



# Artificial Intelligence Project: POS Tagging

September 28, 2020

RITIK JAIN 18114068  
PRIYANSHU GARG 181140058  
YUGANTAR ARYA 18114084

# Contents

<b>1</b>	<b>Week 1</b>	<b>i</b>
1.1	Objective . . . . .	i
1.2	Procedure . . . . .	i
1.2.1	Phase 1 Filtration . . . . .	i
1.2.2	Phase 2 Collection . . . . .	i
1.2.3	Phase 3 Verification . . . . .	i
1.3	Code Screenshots . . . . .	ii
1.4	Execution . . . . .	v
1.5	Output Files . . . . .	vi
1.6	Results . . . . .	ix

# 1 Week 1

**Date:** September 28, 2020

## 1.1 Objective

To filter the text by extracting word and POS tag for the word

## 1.2 Procedure

The program does its job in 3 phases. The number of phases could have been reduced to 1 but I went with this code since it doesn't take much time to pre-process all the files and it has to be done only once; hence I preferred clarity of code over the small extra time it takes.

### 1.2.1 Phase 1 Filtration

The program processes all the raw xml files and converts them into filtered files. The program makes use of the regex, which is the heart of the program:

$$<\backslash s^*w[^>]^*pos="([A-Za-z]^*)"[^>]^*>\backslash s^*([^\backslash s]^*)\backslash s^*<\backslash s^*/w\backslash s^*>$$

The above regex is a flexible one. It could have been made stricter, given the consistency of the dataset, but I still went with the flexible version since it is more general. A filtered file contains all the word\_tag combinations for its xml file. The word\_tag combinations are stored in a case sensitive manner.

### 1.2.2 Phase 2 Collection

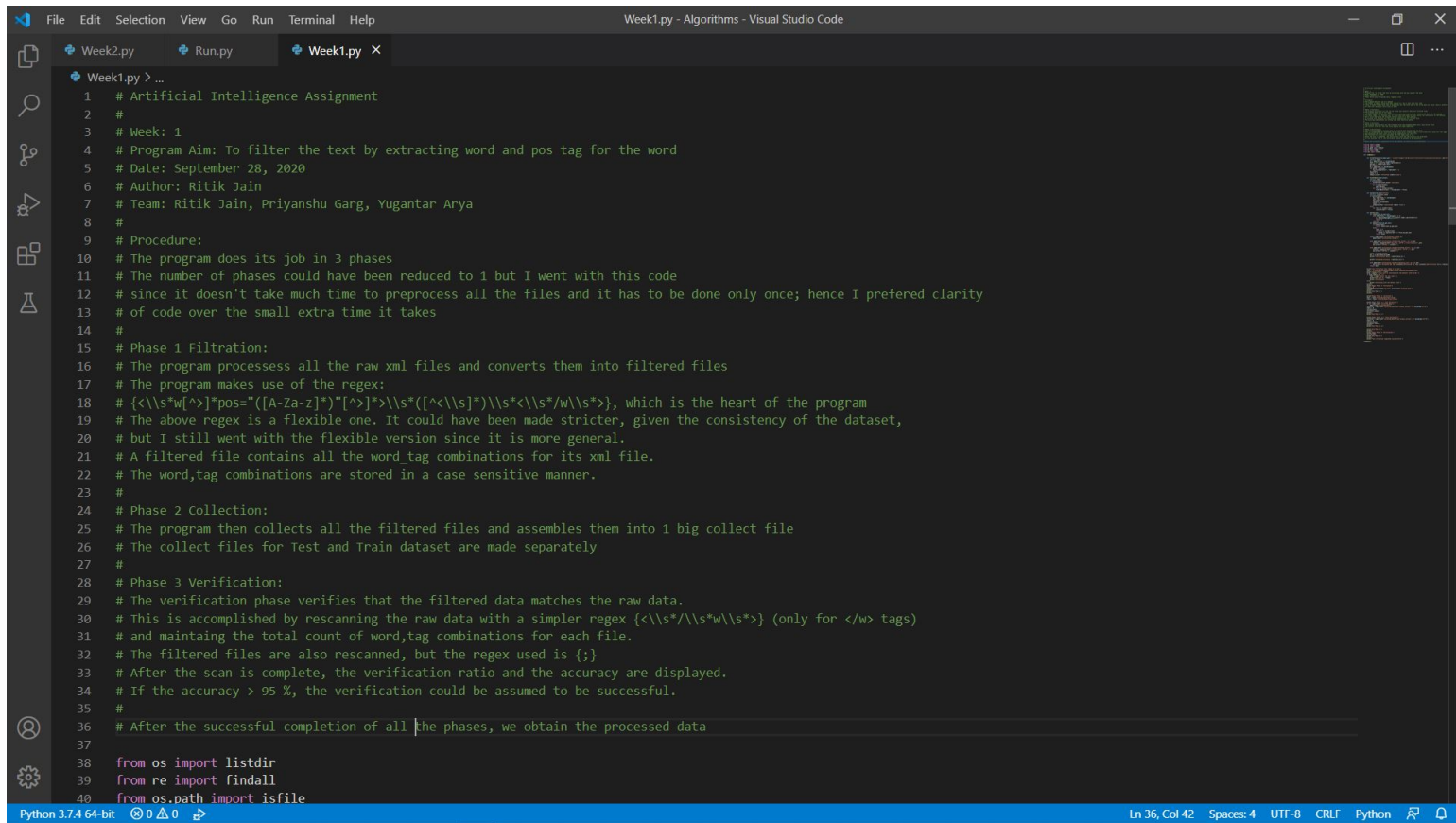
The program then collects all the filtered files and assembles them into 1 big collect file. The collect files for Test and Train dataset are made separately.

### 1.2.3 Phase 3 Verification

The verification phase verifies that the filtered data matches the raw data. This is accomplished by re-scanning the raw data with a simpler regex  $<\backslash s^*/\backslash s^*w\backslash s^*>$  (only for  $</w>$  tags) and maintaining the total count of word\_tag combinations for each file. The filtered files are also re-scanned, but the regex used is  $;$ . After the scan is complete, the verification ratio and the accuracy are displayed. If the accuracy is more than 95%, the verification could be assumed to be successful.

After the successful completion of all the phases, we obtain the processed data.

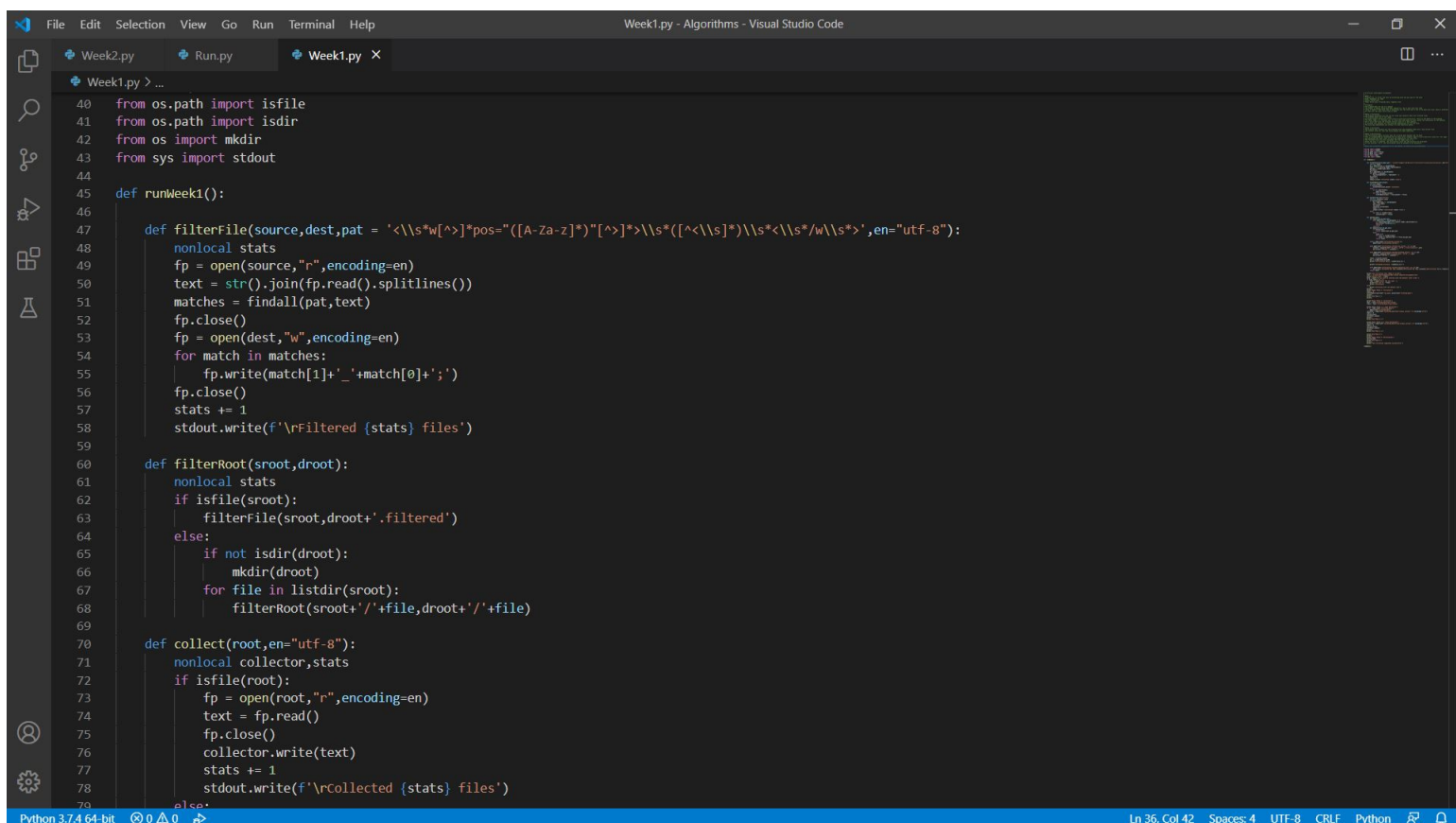
## 1.3 Code Screenshots



```
File Edit Selection View Go Run Terminal Help
Week1.py - Algorithms - Visual Studio Code

Week1.py > ...
1 # Artificial Intelligence Assignment
2 #
3 # Week: 1
4 # Program Aim: To filter the text by extracting word and pos tag for the word
5 # Date: September 28, 2020
6 # Author: Ritik Jain
7 # Team: Ritik Jain, Priyanshu Garg, Yugantar Arya
8 #
9 # Procedure:
10 # The program does its job in 3 phases
11 # The number of phases could have been reduced to 1 but I went with this code
12 # since it doesn't take much time to preprocess all the files and it has to be done only once; hence I preferred clarity
13 # of code over the small extra time it takes
14 #
15 # Phase 1 Filtration:
16 # The program processes all the raw xml files and converts them into filtered files
17 # The program makes use of the regex:
18 # {<\s*w[>]*pos="([A-Za-z]*)"[>]*>\s*([^\s]*)\s*<\s*w\s*>}, which is the heart of the program
19 # The above regex is a flexible one. It could have been made stricter, given the consistency of the dataset,
20 # but I still went with the flexible version since it is more general.
21 # A filtered file contains all the word_tag combinations for its xml file.
22 # The word_tag combinations are stored in a case sensitive manner.
23 #
24 # Phase 2 Collection:
25 # The program then collects all the filtered files and assembles them into 1 big collect file
26 # The collect files for Test and Train dataset are made separately
27 #
28 # Phase 3 Verification:
29 # The verification phase verifies that the filtered data matches the raw data.
30 # This is accomplished by rescanning the raw data with a simpler regex {<\s*/\s*w\s*>} (only for </w> tags)
31 # and maintaing the total count of word_tag combinations for each file.
32 # The filtered files are also rescanned, but the regex used is {;}
33 # After the scan is complete, the verification ratio and the accuracy are displayed.
34 # If the accuracy > 95 %, the verification could be assumed to be successful.
35 #
36 # After the successful completion of all the phases, we obtain the processed data
37
38 from os import listdir
39 from re import findall
40 from os.path import isfile
```

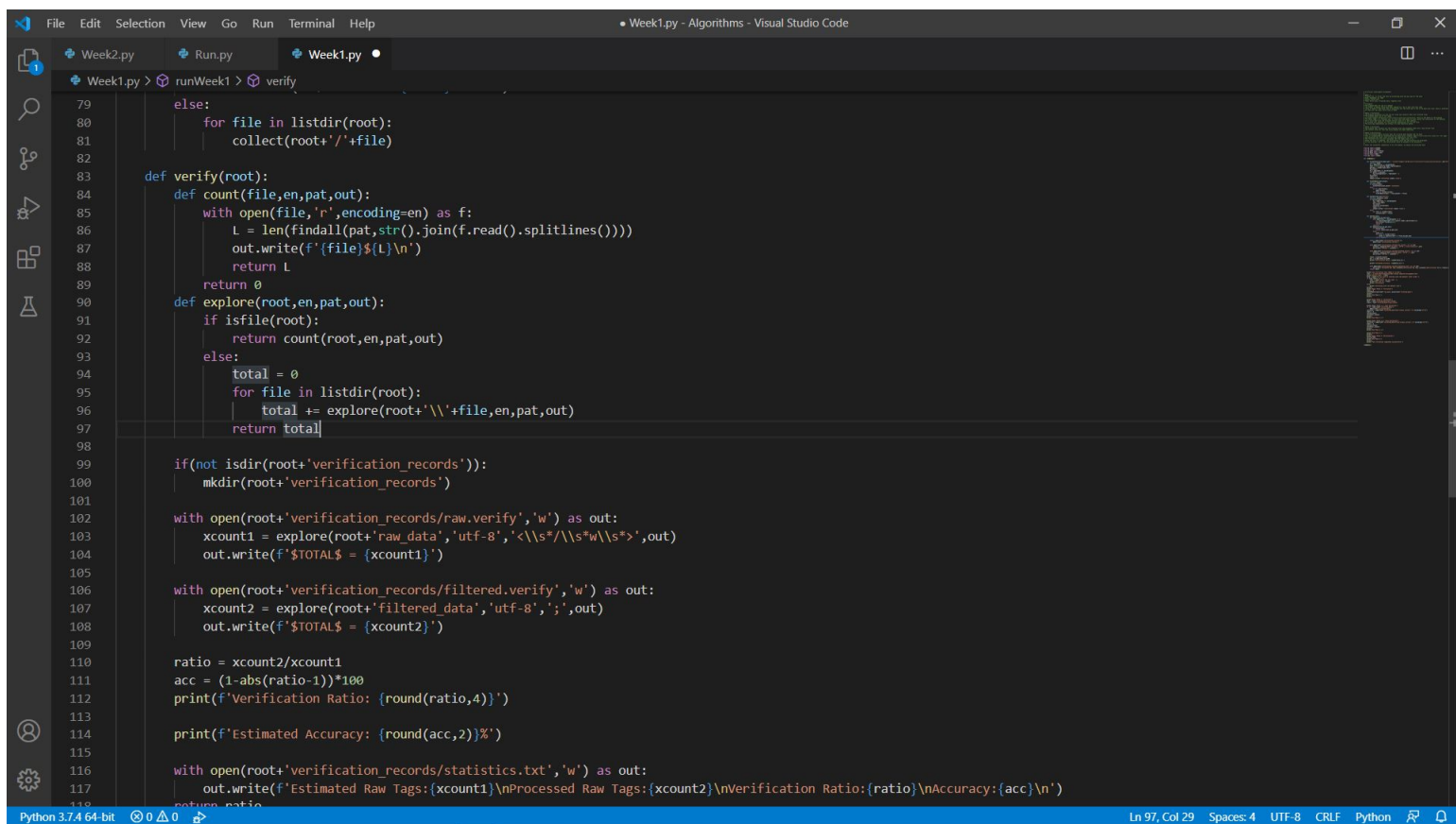
Figure 1: Week-1 File: Section 1



```
File Edit Selection View Go Run Terminal Help
Week1.py - Algorithms - Visual Studio Code

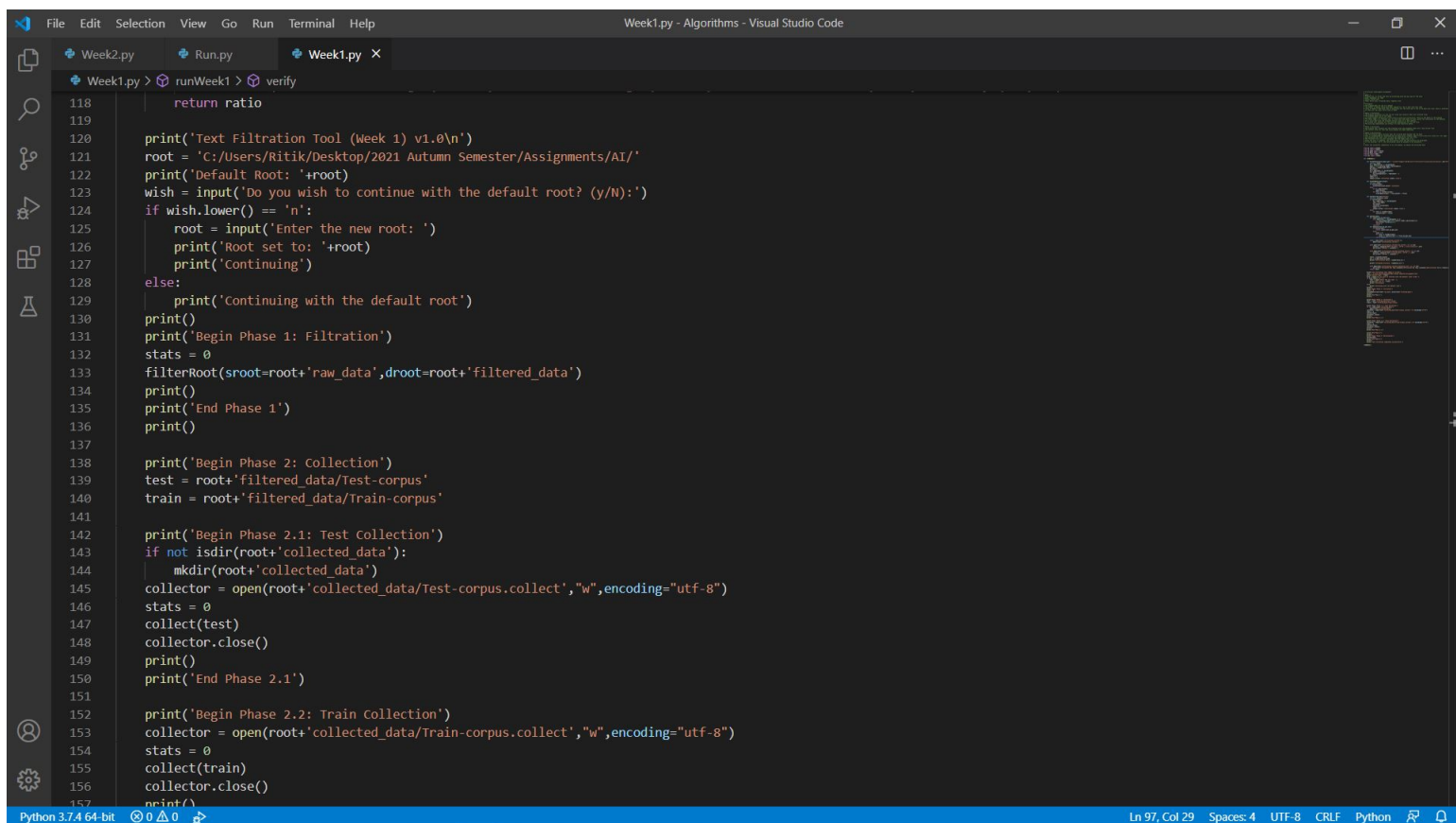
Week1.py > ...
40 from os.path import isfile
41 from os.path import isdir
42 from os import mkdir
43 from sys import stdout
44
45 def runWeek1():
46
47     def filterFile(source,dest,pat = '<\s*w[>]*pos="([A-Za-z]*)"[>]*>\s*([^\s]*)\s*<\s*w\s*>','en="utf-8")':
48         nonlocal stats
49         fp = open(source,"r",encoding=en)
50         text = str().join(fp.read().splitlines())
51         matches = findall(pat,text)
52         fp.close()
53         fp = open(dest,"w",encoding=en)
54         for match in matches:
55             fp.write(match[1]+'_'+match[0]+';')
56         fp.close()
57         stats += 1
58         stdout.write(f'\rFiltered {stats} files')
59
60     def filterRoot(sroot,droot):
61         nonlocal stats
62         if isfile(sroot):
63             filterFile(sroot,droot+'.filtered')
64         else:
65             if not isdir(droot):
66                 mkdir(droot)
67             for file in listdir(sroot):
68                 filterRoot(sroot+'/'+file,droot+'/'+file)
69
70     def collect(root,en="utf-8"):
71         nonlocal collector,stats
72         if isfile(root):
73             fp = open(root,"r",encoding=en)
74             text = fp.read()
75             fp.close()
76             collector.write(text)
77             stats += 1
78             stdout.write(f'\rCollected {stats} files')
79         else:
```

Figure 2: Week-1 File: Section 2



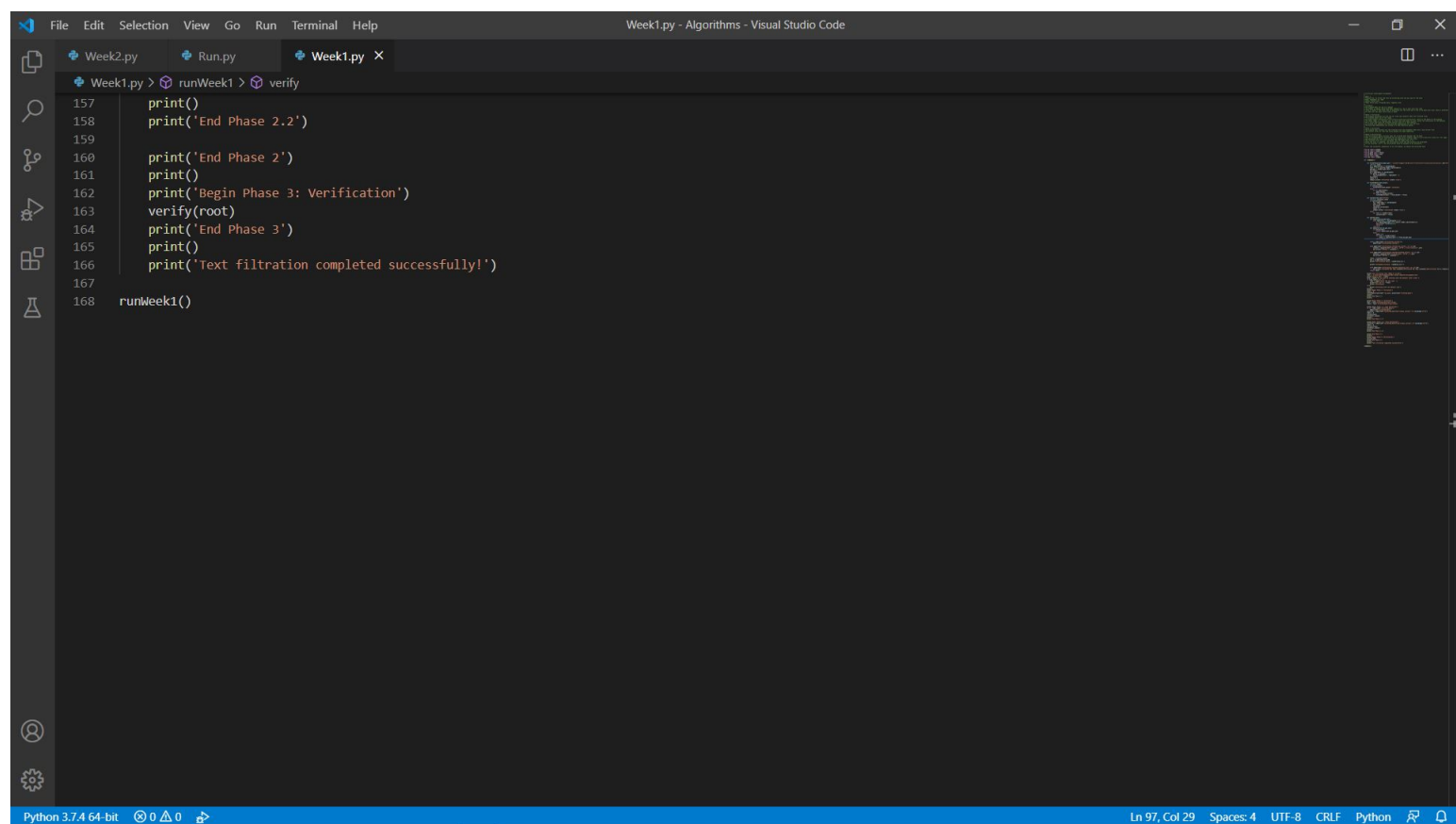
```
79         else:
80             for file in listdir(root):
81                 collect(root+'/'+file)
82
83     def verify(root):
84         def count(file,en,pat,out):
85             with open(file,'r',encoding=en) as f:
86                 L = len(findall(pat,str(f.read()).splitlines()))
87                 out.write(f'{file}${L}\n')
88                 return L
89             return 0
90         def explore(root,en,pat,out):
91             if isfile(root):
92                 return count(root,en,pat,out)
93             else:
94                 total = 0
95                 for file in listdir(root):
96                     total += explore(root+'\\'+file,en,pat,out)
97             return total
98
99         if(not isdir(root+'verification_records')):
100             mkdir(root+'verification_records')
101
102         with open(root+'verification_records/raw.verify','w') as out:
103             xcount1 = explore(root+'raw_data','utf-8','<\\s*/\\s*w\\s*>',out)
104             out.write(f'$TOTAL$ = {xcount1}')
105
106         with open(root+'verification_records/filtered.verify','w') as out:
107             xcount2 = explore(root+'filtered_data','utf-8',';',out)
108             out.write(f'$TOTAL$ = {xcount2}')
109
110         ratio = xcount2/xcount1
111         acc = (1-abs(ratio-1))*100
112         print(f'Verification Ratio: {round(ratio,4)}')
113
114         print(f'Estimated Accuracy: {round(acc,2)}%')
115
116         with open(root+'verification_records/statistics.txt','w') as out:
117             out.write(f'Estimated Raw Tags:{xcount1}\nProcessed Raw Tags:{xcount2}\nVerification Ratio:{ratio}\nAccuracy:{acc}\n')
118             return ratio
```

Figure 3: Week-1 File: Section 3



```
118         return ratio
119
120     print('Text Filtration Tool (Week 1) v1.0\n')
121     root = 'C:/Users/Ritik/Desktop/2021 Autumn Semester/Assignments/AI/'
122     print('Default Root: '+root)
123     wish = input('Do you wish to continue with the default root? (y/N):')
124     if wish.lower() == 'n':
125         root = input('Enter the new root: ')
126         print('Root set to: '+root)
127         print('Continuing')
128     else:
129         print('Continuing with the default root')
130     print()
131     print('Begin Phase 1: Filtration')
132     stats = 0
133     filterRoot(sroot=root+'raw_data',droot=root+'filtered_data')
134     print()
135     print('End Phase 1')
136     print()
137
138     print('Begin Phase 2: Collection')
139     test = root+'filtered_data/Test-corpus'
140     train = root+'filtered_data/Train-corpus'
141
142     print('Begin Phase 2.1: Test Collection')
143     if not isdir(root+'collected_data'):
144         mkdir(root+'collected_data')
145     collector = open(root+'collected_data/Test-corpus.collect','w',encoding='utf-8')
146     stats = 0
147     collect(test)
148     collector.close()
149     print()
150     print('End Phase 2.1')
151
152     print('Begin Phase 2.2: Train Collection')
153     collector = open(root+'collected_data/Train-corpus.collect','w',encoding='utf-8')
154     stats = 0
155     collect(train)
156     collector.close()
157     print()
```

Figure 4: Week-1 File: Section 4

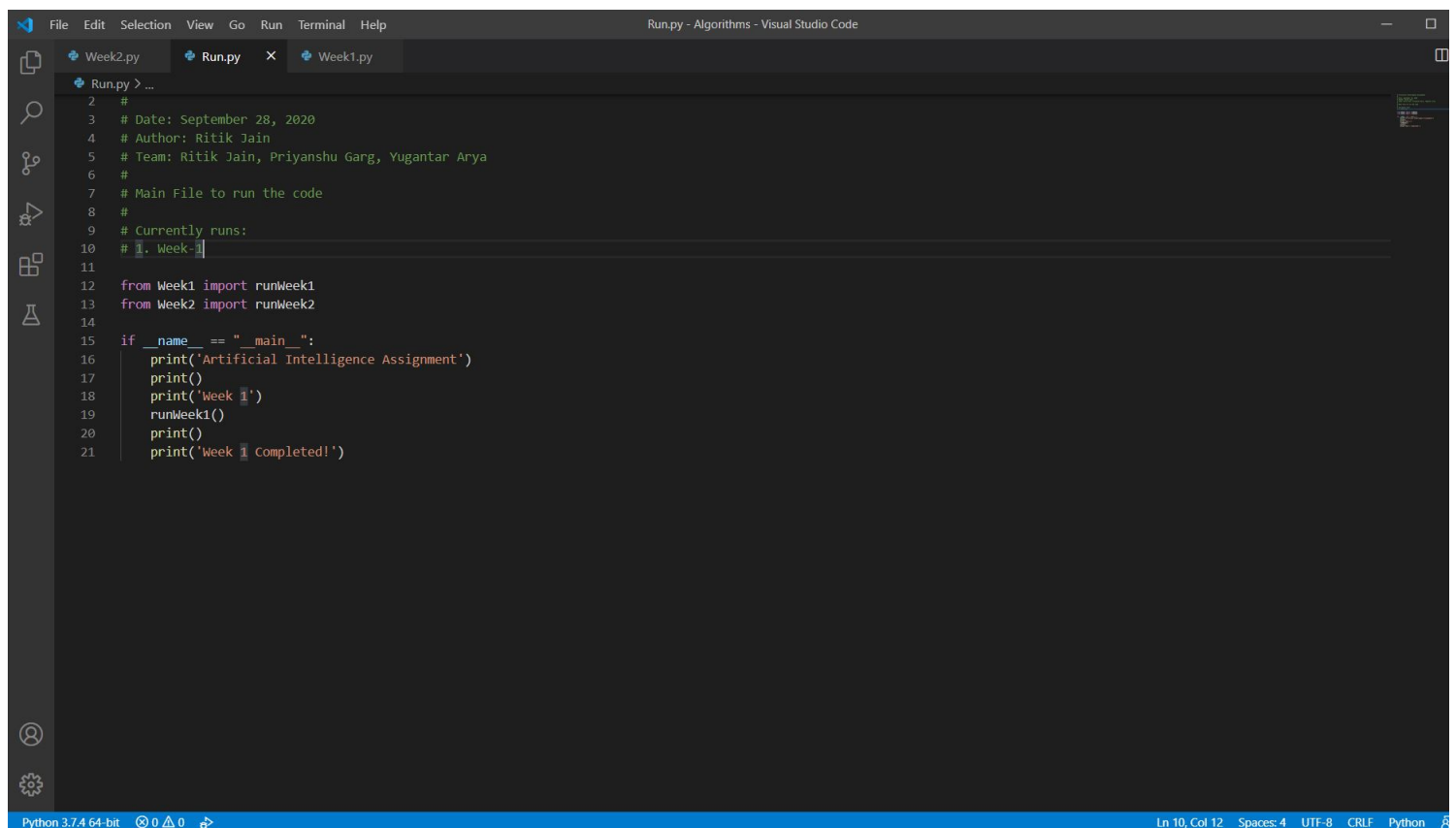


The screenshot shows the Visual Studio Code editor with the file 'Week1.py' open. The editor has a dark theme. The file explorer on the left shows three files: 'Week2.py', 'Run.py', and 'Week1.py'. The 'Week1.py' file is selected and its content is displayed in the editor. The code is as follows:

```
157 print()
158 print('End Phase 2.2')
159
160 print('End Phase 2')
161 print()
162 print('Begin Phase 3: Verification')
163 verify(root)
164 print('End Phase 3')
165 print()
166 print('Text filtration completed successfully!')
167
168 runWeek1()
```

The status bar at the bottom indicates 'Python 3.7.4 64-bit', 'Ln 97, Col 29', 'Spaces: 4', 'UTF-8', 'CRLF', and 'Python'.

Figure 5: Week-1 File: Section 5



The screenshot shows the Visual Studio Code editor with the file 'Run.py' open. The editor has a dark theme. The file explorer on the left shows three files: 'Week2.py', 'Run.py', and 'Week1.py'. The 'Run.py' file is selected and its content is displayed in the editor. The code is as follows:

```
2 #
3 # Date: September 28, 2020
4 # Author: Ritik Jain
5 # Team: Ritik Jain, Priyanshu Garg, Yugantar Arya
6 #
7 # Main File to run the code
8 #
9 # Currently runs:
10 # 1. Week-1
11
12 from Week1 import runWeek1
13 from Week2 import runWeek2
14
15 if __name__ == "__main__":
16     print('Artificial Intelligence Assignment')
17     print()
18     print('Week 1')
19     runWeek1()
20     print()
21     print('Week 1 Completed!')
```

The status bar at the bottom indicates 'Python 3.7.4 64-bit', 'Ln 10, Col 12', 'Spaces: 4', 'UTF-8', 'CRLF', and 'Python'.

Figure 6: Main File: Run.py



## 1.4 Execution

C:\windows\system32\cmd.exe

```
C:\Users\Ritik\Desktop\2021 Autumn Semester\Assignments\AI\ArtificialIntelligenceProject\Algorithms>python run  
Artificial Intelligence Assignment
```

Week 1

Text Filtration Tool (Week 1) v1.0

Default Root: C:/Users/Ritik/Desktop/2021 Autumn Semester/Assignments/AI/

Do you wish to continue with the default root? (y/N):y

Continuing with the default root

Begin Phase 1: Filtration

Filtered 635 files

End Phase 1

Begin Phase 2: Collection

Begin Phase 2.1: Test Collection

Collected 115 files

End Phase 2.1

Begin Phase 2.2: Train Collection

Collected 520 files

End Phase 2.2

End Phase 2

Begin Phase 3: Verification

Verification Ratio: 1.0

Estimated Accuracy: 100.0%

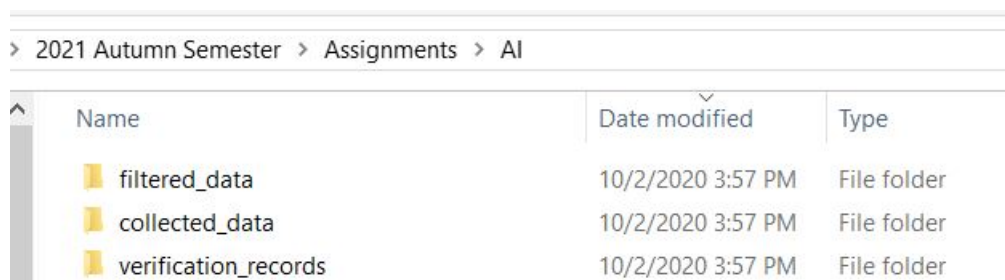
End Phase 3

Text filtration completed successfully!

Week 1 Completed!

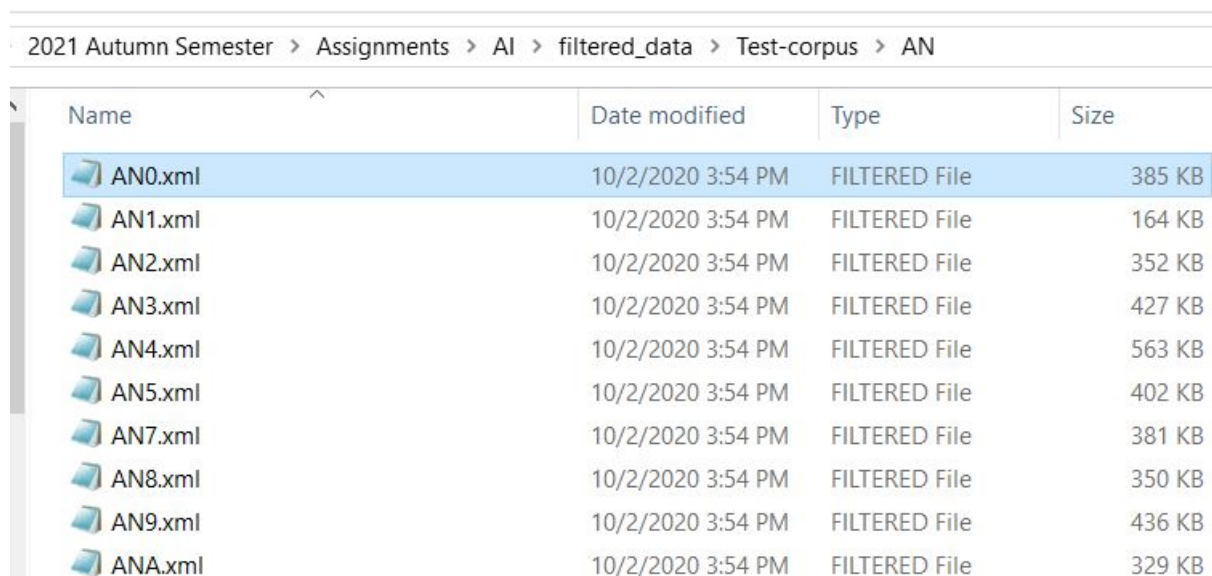
Figure 7: Execution: Run.py

## 1.5 Output Files



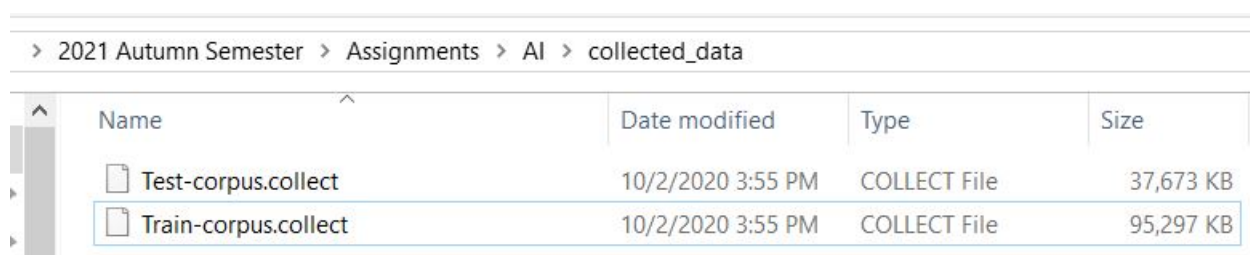
> 2021 Autumn Semester > Assignments > AI			
Name	Date modified	Type	
filtered_data	10/2/2020 3:57 PM	File folder	
collected_data	10/2/2020 3:57 PM	File folder	
verification_records	10/2/2020 3:57 PM	File folder	

Figure 8: Generated Folders



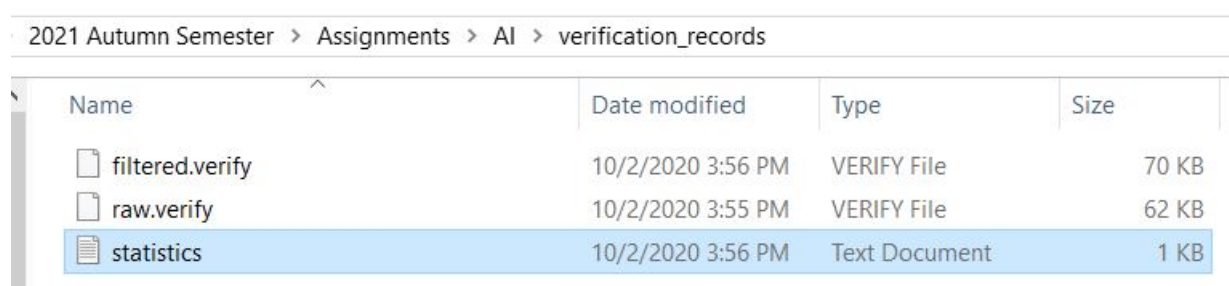
2021 Autumn Semester > Assignments > AI > filtered_data > Test-corpus > AN				
Name	Date modified	Type	Size	
AN0.xml	10/2/2020 3:54 PM	FILTERED File	385 KB	
AN1.xml	10/2/2020 3:54 PM	FILTERED File	164 KB	
AN2.xml	10/2/2020 3:54 PM	FILTERED File	352 KB	
AN3.xml	10/2/2020 3:54 PM	FILTERED File	427 KB	
AN4.xml	10/2/2020 3:54 PM	FILTERED File	563 KB	
AN5.xml	10/2/2020 3:54 PM	FILTERED File	402 KB	
AN7.xml	10/2/2020 3:54 PM	FILTERED File	381 KB	
AN8.xml	10/2/2020 3:54 PM	FILTERED File	350 KB	
AN9.xml	10/2/2020 3:54 PM	FILTERED File	436 KB	
ANA.xml	10/2/2020 3:54 PM	FILTERED File	329 KB	

Figure 9: Phase 1 Filtered Files



> 2021 Autumn Semester > Assignments > AI > collected_data				
Name	Date modified	Type	Size	
Test-corpus.collect	10/2/2020 3:55 PM	COLLECT File	37,673 KB	
Train-corpus.collect	10/2/2020 3:55 PM	COLLECT File	95,297 KB	

Figure 10: Phase 2 Collected Files



2021 Autumn Semester > Assignments > AI > verification_records				
Name	Date modified	Type	Size	
filtered.verify	10/2/2020 3:56 PM	VERIFY File	70 KB	
raw.verify	10/2/2020 3:55 PM	VERIFY File	62 KB	
statistics	10/2/2020 3:56 PM	Text Document	1 KB	

Figure 11: Phase 3 Verification Files



# AN5.xml - Notepad

File Edit Format View Help

Parental\_ADJ;rights\_SUBST;and\_CONJ;responsibilities\_SUBST;INTRODUCTION\_SUBST  
J;policy\_SUBST;considerations\_SUBST;might\_VERB;outweigh\_VERB;individual\_ADJ;  
RB;have\_VERB;regard\_SUBST;when\_CONJ;exercising\_VERB;their\_PRON;functions\_SUB  
dren\_SUBST;'s\_UNC;and\_CONJ;state\_VERB;interests\_SUBST;and\_CONJ;rights\_SUBST;  
n\_ADJ;decisions\_SUBST;in\_PREP;respect\_SUBST;of\_PREP;their\_PRON;own\_ADJ;lives  
NJ;child\_SUBST;interrelate\_SUBST;Put\_VERB;simply\_ADV;LEAs\_SUBST;have\_VERB;a\_  
BST;that\_CONJ;the\_ART;state\_SUBST;system\_SUBST;of\_PREP;education\_SUBST;is\_VE  
and\_CONJ;shameful\_ADJ;to\_PREP;himself\_PRON;The\_ART;distribution\_SUBST;of\_PRE  
ERB;not\_ADV;surprising\_ADJ;then\_ADV;that\_CONJ;areas\_SUBST;such\_ADJ;as\_PREP;s  
er\_PRON;from\_PREP;school\_SUBST;The\_ART;case\_SUBST;of\_PREP;RE\_SUBST;S\_SUBST;A  
\_CONJ;local\_ADJ;authority\_SUBST;rather\_ADV;than\_CONJ;those\_ADJ;of\_PREP;the\_A  
BST;in\_PREP;the\_ART;balance\_SUBST;of\_PREP;power\_SUBST;between\_PREP;parent\_SU  
J;there\_PRON;is\_VERB;inevitably\_ADV;dispute\_SUBST;between\_PREP;parent\_SUBST;  
With\_PREP;the\_ART;exception\_SUBST;of\_PREP;the\_ART;courts\_SUBST;those\_ADJ;det  
;challenged\_VERB;the\_ART;LEA\_SUBST;'s\_UNC;decisions\_SUBST;both\_ADV;individua  
the\_ART;case\_SUBST;at\_PREP;the\_ART;school\_SUBST;designated\_VERB;by\_PREP;the\_  
NJ;found\_VERB;places\_SUBST;for\_PREP;the\_ART;children\_SUBST;at\_PREP;two\_ADJ;s

Figure 12: A filtered File

# Test-corpus.collect - Notepad

File Edit Format View Help

THE\_ART;PLAYERS\_SUBST;Virtually\_ADV;every\_ART;country\_SUBST;in\_PREP;the\_ART;world\_SUBST;has\_VERB;so  
RT;process\_SUBST;exaggerates\_VERB;and\_CONJ;distorts\_VERB;its\_PRON;reports\_SUBST;so\_ADV;as\_CONJ;to\_F  
B;that\_CONJ;this\_ADJ;is\_VERB;only\_ADV;fiction\_SUBST;there\_ADV;nevertheless\_ADV;remains\_VERB;a\_ART;f  
;the\_ART;endless\_ADJ;paperwork\_SUBST;that\_CONJ;forms\_VERB;the\_ART;major\_ADJ;part\_SUBST;of\_PREP;any\_  
ne\_ADJ;recent\_ADJ;advertisement\_SUBST;carried\_VERB;a\_ART;photograph\_SUBST;of\_PREP;a\_ART;rather\_ADV;s  
nt\_SUBST;Six\_ADJ;MI6\_UNC;often\_ADV;referred\_VERB;to\_PREP;as\_PREP;the\_ART;Secret\_ADJ;Intelligence\_SU  
e\_ART;Foreign\_ADJ;Office\_SUBST;MI5\_UNC;'s\_VERB;initial\_ADJ;task\_SUBST;was\_VERB;to\_PREP;counter\_VERB  
a\_ART;particularly\_ADV;difficult\_ADJ;task\_SUBST;since\_CONJ;they\_PRON;were\_VERB;all\_ADJ;pathetically  
uestion\_SUBST;Whether\_CONJ;the\_ART;Germans\_SUBST;were\_VERB;quite\_ADV;as\_ADV;easily\_ADV;fooled\_VERB;  
RON;own\_ADJ;views\_SUBST;and\_CONJ;no\_ART;one\_PRON;was\_VERB;willing\_ADJ;to\_PREP;disagree\_VERB;with\_PR  
PREP;1952\_ADJ;the\_ART;then\_ADJ;Home\_SUBST;Secretary\_SUBST;Sir\_SUBST;David\_SUBST;Maxwell\_SUBST;Fyfe\_  
\_VERB;the\_ART;Defence\_SUBST;of\_PREP;the\_ART;Realm\_SUBST;as\_PREP;a\_ART;whole\_SUBST;from\_PREP;externa  
VERB;colour\_SUBST;to\_PREP;any\_ADJ;suggestion\_SUBST;that\_CONJ;it\_PRON;is\_VERB;concerned\_ADJ;with\_PRE  
RT;Security\_SUBST;Service\_SUBST;in\_PREP;particular\_ADJ;cases\_SUBST;but\_CONJ;are\_VERB;furnished\_VERB  
BST;First\_ADJ;though\_CONJ;MI5\_UNC;is\_VERB;notionally\_ADV;under\_PREP;the\_ART;control\_SUBST;of\_PREP;t  
BST;is\_VERB;consistently\_ADV;and\_CONJ;deliberately\_ADV;ignored\_VERB;The\_ART;absence\_SUBST;of\_PREP;a  
nthonny\_SUBST;Blunt\_SUBST;who\_PRON;finally\_ADV;confessed\_VERB;in\_PREP;1964\_ADJ;to\_PREP;being\_VERB;a\_  
al\_ADJ;secrets\_SUBST;while\_CONJ;working\_VERB;at\_PREP;the\_ART;Admiralty\_SUBST;Frank\_SUBST;Bossard\_SU  
e\_ART;network\_SUBST;consisted\_VERB;of\_PREP;Harry\_SUBST;Houghton\_SUBST;and\_CONJ;Ethel\_SUBST;Gee\_INTE  
SUBST;is\_VERB;that\_CONJ;MI5\_UNC;has\_VERB;been\_VERB;penetrated\_VERB;by\_PREP;Russian\_ADJ;intelligence  
PREP;a\_ART;homosexual\_SUBST;and\_CONJ;blackmailed\_VERB;He\_PRON;later\_ADV;lived\_VERB;in\_PREP;an\_ART;e  
5\_UNC;'s\_UNC;inefficiency\_SUBST;is\_VERB;that\_CONJ;it\_PRON;wastes\_VERB;far\_ADV;too\_ADV;much\_ADJ;time  
;student\_SUBST;and\_CONJ;industrial\_ADJ;militancy\_SUBST;in\_PREP;Britain\_SUBST;the\_ART;numbers\_SUBST;  
\_UNC;thinks\_VERB;there\_PRON;are\_VERB;that\_ADV;many\_ADJ;subversives\_SUBST;spies\_SUBST;and\_CONJ;trait  
SUBST;that\_CONJ;MI5\_UNC;and\_CONJ;Special\_ADJ;Branch\_SUBST;alone\_ADV;decide\_VERB;who\_PRON;merits\_VER  
J;no\_ART;minister\_SUBST;would\_VERB;authorise\_VERB;a\_ART;warrant\_SUBST;for\_PREP;such\_ADJ;tapping\_VER  
f\_PREP;£1.6\_UNC;billion\_ADJ;on\_ADV;to\_PREP;what\_PRON;will\_VERB;become\_VERB;the\_ART;country\_SUBST;'s

Figure 13: A collected File



raw.verify - Notepad

File Edit Format View Help

```
C:/Users/Ritik/Desktop/2021 Autumn Semester/Assignments/AI/raw_data/Test-corpus\AN\AN0.xml$36451
C:/Users/Ritik/Desktop/2021 Autumn Semester/Assignments/AI/raw_data/Test-corpus\AN\AN1.xml$15823
C:/Users/Ritik/Desktop/2021 Autumn Semester/Assignments/AI/raw_data/Test-corpus\AN\AN2.xml$35145
C:/Users/Ritik/Desktop/2021 Autumn Semester/Assignments/AI/raw_data/Test-corpus\AN\AN3.xml$39292
C:/Users/Ritik/Desktop/2021 Autumn Semester/Assignments/AI/raw_data/Test-corpus\AN\AN4.xml$53723
C:/Users/Ritik/Desktop/2021 Autumn Semester/Assignments/AI/raw_data/Test-corpus\AN\AN5.xml$37421
C:/Users/Ritik/Desktop/2021 Autumn Semester/Assignments/AI/raw_data/Test-corpus\AN\AN7.xml$38382
C:/Users/Ritik/Desktop/2021 Autumn Semester/Assignments/AI/raw_data/Test-corpus\AN\AN8.xml$35287
C:/Users/Ritik/Desktop/2021 Autumn Semester/Assignments/AI/raw_data/Test-corpus\AN\AN9.xml$40765
C:/Users/Ritik/Desktop/2021 Autumn Semester/Assignments/AI/raw_data/Test-corpus\AN\ANA.xml$30970
C:/Users/Ritik/Desktop/2021 Autumn Semester/Assignments/AI/raw_data/Test-corpus\AN\ANB.xml$31242
C:/Users/Ritik/Desktop/2021 Autumn Semester/Assignments/AI/raw_data/Test-corpus\AN\ANC.xml$37904
C:/Users/Ritik/Desktop/2021 Autumn Semester/Assignments/AI/raw_data/Test-corpus\AN\AND.xml$34872
C:/Users/Ritik/Desktop/2021 Autumn Semester/Assignments/AI/raw_data/Test-corpus\AN\ANF.xml$32813
```

Figure 14: Raw Data Verification File

filtered.verify - Notepad

File Edit Format View Help

```
C:/Users/Ritik/Desktop/2021 Autumn Semester/Assignments/AI/filtered_data/Test-corpus\AN\AN0.xml.filtered$36454
C:/Users/Ritik/Desktop/2021 Autumn Semester/Assignments/AI/filtered_data/Test-corpus\AN\AN1.xml.filtered$15823
C:/Users/Ritik/Desktop/2021 Autumn Semester/Assignments/AI/filtered_data/Test-corpus\AN\AN2.xml.filtered$35147
C:/Users/Ritik/Desktop/2021 Autumn Semester/Assignments/AI/filtered_data/Test-corpus\AN\AN3.xml.filtered$39293
C:/Users/Ritik/Desktop/2021 Autumn Semester/Assignments/AI/filtered_data/Test-corpus\AN\AN4.xml.filtered$53737
C:/Users/Ritik/Desktop/2021 Autumn Semester/Assignments/AI/filtered_data/Test-corpus\AN\AN5.xml.filtered$37421
C:/Users/Ritik/Desktop/2021 Autumn Semester/Assignments/AI/filtered_data/Test-corpus\AN\AN7.xml.filtered$38383
C:/Users/Ritik/Desktop/2021 Autumn Semester/Assignments/AI/filtered_data/Test-corpus\AN\AN8.xml.filtered$35287
C:/Users/Ritik/Desktop/2021 Autumn Semester/Assignments/AI/filtered_data/Test-corpus\AN\AN9.xml.filtered$40765
C:/Users/Ritik/Desktop/2021 Autumn Semester/Assignments/AI/filtered_data/Test-corpus\AN\ANA.xml.filtered$30975
C:/Users/Ritik/Desktop/2021 Autumn Semester/Assignments/AI/filtered_data/Test-corpus\AN\ANB.xml.filtered$31242
C:/Users/Ritik/Desktop/2021 Autumn Semester/Assignments/AI/filtered_data/Test-corpus\AN\ANC.xml.filtered$37911
C:/Users/Ritik/Desktop/2021 Autumn Semester/Assignments/AI/filtered_data/Test-corpus\AN\AND.xml.filtered$34872
C:/Users/Ritik/Desktop/2021 Autumn Semester/Assignments/AI/filtered_data/Test-corpus\AN\ANF.xml.filtered$32813
C:/Users/Ritik/Desktop/2021 Autumn Semester/Assignments/AI/filtered_data/Test-corpus\AN\ANH.xml.filtered$35106
C:/Users/Ritik/Desktop/2021 Autumn Semester/Assignments/AI/filtered_data/Test-corpus\AN\ANJ.xml.filtered$14256
C:/Users/Ritik/Desktop/2021 Autumn Semester/Assignments/AI/filtered_data/Test-corpus\AN\ANK.xml.filtered$35772
```

Figure 15: Filtered Data Verification File

statistics - Notepad

File Edit Format View Help

```
Estimated Raw Tags:12744948
Processed Raw Tags:12744946
Verification Ratio:0.999998430750757
Accuracy:99.99998430750757
```

Figure 16: Statistics File

## 1.6 Results

The raw data was processed with an estimated accuracy of more than 99.99%. We could have made the data more accurate (to exactly 100%) by improving the regex for scanning raw files, but it would simply not be worth the trouble of investigating the patterns in the huge amount of data, only to correct an error of  $1.6 \times 10^{-5}\%$  or less, which would hardly affect the accuracy of the application. This data is now ready to be converted into a frequency dictionary in week 2.