294868	0.013030179802180504
annual	0.022062012194251066	0.0	0.022062012194251066
approval	0.026060359604361007	0.01938463972647434	0.026060359604361007
are	0.04412402438850213	0.0	0.04412402438850213
around	0.013030179802180504	0.03876927945294868	0.013030179802180504
arranged	0.022062012194251066	0.0	0.022062012194251066
as	0.04412402438850213	0.0	0.04412402438850213
asked	0.022062012194251066	0.0	0.022062012194251066
at	0.09121125861526352	0.05815391917942302	0.09121125861526352
available	0.022062012194251066	0.0	0.022062012194251066
avoiding	0.022062012194251066	0.0	0.022062012194251066
awaited	0.013030179802180504	0.01938463972647434	0.013030179802180504
back	0.013030179802180504	0.01938463972647434	0.013030179802180504
bank	0.03909053940654151	0.01938463972647434	0.03909053940654151
bankruptcy	0.022062012194251066	0.0	0.022062012194251066
banks	0.052120719208722015	0.07753855890589736	0.052120719208722015

be 0.11727161821962452 0.1356924780853204 0.11727161821962452
become 0.022062012194251066 0.022062012194251066 0.022062012194251066
before 0.026060359604361007 0.05815391917942302 0.026060359604361007
below 0.013030179802180504 0.01938463972647434 0.013030179802180504
benefits 0.0 0.0249612530567176 0.0
between 0.013030179802180504 0.03876927945294868 0.013030179802180504
bill 0.04412402438850213 0.0 0.04412402438850213
billion 0.06515089901090251 0.11630783835884605 0.06515089901090251
bonds 0.022062012194251066 0.022062012194251066 0.022062012194251066
brighter 0.013030179802180504 0.01938463972647434 0.013030179802180504
brings 0.013030179802180504 0.03876927945294868 0.013030179802180504
brink 0.013030179802180504 0.01938463972647434 0.013030179802180504
busiest 0.022062012194251066 0.022062012194251066 0.022062012194251066
but 0.03909053940654151 0.01938463972647434 0.03909053940654151
by 0.013030179802180504 0.03876927945294868 0.013030179802180504
called 0.0 0.0249612530567176 0.0
can 0.03909053940654151 0.01938463972647434 0.03909053940654151
capped 0.022062012194251066 0.022062012194251066 0.022062012194251066
cash 0.04412402438850213 0.0 0.04412402438850213
chairman 0.026060359604361007 0.03876927945294868 0.026060359604361007
channel 0.026060359604361007 0.03876927945294868 0.026060359604361007
choose 0.013030179802180504 0.01938463972647434 0.013030179802180504
co 0.026060359604361007 0.03876927945294868 0.026060359604361007
collapse 0.022062012194251066 0.022062012194251066 0.022062012194251066
come 0.04412402438850213 0.0 0.04412402438850213
commercial 0.013030179802180504 0.01938463972647434 0.013030179802180504
companies 0.013030179802180504 0.01938463972647434 0.013030179802180504
company 0.14333197782398555 0.15507711781179473 0.14333197782398555
complex 0.022062012194251066 0.022062012194251066 0.022062012194251066
compromise 0.0 0.0249612530567176 0.0
conference 0.04412402438850213 0.0 0.04412402438850213
considerable 0.013030179802180504 0.01938463972647434 0.013030179802180504
considerably 0.013030179802180504 0.01938463972647434 0.013030179802180504
consolidate 0.013030179802180504 0.01938463972647434 0.013030179802180504
constitute 0.022062012194251066 0.022062012194251066 0.022062012194251066
construction 0.013030179802180504 0.01938463972647434 0.013030179802180504
converted 0.013030179802180504 0.01938463972647434 0.013030179802180504
costs 0.022062012194251066 0.022062012194251066 0.022062012194251066
could 0.03909053940654151 0.03876927945294868 0.03909053940654151
creditor 0.013030179802180504 0.01938463972647434 0.013030179802180504
creditors 0.022062012194251066 0.022062012194251066 0.022062012194251066
crippling 0.013030179802180504 0.01938463972647434 0.013030179802180504
cross 0.013030179802180504 0.01938463972647434 0.013030179802180504
current 0.04412402438850213 0.0 0.04412402438850213
date 0.0 0.0249612530567176 0.0
deal 0.09121125861526352 0.05815391917942302 0.09121125861526352
debt 0.07818107881308302 0.0969231986323717 0.07818107881308302
debts 0.022062012194251066 0.022062012194251066 0.022062012194251066
december 0.013030179802180504 0.01938463972647434 0.013030179802180504
depreciation 0.022062012194251066 0.022062012194251066 0.022062012194251066
despite 0.022062012194251066 0.022062012194251066 0.022062012194251066
details 0.026060359604361007 0.01938463972647434 0.026060359604361007
develop 0.013030179802180504 0.01938463972647434 0.013030179802180504
difficult 0.022062012194251066 0.022062012194251066 0.022062012194251066
diluted 0.013030179802180504 0.01938463972647434 0.013030179802180504
dividend 0.026060359604361007 0.01938463972647434 0.026060359604361007
do 0.022062012194251066 0.022062012194251066 0.022062012194251066
done 0.013030179802180504 0.01938463972647434 0.013030179802180504
doomsday 0.022062012194251066 0.022062012194251066 0.022062012194251066
during 0.013030179802180504 0.01938463972647434 0.013030179802180504
dwindle 0.022062012194251066 0.022062012194251066 0.022062012194251066
early 0.03909053940654151 0.01938463972647434 0.03909053940654151
eight 0.022062012194251066 0.022062012194251066 0.022062012194251066
end 0.013030179802180504 0.03876927945294868 0.013030179802180504
entitling 0.013030179802180504 0.01938463972647434 0.013030179802180504
equity 0.026060359604361007 0.01938463972647434 0.026060359604361007
european 0.013030179802180504 0.01938463972647434 0.013030179802180504

eurotunnel 0.13030179802180503 0.21323103699121776 0.13030179802180503
existing 0.013030179802180504 0.01938463972647434 0.013030179802180504
extended 0.0 0.0249612530567176 0.0
ferry 0.013030179802180504 0.01938463972647434 0.013030179802180504
few 0.0220620121942510660.0 0.022062012194251066
finalise 0.013030179802180504 0.01938463972647434 0.013030179802180504
finance 0.0220620121942510660.0 0.022062012194251066
financial 0.026060359604361007 0.03876927945294868 0.026060359604361007
find 0.0220620121942510660.0 0.022062012194251066
firm 0.013030179802180504 0.01938463972647434 0.013030179802180504
first 0.013030179802180504 0.01938463972647434 0.013030179802180504
fixed 0.0220620121942510660.0 0.022062012194251066
flow 0.0220620121942510660.0 0.022062012194251066
followed 0.0 0.0249612530567176 0.0
for 0.11727161821962452 0.07753855890589736 0.11727161821962452
formula 0.0220620121942510660.0 0.022062012194251066
forward 0.0220620121942510660.0 0.022062012194251066
france 0.0220620121942510660.0 0.022062012194251066
francs 0.013030179802180504 0.01938463972647434 0.013030179802180504
free 0.013030179802180504 0.01938463972647434 0.013030179802180504
french 0.026060359604361007 0.03876927945294868 0.026060359604361007
from 0.026060359604361007 0.03876927945294868 0.026060359604361007
full 0.0 0.0249612530567176 0.0
further 0.026060359604361007 0.03876927945294868 0.026060359604361007
future 0.026060359604361007 0.01938463972647434 0.026060359604361007
game 0.0220620121942510660.0 0.022062012194251066
get 0.0220620121942510660.0 0.022062012194251066
giving 0.013030179802180504 0.01938463972647434 0.013030179802180504
going 0.0220620121942510660.0 0.022062012194251066
granted 0.0220620121942510660.0 0.022062012194251066
group 0.0220620121942510660.0 0.022062012194251066
half 0.013030179802180504 0.01938463972647434 0.013030179802180504
happen 0.0220620121942510660.0 0.022062012194251066
has 0.013030179802180504 0.01938463972647434 0.013030179802180504
have 0.052120719208722015 0.01938463972647434 0.052120719208722015
having 0.0220620121942510660.0 0.022062012194251066
he 0.013030179802180504 0.03876927945294868 0.013030179802180504
help 0.0220620121942510660.0 0.022062012194251066
here 0.0220620121942510660.0 0.022062012194251066
highly 0.013030179802180504 0.01938463972647434 0.013030179802180504
holders 0.0 0.0249612530567176 0.0
holdings 0.0220620121942510660.0 0.022062012194251066
hopeful 0.0220620121942510660.0 0.022062012194251066
if 0.06515089901090251 0.01938463972647434 0.06515089901090251
in 0.11727161821962452 0.1356924780853204 0.11727161821962452
ingenious 0.0220620121942510660.0 0.022062012194251066
insolvency 0.0 0.0249612530567176 0.0
instructing 0.0 0.0249612530567176 0.0
instruments 0.013030179802180504 0.01938463972647434 0.013030179802180504
interest 0.07818107881308302 0.03876927945294868 0.07818107881308302
interests 0.013030179802180504 0.03876927945294868 0.013030179802180504
into 0.026060359604361007 0.01938463972647434 0.026060359604361007
investment 0.04412402438850213 0.0 0.04412402438850213
is 0.052120719208722015 0.01938463972647434 0.052120719208722015
issue 0.03909053940654151 0.01938463972647434 0.03909053940654151
it 0.052120719208722015 0.07753855890589736 0.052120719208722015
its 0.07818107881308302 0.0969231986323717 0.07818107881308302
just 0.013030179802180504 0.01938463972647434 0.013030179802180504
keep 0.0220620121942510660.0 0.022062012194251066
last 0.013030179802180504 0.03876927945294868 0.013030179802180504
late 0.0 0.0249612530567176 0.0
leading 0.0 0.0249612530567176 0.0
least 0.0220620121942510660.0 0.022062012194251066
leaves 0.013030179802180504 0.01938463972647434 0.013030179802180504
less 0.0220620121942510660.0 0.022062012194251066
level 0.013030179802180504 0.01938463972647434 0.013030179802180504
lifeline 0.0220620121942510660.0 0.022062012194251066

likely 0.0220620121942510660.0 0.022062012194251066
london 0.013030179802180504.0387692794529486 8 0.013030179802180504
long 0.013030179802180504.0193846397264743 4 0.013030179802180504
look 0.0220620121942510660.0 0.022062012194251066
major 0.0220620121942510660.0 0.022062012194251066
making 0.013030179802180504.0193846397264743 4 0.013030179802180504
market 0.013030179802180504.0193846397264743 4 0.013030179802180504
massive 0.0220620121942510660.0 0.022062012194251066
maximum 0.0220620121942510660.0 0.022062012194251066
may 0.013030179802180504.0193846397264743 4 0.013030179802180504
medium 0.013030179802180504.0193846397264743 4 0.013030179802180504
meet 0.0220620121942510660.0 0.022062012194251066
meeting 0.0 0.0249612530567176 0.0
million 0.0661860365827532 0.0 0.0661860365827532
monday 0.026060359604361007.058153919179423 02 0.026060359604361007
months 0.013030179802180504.0387692794529486 8 0.013030179802180504
morton 0.052120719208722015.0193846397264743 4 0.052120719208722015
mountain 0.0 0.03282104809905034 0.0
name 0.0220620121942510660.0 0.022062012194251066
nearly 0.026060359604361007.0193846397264743 4 0.026060359604361007
negotiated 0.013030179802180504.0193846397264743 4 0.013030179802180504
negotiations 0.0 0.0249612530567176 0.0
new 0.013030179802180504.0193846397264743 4 0.013030179802180504
news 0.04412402438850213 0.0 0.04412402438850213
next 0.013030179802180504.0193846397264743 4 0.013030179802180504
nine 0.026060359604361007.0193846397264743 4 0.026060359604361007
nobody 0.04412402438850213 0.0 0.04412402438850213
not 0.052120719208722015.0193846397264743 4 0.052120719208722015
notes 0.0661860365827532 0.0 0.0661860365827532
now 0.013030179802180504.0193846397264743 4 0.013030179802180504
of 0.29969413545015156 0.32953887535006376 0.29969413545015156
offered 0.013030179802180504.0193846397264743 4 0.013030179802180504
on 0.03909053940654151 0.03876927945294868 0.03909053940654151
one 0.013030179802180504.0193846397264743 4 0.013030179802180504
only 0.026060359604361007.0193846397264743 4 0.026060359604361007
operating 0.013030179802180504.0193846397264743 4 0.013030179802180504
operations 0.013030179802180504.0193846397264743 4 0.013030179802180504
operator 0.013030179802180504.0193846397264743 4 0.013030179802180504
or 0.013030179802180504.0193846397264743 4 0.013030179802180504
out 0.013030179802180504.0193846397264743 4 0.013030179802180504
outstanding 0.0 0.0249612530567176 0.0
over 0.026060359604361007.0193846397264743 4 0.026060359604361007
owes 0.0 0.0249612530567176 0.0
owning 0.013030179802180504.0193846397264743 4 0.013030179802180504
paris 0.026060359604361007.0193846397264743 4 0.026060359604361007
participate 0.013030179802180504.0193846397264743 4 0.013030179802180504
patient 0.013030179802180504.0193846397264743 4 0.013030179802180504
patrick 0.013030179802180504.0193846397264743 4 0.013030179802180504
pay 0.03909053940654151 0.0193846397264743 4 0.03909053940654151
payments 0.026060359604361007.0193846397264743 4 0.026060359604361007
pence 0.03909053940654151 0.05815391917942302 0.03909053940654151
per 0.013030179802180504.0193846397264743 4 0.013030179802180504
percent 0.03909053940654151 0.05815391917942302 0.03909053940654151
performance 0.013030179802180504.0193846397264743 4 0.013030179802180504
performed 0.0220620121942510660.0 0.022062012194251066
plan 0.026060359604361007.058153919179423 02 0.026060359604361007
ponsolle 0.013030179802180504.0193846397264743 4 0.013030179802180504
position 0.0220620121942510660.0 0.022062012194251066
pound 0.016778728932091126.0249612530567176 0.016778728932091126
pounds 0.09121125861526352 0.07753855890589736 0.09121125861526352
pretty 0.0220620121942510660.0 0.022062012194251066
probably 0.013030179802180504.0193846397264743 4 0.013030179802180504
profit 0.013030179802180504.0193846397264743 4 0.013030179802180504
prospect 0.013030179802180504.0193846397264743 4 0.013030179802180504
provides 0.013030179802180504.0193846397264743 4 0.013030179802180504
reap 0.0 0.0249612530567176 0.0
reduce 0.013030179802180504.0387692794529486 8 0.013030179802180504

reduced 0.013030179802180504 0.01938463972647434 0.013030179802180504
reject 0.022062012194251066 0.0220620121942510660.0 0.022062012194251066
reported 0.013030179802180504 0.01938463972647434 0.013030179802180504
reporters 0.04412402438850213 0.0 0.04412402438850213
restructuring 0.013030179802180504 0.05815391917942302 0.013030179802180504
resume 0.013030179802180504 0.01938463972647434 0.013030179802180504
return 0.013030179802180504 0.01938463972647434 0.013030179802180504
revenue 0.022062012194251066 0.0220620121942510660.0 0.022062012194251066
roll 0.022062012194251066 0.0220620121942510660.0 0.022062012194251066
route 0.022062012194251066 0.0220620121942510660.0 0.022062012194251066
run 0.016778728932091126 0.02496125305671760.016778728932091126
said 0.11727161821962452 0.15507711781179473 0.11727161821962452
scenario 0.022062012194251066 0.0220620121942510660.0 0.022062012194251066
secretive 0.0 0.02496125305671760.0
secure 0.022062012194251066 0.0220620121942510660.0 0.022062012194251066
see 0.013030179802180504 0.01938463972647434 0.013030179802180504
seen 0.022062012194251066 0.0220620121942510660.0 0.022062012194251066
seven 0.022062012194251066 0.0220620121942510660.0 0.022062012194251066
share 0.013030179802180504 0.01938463972647434 0.013030179802180504
shareholders 0.09121125861526352 0.11630783835884605 0.09121125861526352
shares 0.013030179802180504 0.03876927945294868 0.013030179802180504
should 0.0 0.02496125305671760.0
simultaneously 0.013030179802180504 0.01938463972647434 0.013030179802180504
six 0.0 0.02496125305671760.0
so 0.022062012194251066 0.0220620121942510660.0 0.022062012194251066
spring 0.022062012194251066 0.0220620121942510660.0 0.022062012194251066
stabilisation 0.04412402438850213 0.0 0.04412402438850213
stability 0.013030179802180504 0.01938463972647434 0.013030179802180504
statement 0.016778728932091126 0.02496125305671760.016778728932091126
still 0.03909053940654151 0.01938463972647434 0.03909053940654151
strong 0.013030179802180504 0.01938463972647434 0.013030179802180504
submitted 0.013030179802180504 0.01938463972647434 0.013030179802180504
subscribe 0.013030179802180504 0.01938463972647434 0.013030179802180504
substantial 0.0 0.02496125305671760.0
success 0.0 0.02496125305671760.0
suspended 0.013030179802180504 0.01938463972647434 0.013030179802180504
sustainable 0.022062012194251066 0.0220620121942510660.0 0.022062012194251066
swap 0.013030179802180504 0.01938463972647434 0.013030179802180504
swaps 0.022062012194251066 0.0220620121942510660.0 0.022062012194251066
syndicate 0.0 0.02496125305671760.0
take 0.013030179802180504 0.01938463972647434 0.013030179802180504
taken 0.013030179802180504 0.01938463972647434 0.013030179802180504
taking 0.022062012194251066 0.0220620121942510660.0 0.022062012194251066
ten 0.022062012194251066 0.0220620121942510660.0 0.022062012194251066
term 0.013030179802180504 0.01938463972647434 0.013030179802180504
than 0.022062012194251066 0.0220620121942510660.0 0.022062012194251066
that 0.06515089901090251 0.01938463972647434 0.06515089901090251
the 0.6515089901090252 0.6009238315207046 0.6515089901090252
their 0.026060359604361007 0.01938463972647434 0.026060359604361007
them 0.013030179802180504 0.01938463972647434 0.013030179802180504
then 0.022062012194251066 0.0220620121942510660.0 0.022062012194251066
there 0.013030179802180504 0.01938463972647434 0.013030179802180504
these 0.022062012194251066 0.0220620121942510660.0 0.022062012194251066
they 0.07818107881308302 0.05815391917942302 0.07818107881308302
this 0.03909053940654151 0.05815391917942302 0.03909053940654151
throws 0.022062012194251066 0.0220620121942510660.0 0.022062012194251066
time 0.022062012194251066 0.0220620121942510660.0 0.022062012194251066
to 0.29969413545015156 0.36830815480301250.29969413545015156
told 0.04412402438850213 0.0 0.04412402438850213
trading 0.013030179802180504 0.01938463972647434 0.013030179802180504
tuesday 0.013030179802180504 0.01938463972647434 0.013030179802180504
tunnel 0.026060359604361007 0.03876927945294868 0.026060359604361007
two 0.0 0.02496125305671760.0
uncertainty 0.013030179802180504 0.01938463972647434 0.013030179802180504
under 0.022062012194251066 0.0220620121942510660.0 0.022062012194251066
unpaid 0.022062012194251066 0.0220620121942510660.0 0.022062012194251066
up 0.03909053940654151 0.03876927945294868 0.03909053940654151

```
urged       0.0220620121942510660.0             0.022062012194251066
very       0.0220620121942510660.0             0.022062012194251066
wants       0.04412402438850213 0.0             0.04412402438850213
warrants      0.013030179802180504.0193846397264743 4 0.013030179802180504
was        0.026060359604361007.05815391917942302 0.026060359604361007
we         0.0220620121942510660.0             0.022062012194251066
week        0.013030179802180504.0387692794529486 8 0.013030179802180504
weeks       0.0             0.0249612530567176  0.0
well       0.0220620121942510660.0             0.022062012194251066
were        0.026060359604361007.0387692794529486 8 0.026060359604361007
what        0.08824804877700426 0.0             0.08824804877700426
when        0.0220620121942510660.0             0.022062012194251066
which       0.052120719208722015.0581539191794230 2 0.052120719208722015
while      0.013030179802180504.0193846397264743 4 0.013030179802180504
who         0.0220620121942510660.0             0.022062012194251066
widely      0.013030179802180504.0193846397264743 4 0.013030179802180504
will       0.06515089901090251 0.0387692794529486 8 0.06515089901090251
wiping      0.013030179802180504.0193846397264743 4 0.013030179802180504
with       0.026060359604361007.0193846397264743 4 0.026060359604361007
within      0.013030179802180504.0193846397264743 4 0.013030179802180504
work        0.013030179802180504.0193846397264743 4 0.013030179802180504
would       0.09121125861526352 0.07753855890589736 0.09121125861526352
wrangling     0.0             0.0249612530567176  0.0
wrestled      0.013030179802180504.0193846397264743 4 0.013030179802180504
year       0.0220620121942510660.0             0.022062012194251066
years       0.06515089901090251 0.0193846397264743 4 0.06515089901090251
you        0.0220620121942510660.0             0.022062012194251066
```

Step 4: Calculate pairwise cosine similarity for the document

Code used:

```
# cosine similarity

output_file.write(text_lines + "Cosine similarity.\n" + text_lines)
for i in range(0, num_docs):
    for j in range(i, num_docs):
        if i != j:
            output_file.write("cosine similarity of %s to %s is %s. \n" % ( doc_names[i], doc_names[j], cosine_similarity(X[i,], X[j,])))
output_file.close()
```

Output of code for calculating cosine similarity:

```
Cosine similarity.

cosine similarity of 100554newsML.txt to 100593newsML.txt is [[0.89622189]].
cosine similarity of 100554newsML.txt to 100618newsML.txt is [[0.90820406]].
cosine similarity of 100593newsML.txt to 100618newsML.txt is [[0.99175097]].
```

This fairly high cosine similarity makes sense because all documents share many words. Additionally, we would expect the 100554newsML, which has 758 words, to have a lower cosine similarity compared to the other two with their lower word counts. The best match is for 100593newsML.txt to 100618newsML.txt at [[0.99175097]]. This

makes sense because they do share many words and their word counts (483 and 492) are similar.

```
/Users/rraven/PycharmProjects/pythonProject1/venv/bin/python /Users/rraven/
file 100554newsML.txt
number of words 758
file 100593newsML.txt
number of words 492
file 100618newsML.txt
number of words 483
```