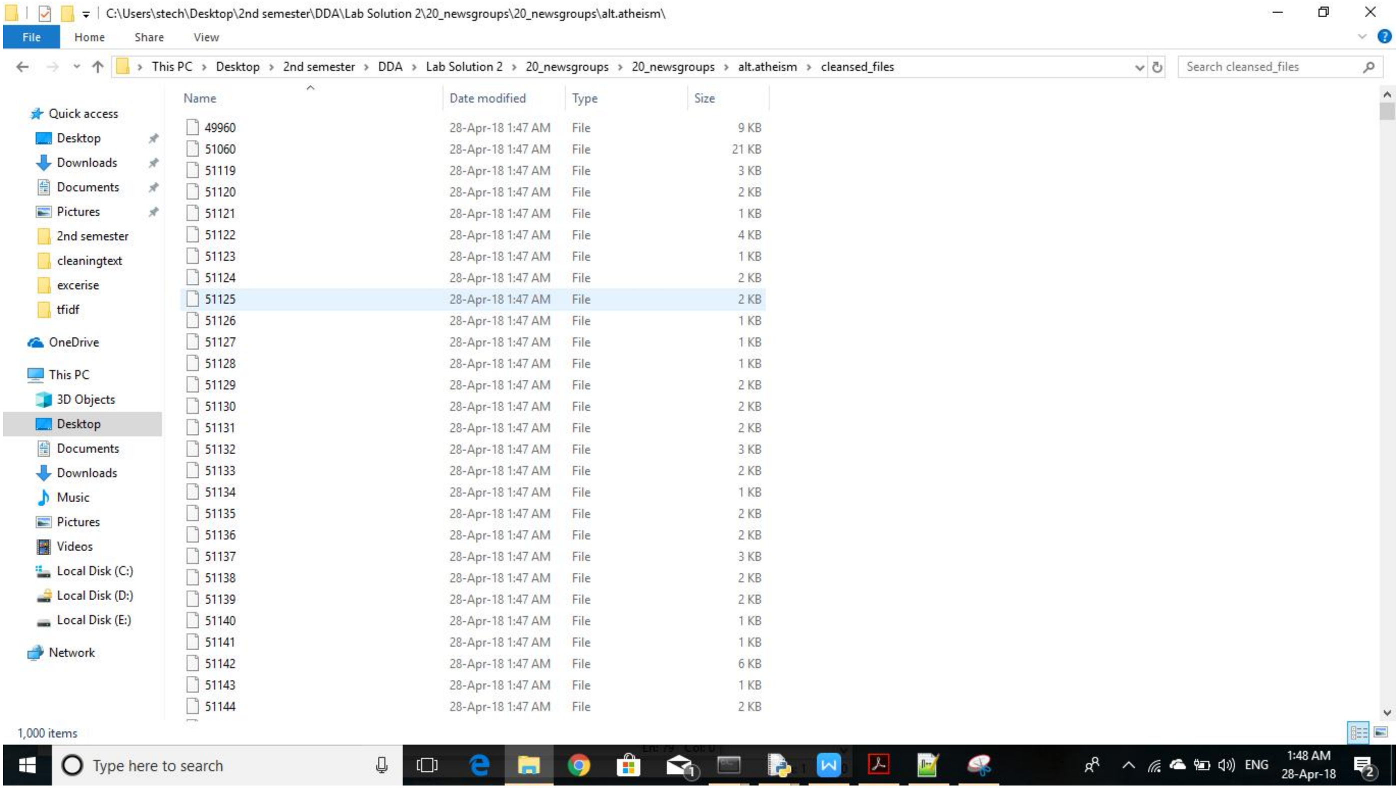
Distributed Data Analytics
Exercise Sheet # 2
Arooba Jamil Khokhar
278077

Exercise 1: Data Cleaning and Text Tokenization

In this exercise, I have first taken the data from the corpus. I have implemented the logic on a folder-by-folder basis such that the root worker goes into each folder and gets all the file paths for the file present in it. It then distributes the filesPath array containing paths for each file by splitting the array into equal chunks.

For example: If a folder has 1000 files, and we have 4 workers, then root sends array of size 250 to each worker to process, meaning that each worker will have to process 250 files. This is the basic parallelization strategy.

Each worker loops through the file array taking each file, cleaning it by removing the stop words and then returns the cleansed string. The returned string is tokenized by a comma and then written onto a file with the same name in folder named 'cleansed_files'



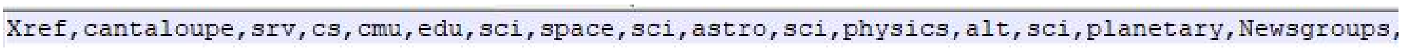
C:\Users\stech\Desktop\2nd semester\DDA\Lab Solution 2\20_newsgroups\20_newsgroups\alt.atheism\

This PC › Desktop › 2nd semester › DDA › Lab Solution 2 › 20_newsgroups › 20_newsgroups › alt.atheism › cleansed_files

Name	Date modified	Type	Size
49960	28-Apr-18 1:47 AM	File	9 KB
51060	28-Apr-18 1:47 AM	File	21 KB
51119	28-Apr-18 1:47 AM	File	3 KB
51120	28-Apr-18 1:47 AM	File	2 KB
51121	28-Apr-18 1:47 AM	File	1 KB
51122	28-Apr-18 1:47 AM	File	4 KB
51123	28-Apr-18 1:47 AM	File	1 KB
51124	28-Apr-18 1:47 AM	File	2 KB
51125	28-Apr-18 1:47 AM	File	2 KB
51126	28-Apr-18 1:47 AM	File	1 KB
51127	28-Apr-18 1:47 AM	File	1 KB
51128	28-Apr-18 1:47 AM	File	1 KB
51129	28-Apr-18 1:47 AM	File	2 KB
51130	28-Apr-18 1:47 AM	File	2 KB
51131	28-Apr-18 1:47 AM	File	2 KB
51132	28-Apr-18 1:47 AM	File	3 KB
51133	28-Apr-18 1:47 AM	File	2 KB
51134	28-Apr-18 1:47 AM	File	1 KB
51135	28-Apr-18 1:47 AM	File	2 KB
51136	28-Apr-18 1:47 AM	File	2 KB
51137	28-Apr-18 1:47 AM	File	3 KB
51138	28-Apr-18 1:47 AM	File	2 KB
51139	28-Apr-18 1:47 AM	File	2 KB
51140	28-Apr-18 1:47 AM	File	1 KB
51141	28-Apr-18 1:47 AM	File	1 KB
51142	28-Apr-18 1:47 AM	File	6 KB
51143	28-Apr-18 1:47 AM	File	1 KB
51144	28-Apr-18 1:47 AM	File	2 KB

1,000 items

Fig 1: Folder containing cleaned and tokenized files



```
Xref,cantaloupe,srv,cs,cmu,edu,sci,space,sci,astro,sci,physics,alt,sci,planetary,Newsgroups,
```

Fig 2: Tokenized values in the file

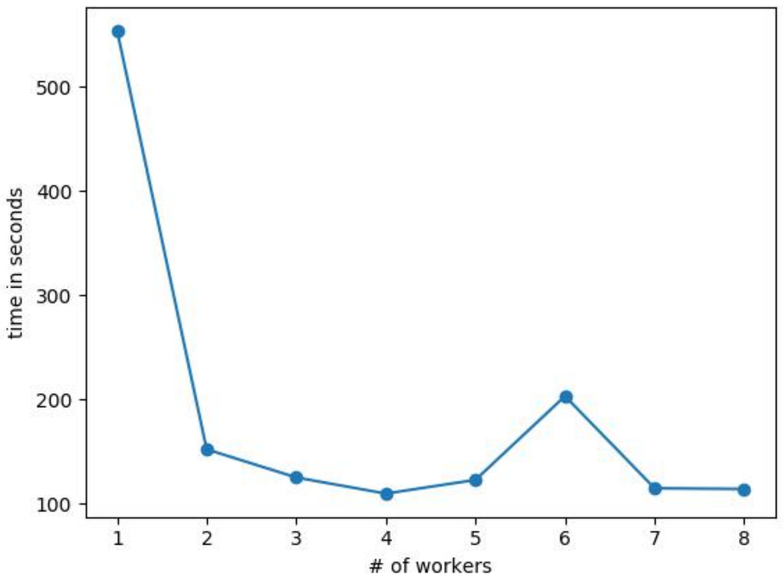


Fig : Time taken to cleanse and tokenize with different number of workers

Exercise 2,3,4: TF-IDF Score

In this question, the same logic to divide the file array among workers is used. The only new addition is the masterDictionary, which holds all the text counts for each file in a folder. The flow of the program is as follows:

1) Each worker gets the file array. It then creates a partial dictionary which holds the word count for each word. The key of the dictionary is the filename. And each filename key holds the word count for that file.

2) After the worker is done creating the partial dictionary, It then sends the dictionary back to the root, which merges all the dictionary to create a master dictionary. This will be used to perform word lookups when calculating IDF.

3) The root again sends the master dictionary back to all the workers.

4) Each worker has access to its partial dictionary as well as the master dictionary. Each worker then calculates the TF and IDF scores on its respective dictionaries, since each worker made dictionary for a chunk of files.

5) The scores for each word for each file of a specific folder is written in the same folder in a directory named 'TFIDF_ScoresPerFile'.

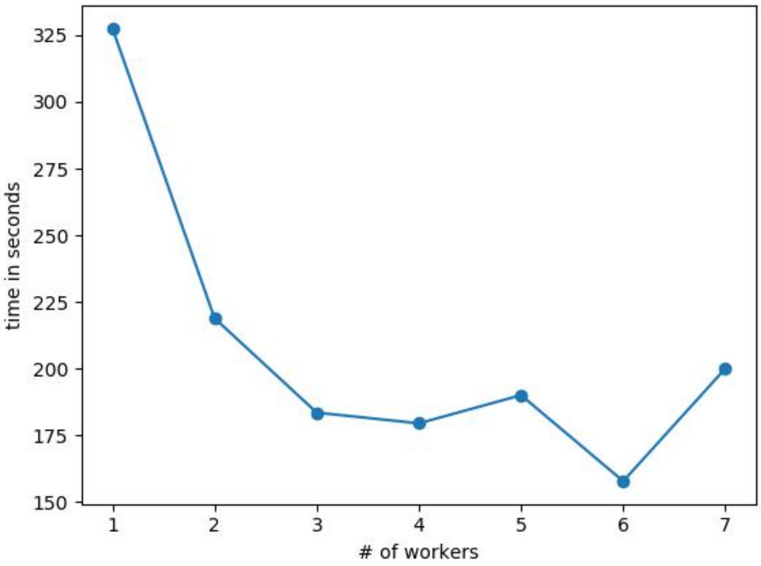
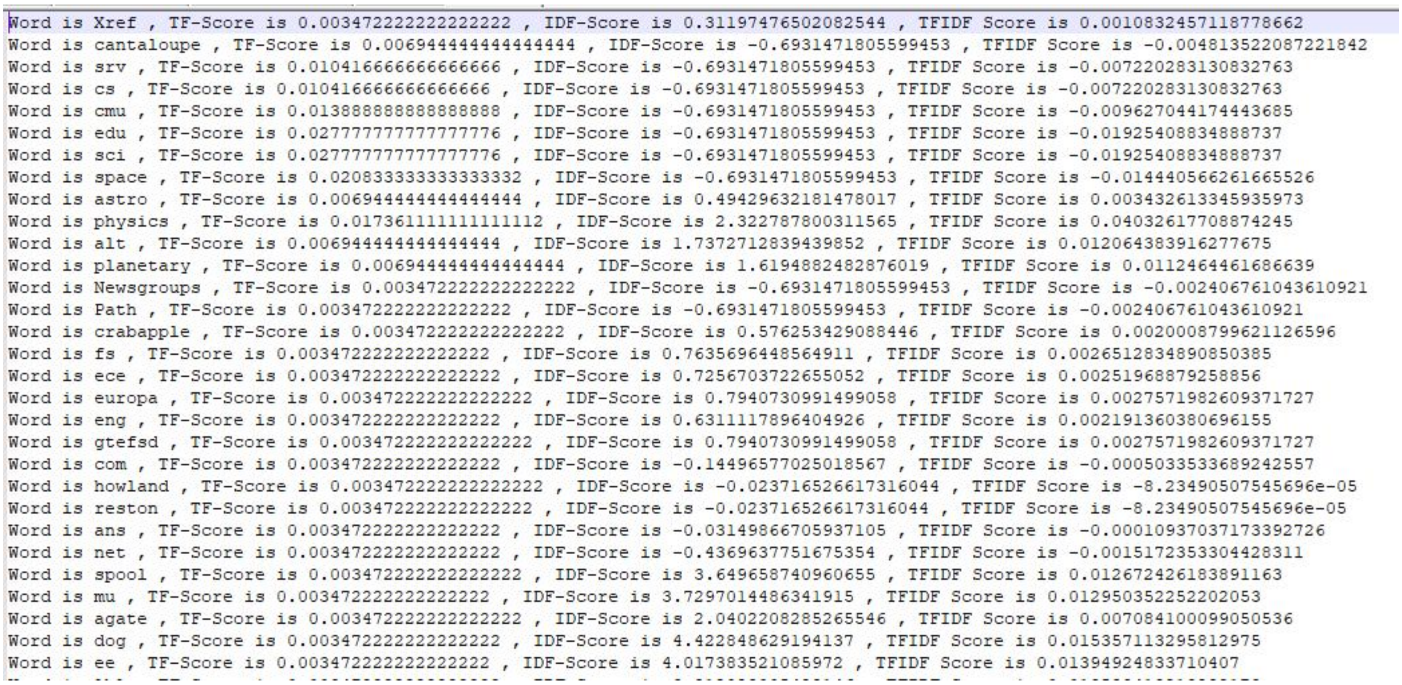
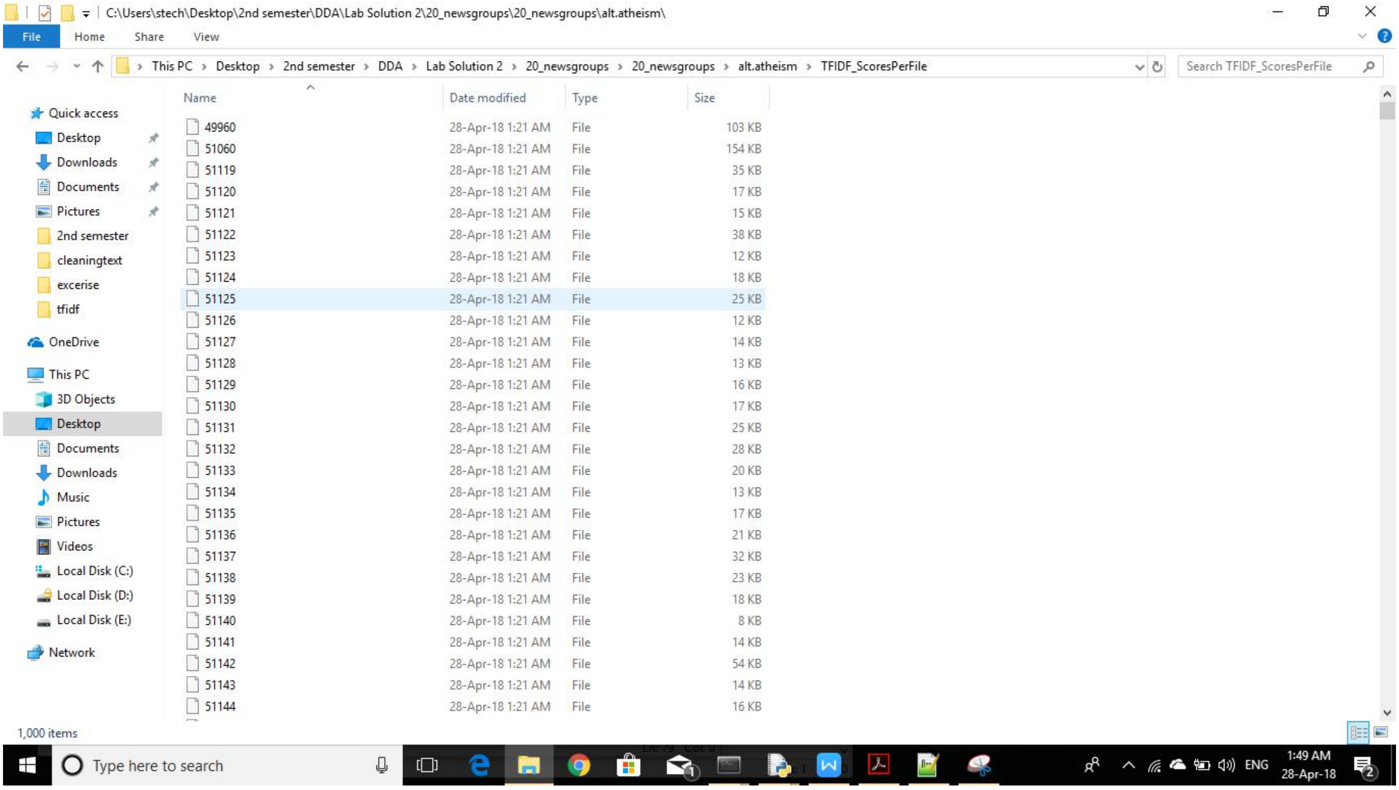


Fig 3: Chart showing number of workers against time



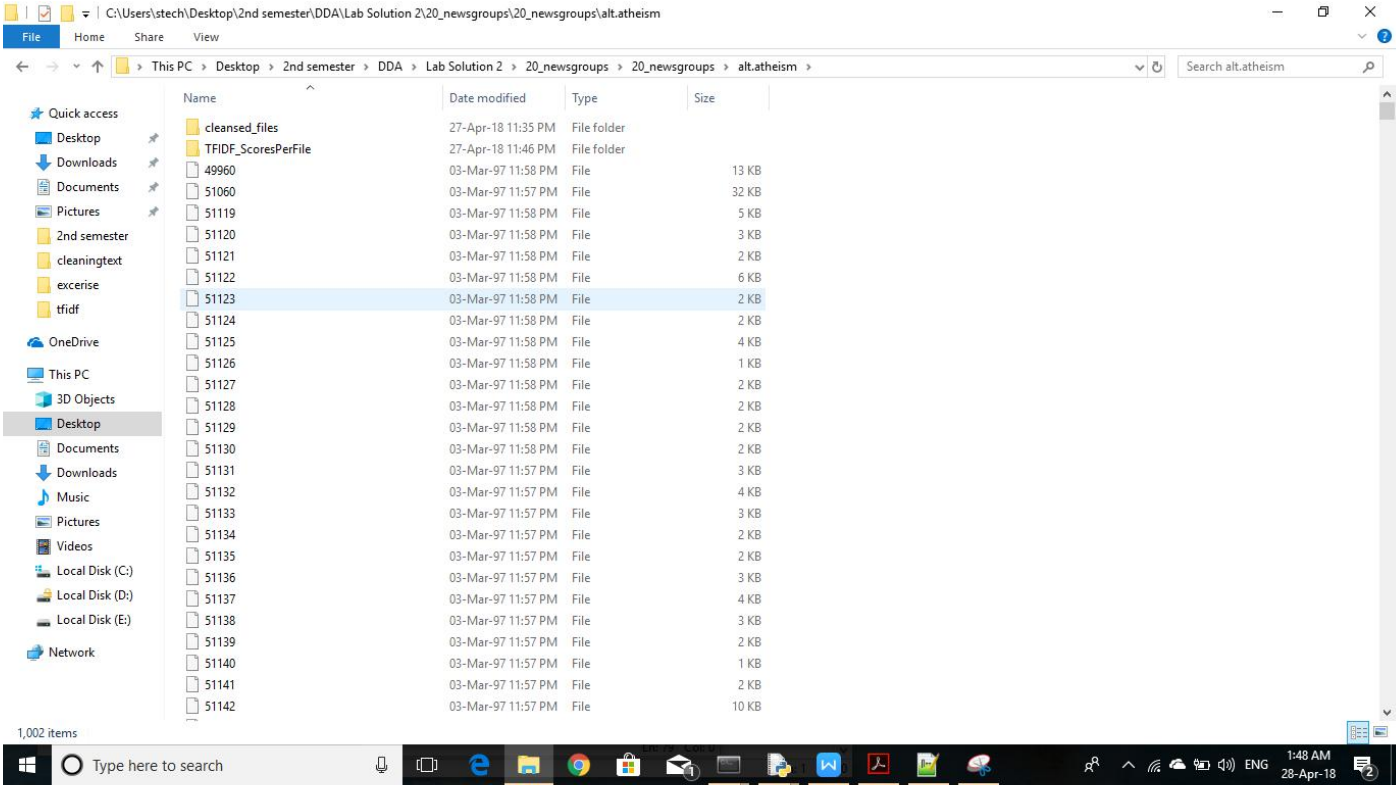
```
Word is Xref , TF-Score is 0.003472222222222222 , IDF-Score is 0.31197476502082544 , TFIDF Score is 0.0010832457118778662
Word is cantaloupe , TF-Score is 0.006944444444444444 , IDF-Score is -0.6931471805599453 , TFIDF Score is -0.004813522087221842
Word is srv , TF-Score is 0.010416666666666666 , IDF-Score is -0.6931471805599453 , TFIDF Score is -0.007220283130832763
Word is cs , TF-Score is 0.010416666666666666 , IDF-Score is -0.6931471805599453 , TFIDF Score is -0.007220283130832763
Word is cmu , TF-Score is 0.013888888888888888 , IDF-Score is -0.6931471805599453 , TFIDF Score is -0.009627044174443685
Word is edu , TF-Score is 0.027777777777777776 , IDF-Score is -0.6931471805599453 , TFIDF Score is -0.01925408834888737
Word is sci , TF-Score is 0.027777777777777776 , IDF-Score is -0.6931471805599453 , TFIDF Score is -0.01925408834888737
Word is space , TF-Score is 0.020833333333333332 , IDF-Score is -0.6931471805599453 , TFIDF Score is -0.014440566261665526
Word is astro , TF-Score is 0.006944444444444444 , IDF-Score is 0.49429632181478017 , TFIDF Score is 0.003432613345935973
Word is physics , TF-Score is 0.017361111111111112 , IDF-Score is 2.322787800311565 , TFIDF Score is 0.04032617708874245
Word is alt , TF-Score is 0.006944444444444444 , IDF-Score is 1.7372712839439852 , TFIDF Score is 0.012064383916277675
Word is planetary , TF-Score is 0.006944444444444444 , IDF-Score is 1.6194882482876019 , TFIDF Score is 0.0112464461686639
Word is Newsgroups , TF-Score is 0.003472222222222222 , IDF-Score is -0.6931471805599453 , TFIDF Score is -0.002406761043610921
Word is Path , TF-Score is 0.003472222222222222 , IDF-Score is -0.6931471805599453 , TFIDF Score is -0.002406761043610921
Word is crabapple , TF-Score is 0.003472222222222222 , IDF-Score is 0.576253429088446 , TFIDF Score is 0.0020008799621126596
Word is fs , TF-Score is 0.003472222222222222 , IDF-Score is 0.7635696448564911 , TFIDF Score is 0.0026512834890850385
Word is ece , TF-Score is 0.003472222222222222 , IDF-Score is 0.7256703722655052 , TFIDF Score is 0.00251968879258856
Word is europa , TF-Score is 0.003472222222222222 , IDF-Score is 0.7940730991499058 , TFIDF Score is 0.0027571982609371727
Word is eng , TF-Score is 0.003472222222222222 , IDF-Score is 0.6311117896404926 , TFIDF Score is 0.002191360380696155
Word is gtefsd , TF-Score is 0.003472222222222222 , IDF-Score is 0.7940730991499058 , TFIDF Score is 0.0027571982609371727
Word is com , TF-Score is 0.003472222222222222 , IDF-Score is -0.14496577025018567 , TFIDF Score is -0.0005033533689242557
Word is howland , TF-Score is 0.003472222222222222 , IDF-Score is -0.023716526617316044 , TFIDF Score is -8.23490507545696e-05
Word is reston , TF-Score is 0.003472222222222222 , IDF-Score is -0.023716526617316044 , TFIDF Score is -8.23490507545696e-05
Word is ans , TF-Score is 0.003472222222222222 , IDF-Score is -0.03149866705937105 , TFIDF Score is -0.00010937037173392726
Word is net , TF-Score is 0.003472222222222222 , IDF-Score is -0.4369637751675354 , TFIDF Score is -0.0015172353304428311
Word is spool , TF-Score is 0.003472222222222222 , IDF-Score is 3.649658740960655 , TFIDF Score is 0.012672426183891163
Word is mu , TF-Score is 0.003472222222222222 , IDF-Score is 3.7297014486341915 , TFIDF Score is 0.012950352252202053
Word is agate , TF-Score is 0.003472222222222222 , IDF-Score is 2.0402208285265546 , TFIDF Score is 0.007084100099050536
Word is dog , TF-Score is 0.003472222222222222 , IDF-Score is 4.422848629194137 , TFIDF Score is 0.015357113295812975
Word is ee , TF-Score is 0.003472222222222222 , IDF-Score is 4.017383521085972 , TFIDF Score is 0.01394924833710407
```

Fig 4: Output of TF, IDF and TFIDF scores



Name	Date modified	Type	Size
49960	28-Apr-18 1:21 AM	File	103 KB
51060	28-Apr-18 1:21 AM	File	154 KB
51119	28-Apr-18 1:21 AM	File	35 KB
51120	28-Apr-18 1:21 AM	File	17 KB
51121	28-Apr-18 1:21 AM	File	15 KB
51122	28-Apr-18 1:21 AM	File	38 KB
51123	28-Apr-18 1:21 AM	File	12 KB
51124	28-Apr-18 1:21 AM	File	18 KB
51125	28-Apr-18 1:21 AM	File	25 KB
51126	28-Apr-18 1:21 AM	File	12 KB
51127	28-Apr-18 1:21 AM	File	14 KB
51128	28-Apr-18 1:21 AM	File	13 KB
51129	28-Apr-18 1:21 AM	File	16 KB
51130	28-Apr-18 1:21 AM	File	17 KB
51131	28-Apr-18 1:21 AM	File	25 KB
51132	28-Apr-18 1:21 AM	File	28 KB
51133	28-Apr-18 1:21 AM	File	20 KB
51134	28-Apr-18 1:21 AM	File	13 KB
51135	28-Apr-18 1:21 AM	File	17 KB
51136	28-Apr-18 1:21 AM	File	21 KB
51137	28-Apr-18 1:21 AM	File	32 KB
51138	28-Apr-18 1:21 AM	File	23 KB
51139	28-Apr-18 1:21 AM	File	18 KB
51140	28-Apr-18 1:21 AM	File	8 KB
51141	28-Apr-18 1:21 AM	File	14 KB
51142	28-Apr-18 1:21 AM	File	54 KB
51143	28-Apr-18 1:21 AM	File	14 KB
51144	28-Apr-18 1:21 AM	File	16 KB

Fig 5: Files of output with TFIDF scores



Name	Date modified	Type	Size
cleansed_files	27-Apr-18 11:35 PM	File folder	
TFIDF_ScoresPerFile	27-Apr-18 11:46 PM	File folder	
49960	03-Mar-97 11:58 PM	File	13 KB
51060	03-Mar-97 11:57 PM	File	32 KB
51119	03-Mar-97 11:58 PM	File	5 KB
51120	03-Mar-97 11:58 PM	File	3 KB
51121	03-Mar-97 11:58 PM	File	2 KB
51122	03-Mar-97 11:58 PM	File	6 KB
51123	03-Mar-97 11:58 PM	File	2 KB
51124	03-Mar-97 11:58 PM	File	2 KB
51125	03-Mar-97 11:58 PM	File	4 KB
51126	03-Mar-97 11:58 PM	File	1 KB
51127	03-Mar-97 11:58 PM	File	2 KB
51128	03-Mar-97 11:58 PM	File	2 KB
51129	03-Mar-97 11:58 PM	File	2 KB
51130	03-Mar-97 11:58 PM	File	2 KB
51131	03-Mar-97 11:57 PM	File	3 KB
51132	03-Mar-97 11:57 PM	File	4 KB
51133	03-Mar-97 11:57 PM	File	3 KB
51134	03-Mar-97 11:57 PM	File	2 KB
51135	03-Mar-97 11:57 PM	File	2 KB
51136	03-Mar-97 11:57 PM	File	3 KB
51137	03-Mar-97 11:57 PM	File	4 KB
51138	03-Mar-97 11:57 PM	File	3 KB
51139	03-Mar-97 11:57 PM	File	2 KB
51140	03-Mar-97 11:57 PM	File	1 KB
51141	03-Mar-97 11:57 PM	File	2 KB
51142	03-Mar-97 11:57 PM	File	10 KB

Fig 6: Two folders containing cleansed files and scores of words for a specific file.