COMP0660 (Malware)

Similarity Measures Coursework Report

Part 1 Experiment Fundamentals

1.1 Core Theory

Normalized Compression Distance (NCD) is used to detect similarities among strings. Based on this method, we can tell how similar two files are by calculating the NCD value between them. NCD value ranges from 0 (means similar) to 1(means different).

1.2 Environment

Operating System: Mac OS X

Programming language: Python 2.7

Main Python libraries used: *lzma*, *os*, *pandas*, *matplotlib.pyplot*

1.3 Chosen Compressor

7zip is a normal compressor holding the conditions of idempotency, monotonicity, symmetry and distributivity. Furthermore, it performs better than other compressors: more accurate behavior for self comparisons and lower compressed length on average.

7zip uses *lzma* as the compression algorithm so *lzma* also performs well in NCD-related experiments. Combined with the fact that *lzma* is supported by Python library *lzma* which is easy to apply in NCD calculation algorithm, *lzma* is chosen to be the compressor in this experiment.

Part 2 Algorithm Abstract

2.1 NCD Calculation Function

According to the definition of NCD, we need to calculate the file size of X, Y and XY after compression.

It can be simply implemented to get the concatenation of X and Y by using *X+Y* in Python.

To compress files, we just need to call function *lzma.compress ()*.

To get the file size, we only have to call function *len()*.

We keep four digits after the decimal point for the NCD value.

2.2 Import Files Based on *os.walk*

As can be seen, names of the target files are complicated and the number of files are not small. Therefore, an automated way of importing files is required. It can be simply implemented by using *os.walk* to traverse the root containing all the target files. After this, *.bin* files are selected and stored in an array.

2.3 Output and Store Results

In this experiment, we have 28 files in total (20 group files and 8 class files). By doing 28*28 = 784 calculations of NCD, we can get the NCD value between any two files.

The results would be a two-dimensional array whose size is 28*28. We can transfer the data structure from array to *dataframe*, which would be a great help for the following work of data visualization.

2.4 Classification of Group Files

Since we have the NCD value between any two files, we can simply check each group file for the NCD values between it and all the class files. We noticed that there exists a NCD value which is very close to 0 while others are close to 1. Therefore, we found that the group file is similar to this class file, which is the classification result.

2.5 Similarity Matrix of All Files

The similarity matrix can be generated by functions provided by library *matplotlib.pyplot*, taking the *dataframe* mentioned in 2.3 as input.

Part 3 Experiment Results

3.1 Table of All NCD Values

	0	1	2	3	...	24	25	26	27
0	0.0005	1.0001	0.9979	1.0011	...	0.9951	0.0604	0.9911	0.9848
1	0.9920	0.0003	0.9994	0.9999	...	0.9661	0.9998	0.0359	0.9978
2	0.9907	1.0002	0.0010	0.9978	...	0.9990	0.9953	0.9988	0.9912
3	0.9953	0.9994	0.9988	0.0008	...	0.9994	0.9969	0.9992	0.9981
4	0.0047	1.0009	0.9977	1.0008	...	0.9982	0.0527	0.9896	0.9841
5	0.9960	0.9641	0.9994	0.9999	...	0.0373	1.0015	0.9659	0.9980
6	0.9935	0.9846	0.9996	0.9999	...	0.9831	1.0024	0.9828	1.0015
7	0.9752	1.0069	0.9951	1.0002	...	0.9848	0.9840	1.0006	0.1492
8	0.9729	1.0034	0.9950	1.0004	...	0.9905	0.9824	1.0010	0.1513
9	0.9928	1.0011	0.9983	0.9998	...	0.9987	1.0009	0.9987	0.9957
10	0.9717	1.0051	0.9955	1.0005	...	0.9959	0.9843	0.9996	0.1516
11	0.9933	0.9998	1.0008	0.0107	...	0.9995	0.9978	0.9968	0.9990
12	0.9928	1.0006	1.0006	0.0103	...	0.9994	0.9955	0.9992	0.9954
13	0.9943	0.9998	0.9992	0.0091	...	0.9995	0.9983	0.9993	0.9965
14	0.9950	1.0007	0.9985	0.9999	...	0.9987	1.0006	0.9987	0.9971
15	0.9913	1.0012	0.0238	0.9983	...	0.9992	0.9956	0.9990	0.9936
16	0.9939	0.9642	0.9994	0.9997	...	0.0360	1.0001	0.9657	0.9982
17	0.9902	0.9842	0.9995	0.9998	...	0.9836	1.0020	0.9833	0.9978
18	0.9929	0.9995	0.0236	0.9980	...	0.9990	0.9948	0.9989	0.9939
19	0.9898	1.0004	0.0217	0.9972	...	0.9991	0.9968	0.9988	0.9916
20	0.9885	1.0006	0.0221	0.9994	...	0.9990	0.9981	0.9988	0.9930
21	1.0011	0.9860	1.0004	1.0003	...	0.9858	0.9996	0.9855	0.9972
22	0.9945	1.0019	0.9983	1.0000	...	0.9989	1.0016	0.9988	0.9982
23	0.9891	1.0001	1.0010	0.0092	...	0.9994	1.0021	0.9993	1.0028
24	0.9987	0.9656	0.9998	0.9998	...	0.0003	1.0011	0.9269	0.9982
25	0.0047	0.9995	0.9969	1.0003	...	0.9988	0.0005	0.9668	0.9824
26	0.9994	0.0165	1.0003	1.0000	...	0.9361	1.0029	0.0003	0.9971
27	0.9697	0.9984	0.9937	0.9998	...	0.9983	0.9852	0.9980	0.0006

No.0-27 respectively represent the following files:

No.0: 4abce6f575e9dd58cf2b131a1713ae91.bin

No.1: 28582df3f38139fffc6918341b49eadf.bin

No.2: 927fcccc5329d36fc168ef0e8fa4bbfd.bin

No.3: 790cb1adc16e0dd71ccc8a69a07a2622.bin

No.4: 7039057d0e801348c5f70fb98836c3af.bin

No.5: c29250c9b4052e62c0dfd65a224392c3.bin

No.6: 6ae6f97b54d0cd333952cde85f68c89b.bin

No.7: 597d8a63341acce1c1246c36df28dc7f.bin

No.8: 4637423926f9de7a46f5cdd7c7e071ea.bin

No.9: 43f53457e4618f46fb54fe31f9a95708.bin

No.10: 6db076a5cb45bf43333db0e63b764435.bin

No.11: 3d25c3c89af1a962fbf4ebc3eab6a1ef.bin

No.12: 851a10450da916fd66e92c2c24dbd711.bin

No.13: 9421a777248d5a2311084a0c73901300.bin

No.14: 7f98011c224ddb32b65f9e0b72e7a669.bin
No.15: 35a2bacbc84feca1293de0ee7784bc5d.bin
No.16: 0b75ad605500729a53ca13b1c2727ab1.bin
No.17: dc3d68be85ed5d49a36e5c5758fadfed.bin
No.18: 75829d9a403713c2b9f19155f4fb67f0.bin
No.19: 3596a3d7abc08b4f34239ac77916e973.bin
No.20: class8.bin
No.21: class1.bin
No.22: class2.bin
No.23: class3.bin
No.24: class7.bin
No.25: class6.bin
No.26: class4.bin
No.27: class5.bin

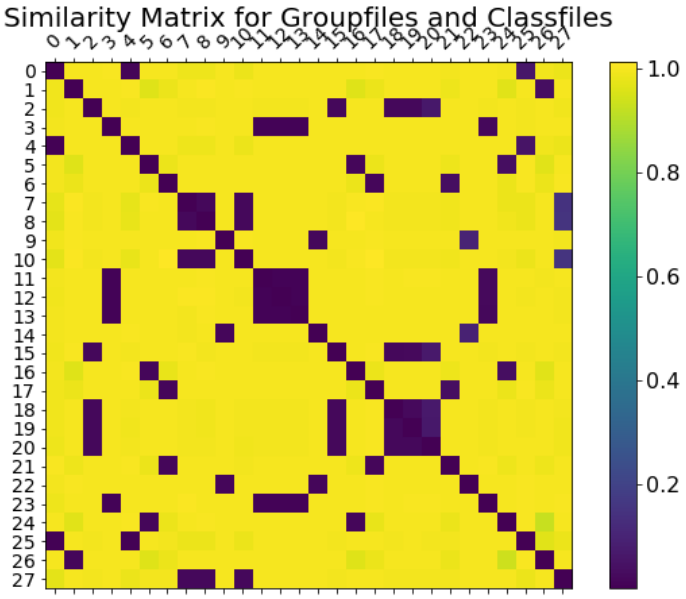
Any single value in this table means the NCD value between the file represented by row No. and the file represented by column No.

*Please refer to the file ***NCD_value.csv*** to view the whole table.*

3.2 Classification Results

```
4abce6f575e9dd58cf2b131a1713ae91.bin is most similar to class6.bin
28582df3f38139fffc6918341b49eadf.bin is most similar to class4.bin
927fcccc5329d36fc168ef0e8fa4bbfd.bin is most similar to class8.bin
790cb1adc16e0dd71ccc8a69a07a2622.bin is most similar to class3.bin
7039057d0e801348c5f70fb98836c3af.bin is most similar to class6.bin
c29250c9b4052e62c0dfd65a224392c3.bin is most similar to class7.bin
6ae6f97b54d0cd333952cde85f68c89b.bin is most similar to class1.bin
597d8a63341acce1c1246c36df28dc7f.bin is most similar to class5.bin
4637423926f9de7a46f5cdd7c7e071ea.bin is most similar to class5.bin
43f53457e4618f46fb54fe31f9a95708.bin is most similar to class2.bin
6db076a5cb45bf43333db0e63b764435.bin is most similar to class5.bin
3d25c3c89af1a962fbf4ebc3eab6a1ef.bin is most similar to class3.bin
851a10450da916fd66e92c2c24dbd711.bin is most similar to class3.bin
9421a777248d5a2311084a0c73901300.bin is most similar to class3.bin
7f98011c224ddb32b65f9e0b72e7a669.bin is most similar to class2.bin
35a2bacbc84feca1293de0ee7784bc5d.bin is most similar to class8.bin
0b75ad605500729a53ca13b1c2727ab1.bin is most similar to class7.bin
dc3d68be85ed5d49a36e5c5758fadfed.bin is most similar to class1.bin
75829d9a403713c2b9f19155f4fb67f0.bin is most similar to class8.bin
3596a3d7abc08b4f34239ac77916e973.bin is most similar to class8.bin
```

3.3 Similarity Matrix



No.0-27 represent the same files as those in the NCD value table.

*Please refer to the file ***similarity_matrix.py*** to view the whole python code.*