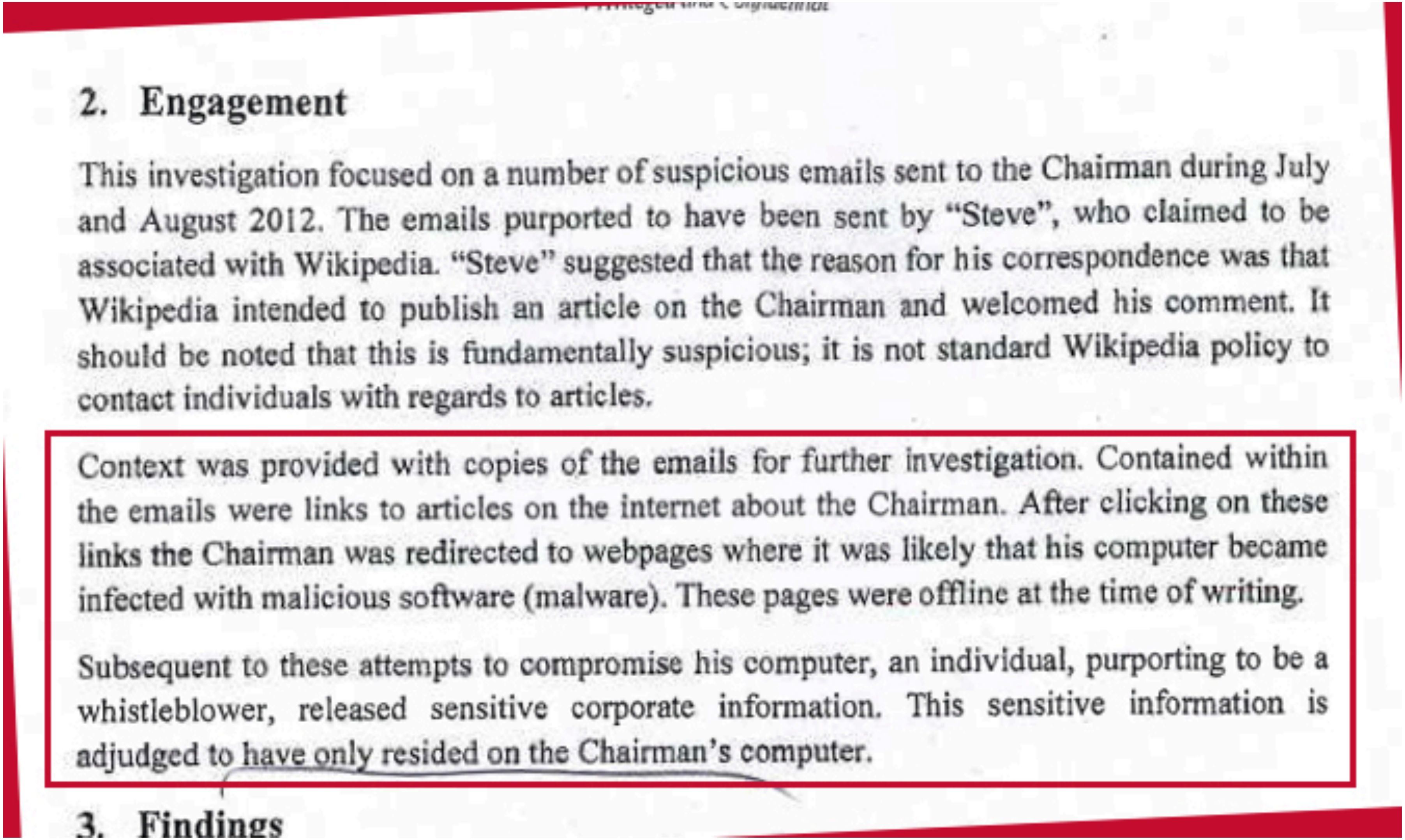
Increasing Security Awareness in Enterprise Using Automated Feature Extraction with Long n-gram Analysis for Filetype Identification

Master Thesis
Burman Noviansyah
MSISPM ’14

December 10th, 2013

Carnegie Mellon
Information Security Policy & Management | Heinz College

Context Information Security: Project MASON – Incident Response.



2. Engagement

This investigation focused on a number of suspicious emails sent to the Chairman during July and August 2012. The emails purported to have been sent by "Steve", who claimed to be associated with Wikipedia. "Steve" suggested that the reason for his correspondence was that Wikipedia intended to publish an article on the Chairman and welcomed his comment. It should be noted that this is fundamentally suspicious; it is not standard Wikipedia policy to contact individuals with regards to articles.

Context was provided with copies of the emails for further investigation. Contained within the emails were links to articles on the internet about the Chairman. After clicking on these links the Chairman was redirected to webpages where it was likely that his computer became infected with malicious software (malware). These pages were offline at the time of writing.

Subsequent to these attempts to compromise his computer, an individual, purporting to be a whistleblower, released sensitive corporate information. This sensitive information is adjudged to have only resided on the Chairman's computer.

3. Findings

Source:
http://www.thetimes.co.uk/tto/multimedia/archive/00372/DOC100113-100120132_372895a.pdf

Problem in Enterprise

- Prevent IPR or confidential documents left the network
- Policy revisited: Data Loss Prevention. Commercial Product of the shelves only use simple recognitions methods to detect file type, that is file extension.
- Anti-forensics case: altering file header and footer, file container to conceal true file type
- File Type Identification (FTI) is a challenging task

File Type Identification (FTI)

Previous Works on FTI:

- Libmagic (Darwin, 1973)
- File Extension (Hiller, 1996)
- Byte Frequency Analysis, Byte Frequency Cross-Correlation, File Header/Trailer Algorithm (McDaniel, 2001)
- Fileprints with 1-gram Analysis (Li, Wang and Stolfo)
- Long summarized n-gram (Mayer)

N-Gram as Features in FTI

- Introduced in (Damashek, 1995) as a subsequence of n consecutive tokens in stream of tokens.
- N-gram in FTI: 1 bytes x n
- Feature: Specific and common n-gram among same file type, and unique between all other file type. This feature can be choose to a identify file type
- Location of feature is anywhere inside the file, not necessarily only in header or footer
- Can occurred once in file, or even higher frequency of occurrences in one single file.

Proposed Solution

Characteristics

- Doesn’t interfere the extraction process to create predictability flag (requires prior knowledge towards file structure)
- Doesn’t summarized (to see exact content without modifying any byte)
- Doesn’t eliminate short n-gram that as part as longer n-gram (the position of short n-gram is unknown, whether a subset of longer n-gram, or indeed stand alone short n-gram)
- Training set was chosen specifically so that only pristine file types were processed
- Prevent the similar and identic file processed more than once.
- Same data set for prediction rate calculation

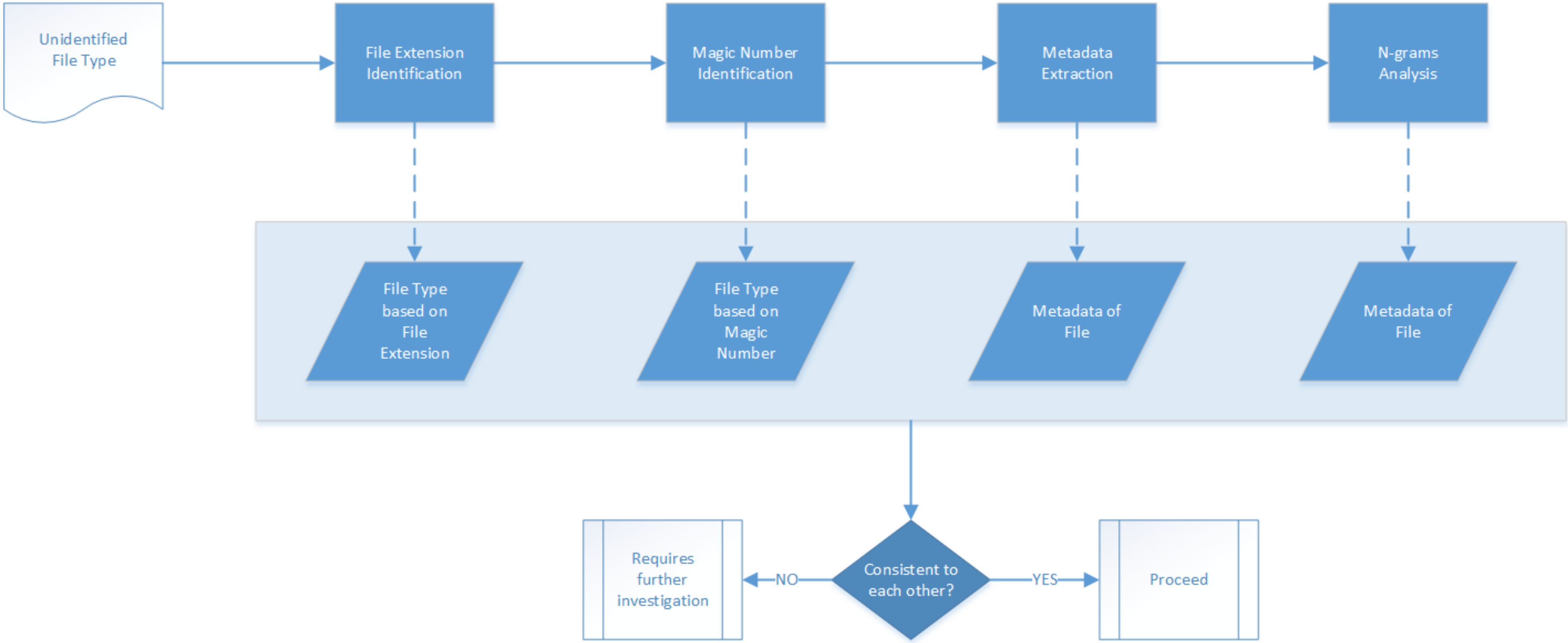
Proposed Solution

Combination of:

- File Extension
- Libmagic 'file' linux command
- 'exiftool' linux command
- N-gram analysis

Outcome: consistency check between all of results from those four techniques

Proposed Solution

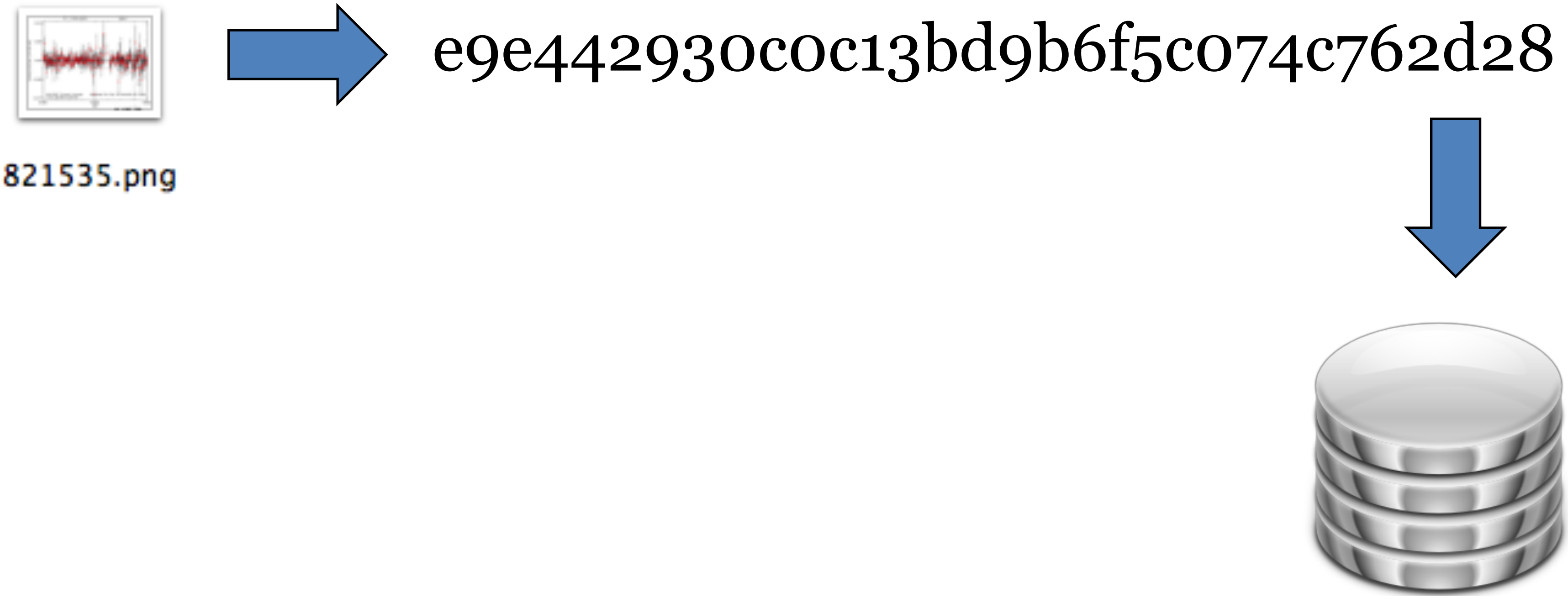

Unidentified File Type
File Extension Identification
Magic Number Identification
Metadata Extraction
N-grams Analysis
File Type based on File Extension
File Type based on Magic Number
Metadata of File
Metadata of File
Consistent to each other?
NO
YES
Requires further investigation
Proceed

Algorithm: Learning

- Determine the maximum size of n-gram will be extracted from file. In this thesis, the author chooses the maximum size is 20-gram.
- For each n-gram size, repeat these steps:
  - Create sliding window with the size of n, started with value of n = 1.
  - Copy the gram inside this current window to memory object as *extracted n-grams*. If this gram has not been stored in memory object, then simply put it on memory object. Otherwise, proceed to next step 2c. This step guarantees that memory object only contains *unique extracted n-grams* that exists on this current file and not repeated between other n-gram.
  - Slide the window 1 byte away, then repeat the step 2b above until sliding window reach the end of the file.
  - Write all *unique extracted n-grams* in memory object to database, along with its file. Until this step, the database contains all *unique extracted n-grams* data from size n = 1 up until n <= 20.
  - Increase the size of n with 1, and then repeat from step 2a until the maximum size of n is reached.

Algorithm: Learning

Check whether to avoid re-learn the identic file using md5 hash value

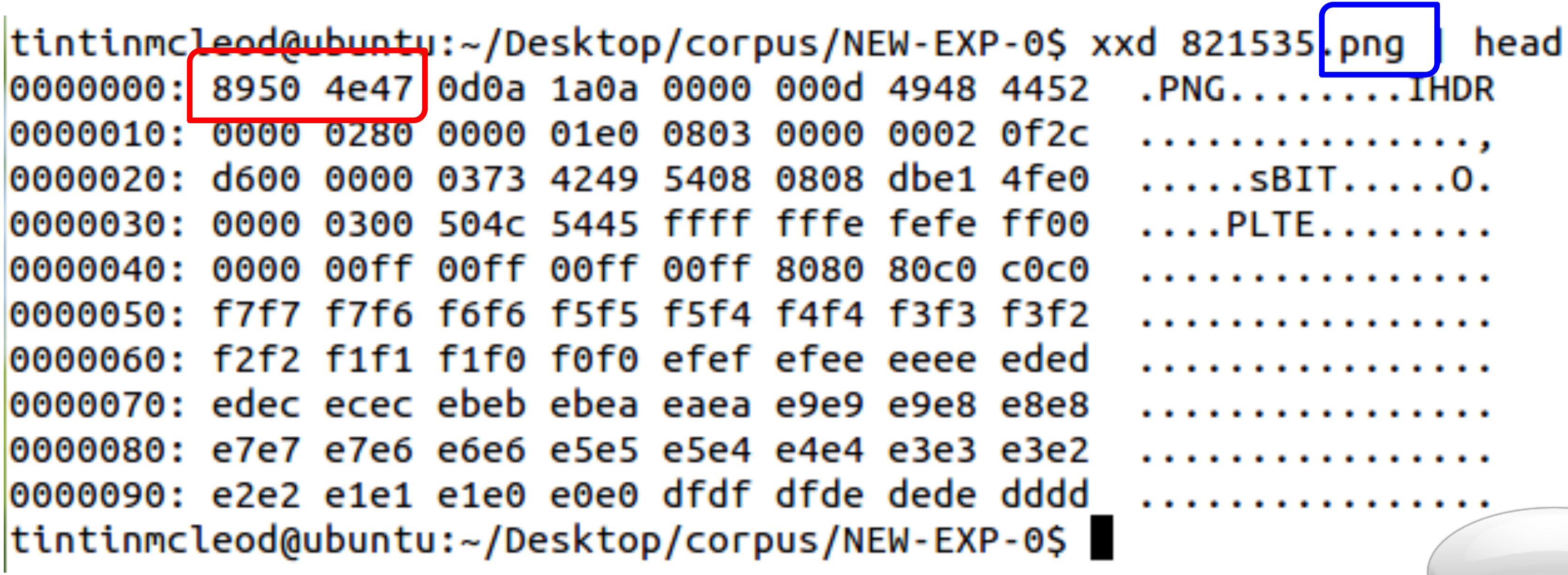


821535.png

e9e442930c0c13bd9b6f5c074c762d28

Algorithm: Learning

Sliding window of n-size to extract the content

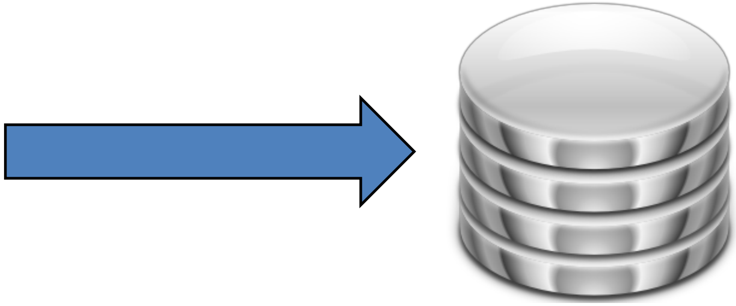


```
tintinmcleod@ubuntu:~/Desktop/corpus/NEW-EXP-0$ xxd 821535.png | head
0000000: 8950 4e47 0d0a 1a0a 0000 000d 4948 4452  .PNG........IHDR
0000010: 0000 0280 0000 01e0 0803 0000 0002 0f2c  ...............,
0000020: d600 0000 0373 4249 5408 0808 dbe1 4fe0  .....sBIT.....O.
0000030: 0000 0300 504c 5445 ffff fffe fefe ff00  ....PLTE........
0000040: 0000 00ff 00ff 00ff 00ff 8080 80c0 c0c0  ................
0000050: f7f7 f7f6 f6f6 f5f5 f5f4 f4f4 f3f3 f3f2  ................
0000060: f2f2 f1f1 f1f0 f0f0 efef efee eeee eded  ................
0000070: edec ecec ebeb ebea eaea e9e9 e9e8 e8e8  ................
0000080: e7e7 e7e6 e6e6 e5e5 e5e4 e4e4 e3e3 e3e2  ................
0000090: e2e2 e1e1 e1e0 e0e0 dfdf dfde dede dddd  ................
tintinmcleod@ubuntu:~/Desktop/corpus/NEW-EXP-0$
```

```
89504E47 for PNG
504E470D for PNG
4E470D0A for PNG
```

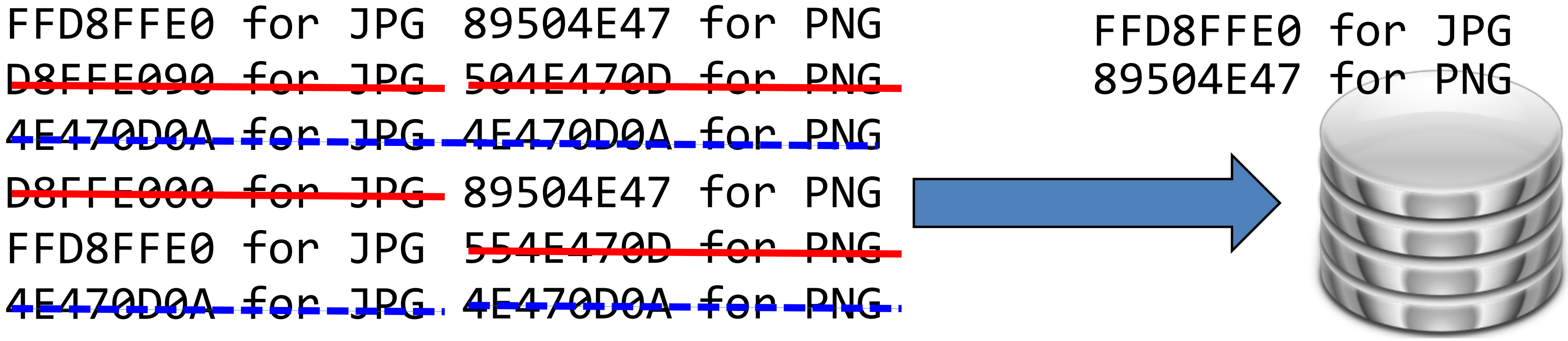
Algorithm: Learning

- Repeat this process for all size of sliding windows that represent n-gram
- Repeat this process for all files in learning data set



Algorithm: Learning

- Find all same n-gram size that belongs to same file type → common n-gram
- Find all common n-gram that only belongs to one file type → features



FFD8FFE0 for JPG 89504E47 for PNG

~~D8FFE090 for JPG~~ ~~504E470D for PNG~~

~~4E470D0A for JPG 4E470D0A for PNG~~

~~D8FFE000 for JPG~~ 89504E47 for PNG

FFD8FFE0 for JPG ~~554E470D for PNG~~

~~4E470D0A for JPG~~ ~~4E470D0A for PNG~~

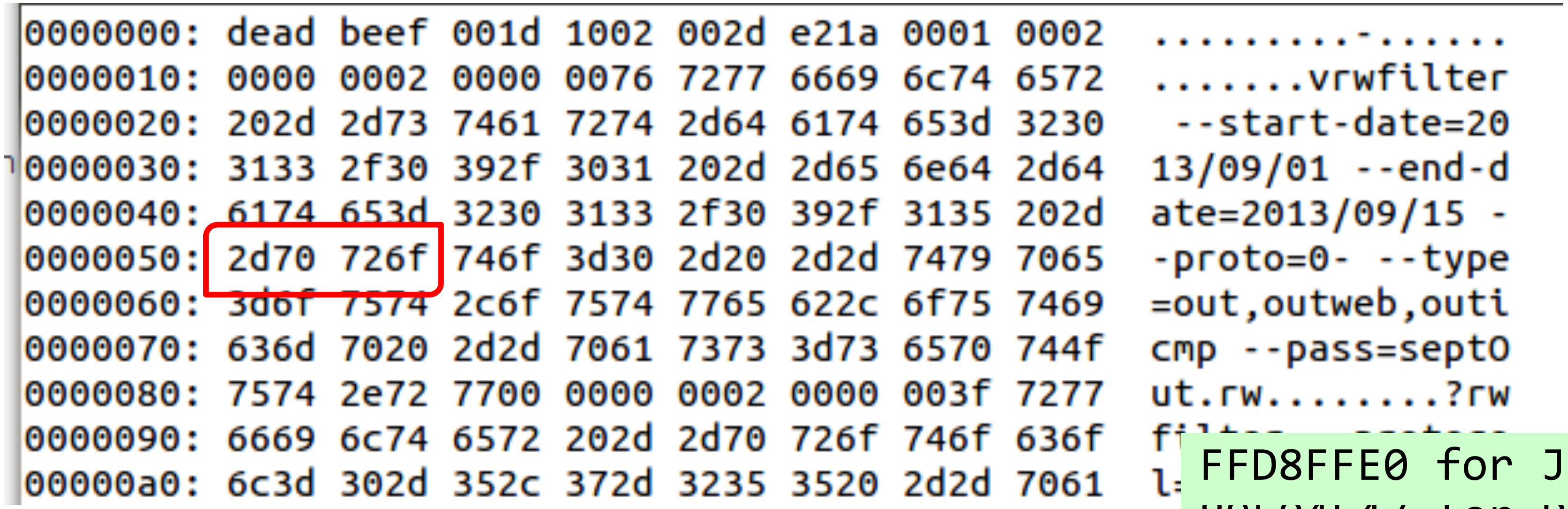
FFD8FFE0 for JPG
89504E47 for PNG

Algorithm: Prediction

- Determine the maximum size of n-gram will be compared with file. In this thesis, the author chooses the maximum size is 20-gram similar with Extraction process.
- For each n-gram size, repeat these steps:
  - Get all *final features* from the database based on n-gram size, regardless what file type associated with it, along with its *total final features counter* for each file type.
  - For each file type in *final features* with current size of n:
    - Create a *final features statistics container* to hold statistics of all matched *final features* per file type. This container will be used to count how many *final features* in occurred in *array of bytes*. This *final features statistics container* contains [file extension, size of n, *total final features counter*, matched *final features counter*].
    - Repeat process 2.b.i until all of file types have each different *final features statistics container*.
  - Create sliding window with size is current n.
  - Check whether this current gram in *final features*. If current gram exists in *final features* then updates its statistics in *final features statistics container*. Otherwise, just skip it.
  - Slide the window 1 byte away from the original.
  - Repeat from 2.d until the window reaches the end of file.
- Calculate the percentage match based on value in *final features statistics container*.
- Determine the acceptance threshold toward percentage match so that this matching feature can be used to determine whether this is accepted as specific file type. For this thesis, author set the threshold of acceptance is 95%. Hence, if a set of features in size of n for file type X matches more than 95% of total features, than this file is classified as file type X.

Algorithm: Prediction

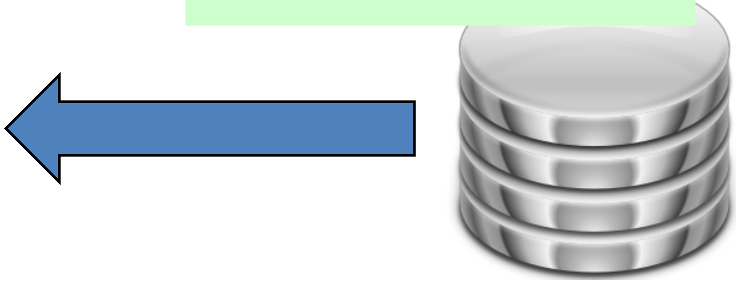
Sliding window of n-size to extract the content



```
0000000: dead beef 001d 1002 002d e21a 0001 0002  .........-......
0000010: 0000 0002 0000 0076 7277 6669 6c74 6572  .......vrwfilter
0000020: 202d 2d73 7461 7274 2d64 6174 653d 3230   --start-date=20
0000030: 3133 2f30 392f 3031 202d 2d65 6e64 2d64  13/09/01 --end-d
0000040: 6174 653d 3230 3133 2f30 392f 3135 202d  ate=2013/09/15 -
0000050: 2d70 726f 746f 3d30 2d20 2d2d 7479 7065  -proto=0- --type
0000060: 3d6f 7574 2c6f 7574 7765 622c 6f75 7469  =out,outweb,outi
0000070: 636d 7020 2d2d 7061 7373 3d73 6570 744f  cmp --pass=septO
0000080: 7574 2e72 7700 0000 0002 0000 003f 7277  ut.rw........?rw
0000090: 6669 6c74 6572 202d 2d70 726f 746f 636f  fi
00000a0: 6c3d 302d 352c 372d 3235 3520 2d2d 7061  l=
```

FFD8FFE0 for JPG
89504E47 for PNG

If it is match with any feature, count for matching rate



Data Set

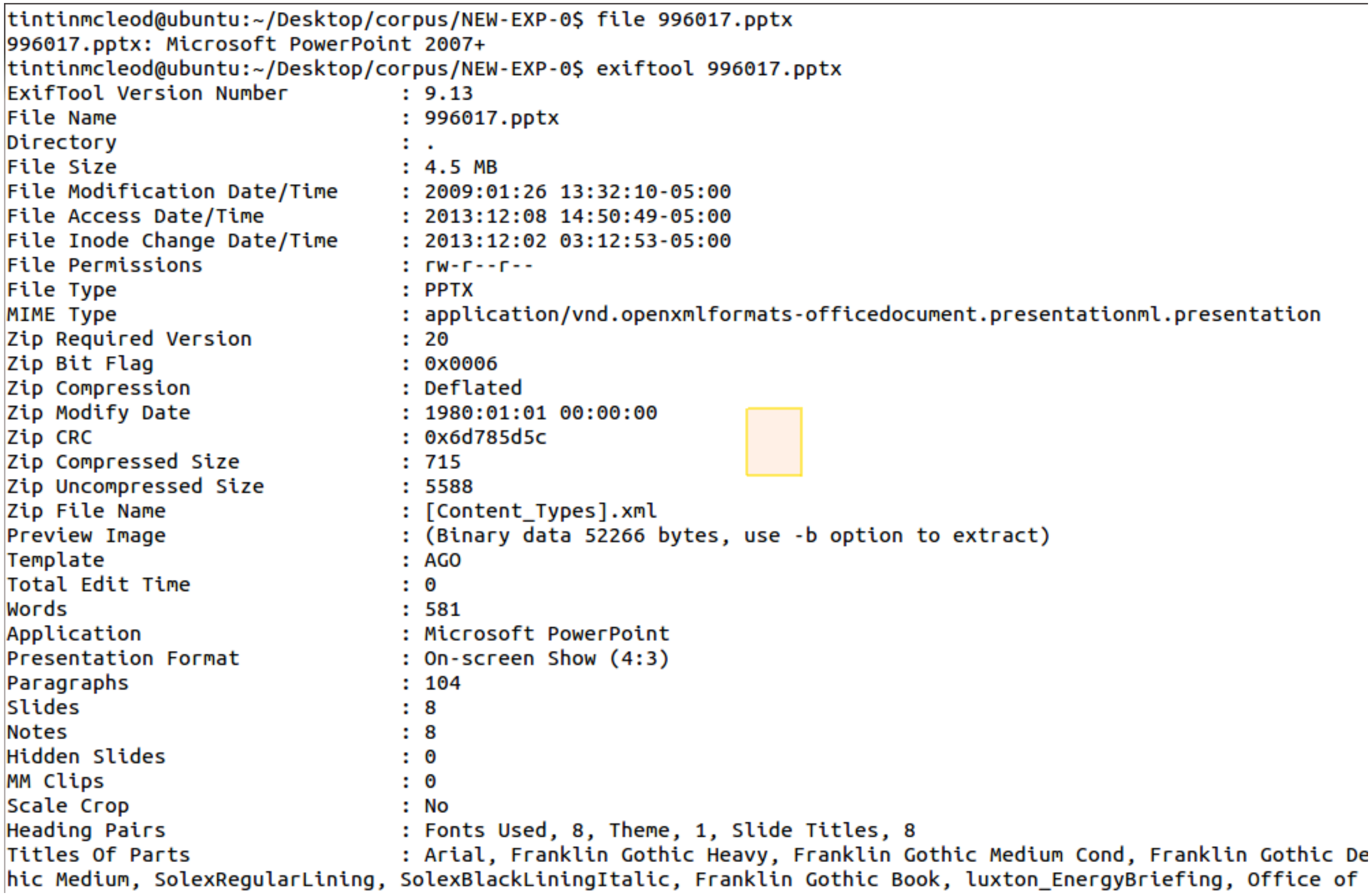
Learning Data Set:

- Standard Corpus from (Garfinkel, Farrell and Roussev)
- Limit to 11 data types with 20 files each

Prediction Data Set:

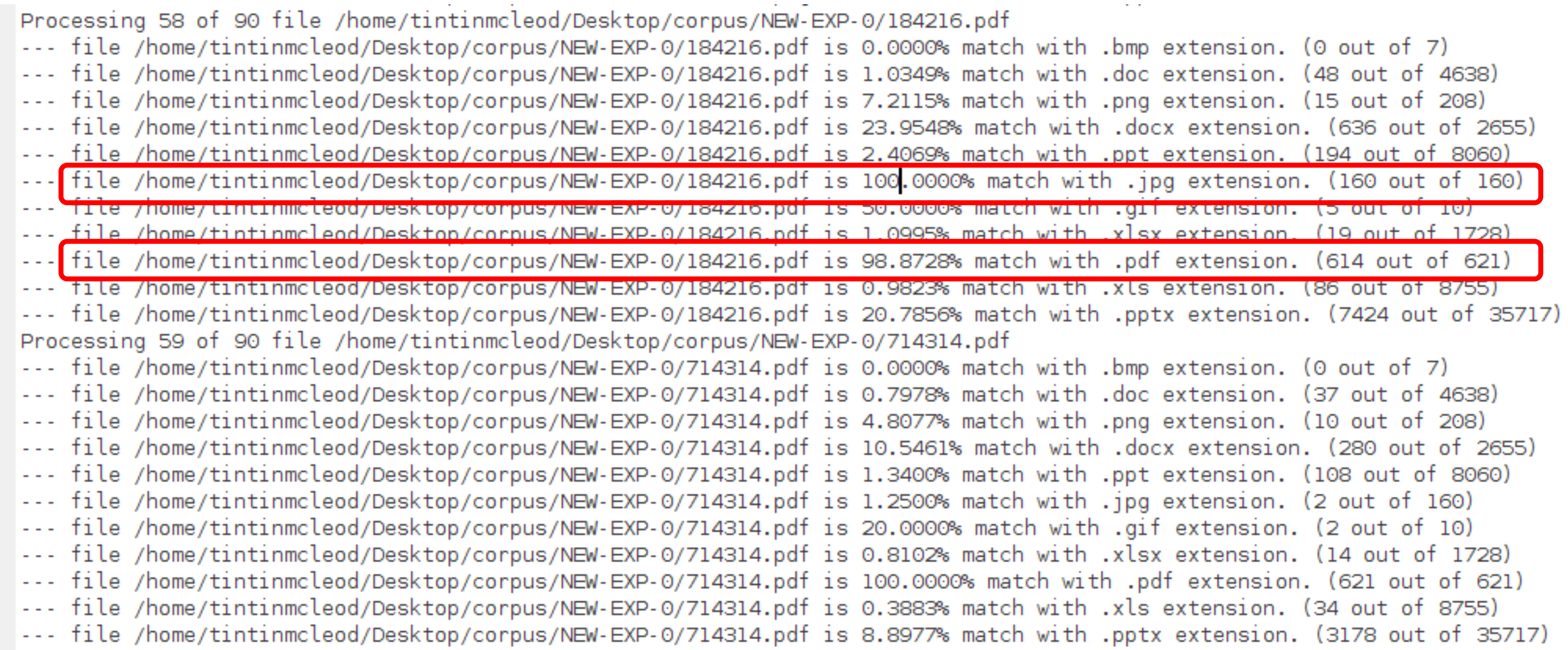
- Axelsson's NEW-EXP-0
- Limit to 11 data types

Result: ‘file’, ‘exiftool’



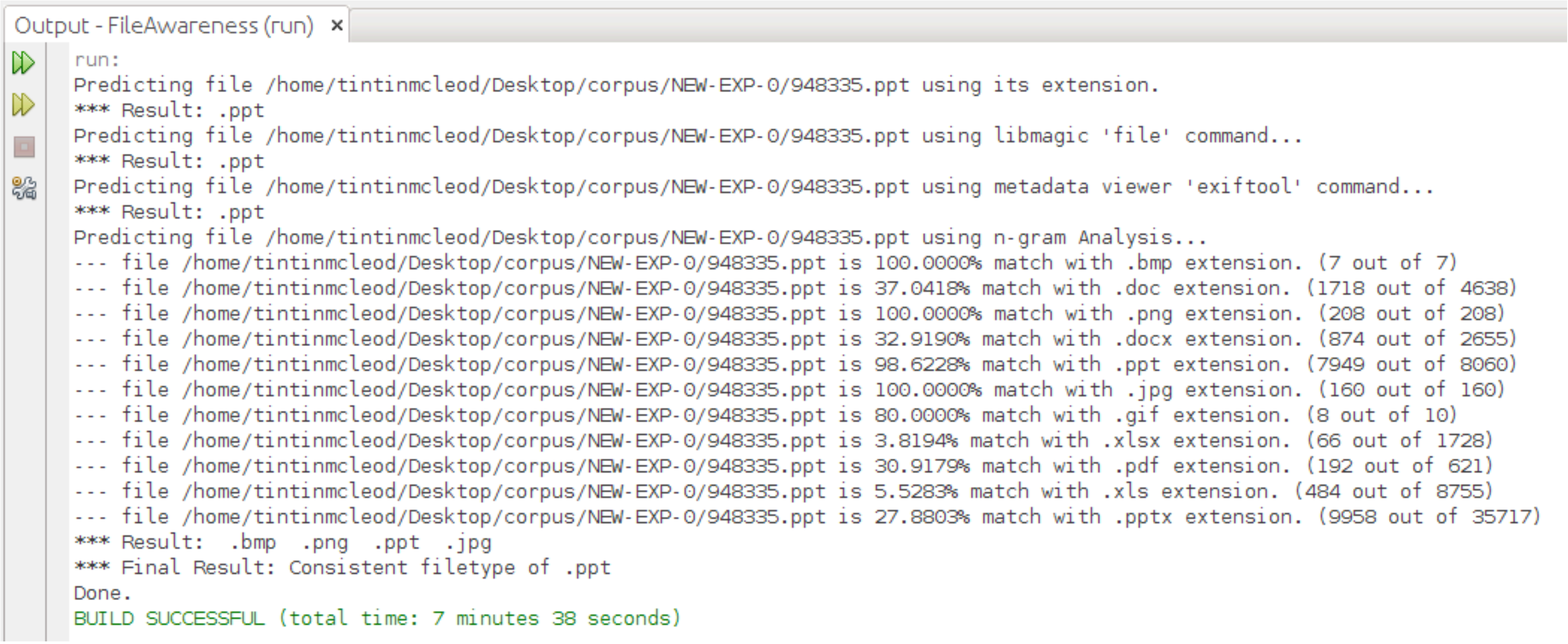
```
tintinmcleod@ubuntu:~/Desktop/corpus/NEW-EXP-0$ file 996017.pptx
996017.pptx: Microsoft PowerPoint 2007+
tintinmcleod@ubuntu:~/Desktop/corpus/NEW-EXP-0$ exiftool 996017.pptx
ExifTool Version Number         : 9.13
File Name                       : 996017.pptx
Directory                       : .
File Size                       : 4.5 MB
File Modification Date/Time     : 2009:01:26 13:32:10-05:00
File Access Date/Time           : 2013:12:08 14:50:49-05:00
File Inode Change Date/Time     : 2013:12:02 03:12:53-05:00
File Permissions                : rw-r--r--
File Type                       : PPTX
MIME Type                       : application/vnd.openxmlformats-officedocument.presentationml.presentation
Zip Required Version            : 20
Zip Bit Flag                    : 0x0006
Zip Compression                 : Deflated
Zip Modify Date                 : 1980:01:01 00:00:00
Zip CRC                         : 0x6d785d5c
Zip Compressed Size             : 715
Zip Uncompressed Size           : 5588
Zip File Name                   : [Content_Types].xml
Preview Image                   : (Binary data 52266 bytes, use -b option to extract)
Template                        : AGO
Total Edit Time                 : 0
Words                           : 581
Application                     : Microsoft PowerPoint
Presentation Format             : On-screen Show (4:3)
Paragraphs                      : 104
Slides                          : 8
Notes                           : 8
Hidden Slides                   : 0
MM Clips                        : 0
Scale Crop                      : No
Heading Pairs                   : Fonts Used, 8, Theme, 1, Slide Titles, 8
Titles Of Parts                 : Arial, Franklin Gothic Heavy, Franklin Gothic Medium Cond, Franklin Gothic De
hic Medium, SolexRegularLining, SolexBlackLiningItalic, Franklin Gothic Book, luxton_EnergyBriefing, Office of
```

Result: n-gram



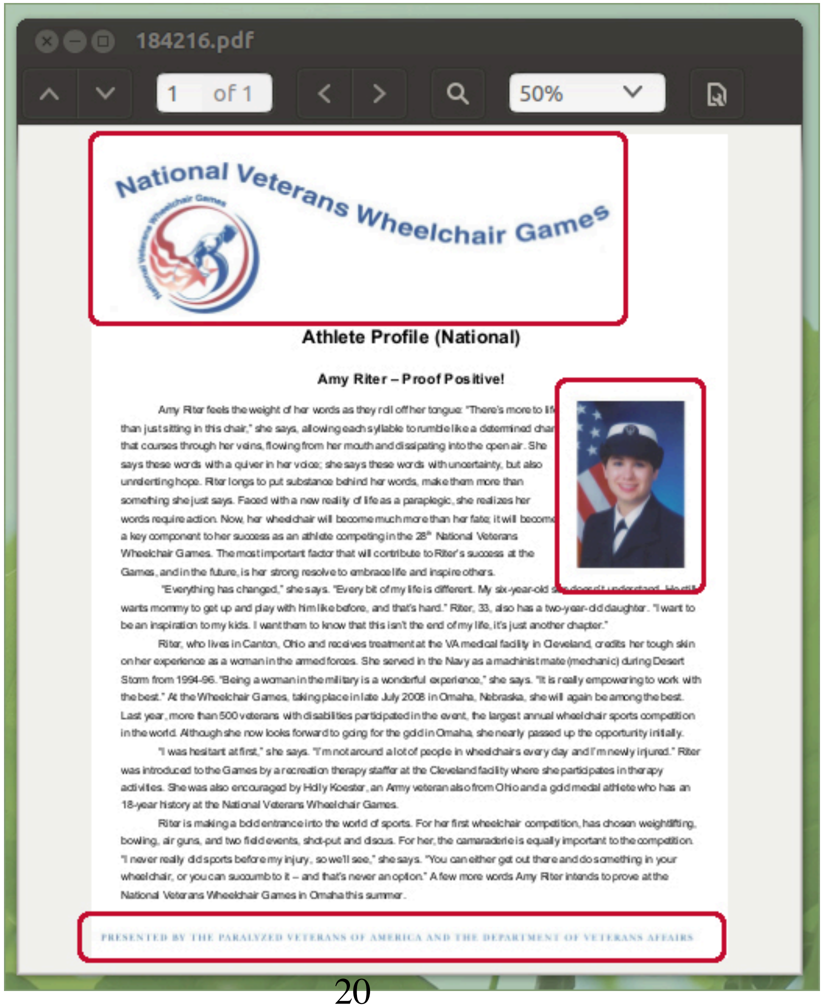
```
Processing 58 of 90 file /home/tintinmcleod/Desktop/corpus/NEW-EXP-0/184216.pdf
--- file /home/tintinmcleod/Desktop/corpus/NEW-EXP-0/184216.pdf is 0.0000% match with .bmp extension. (0 out of 7)
--- file /home/tintinmcleod/Desktop/corpus/NEW-EXP-0/184216.pdf is 1.0349% match with .doc extension. (48 out of 4638)
--- file /home/tintinmcleod/Desktop/corpus/NEW-EXP-0/184216.pdf is 7.2115% match with .png extension. (15 out of 208)
--- file /home/tintinmcleod/Desktop/corpus/NEW-EXP-0/184216.pdf is 23.9548% match with .docx extension. (636 out of 2655)
--- file /home/tintinmcleod/Desktop/corpus/NEW-EXP-0/184216.pdf is 2.4069% match with .ppt extension. (194 out of 8060)
--- file /home/tintinmcleod/Desktop/corpus/NEW-EXP-0/184216.pdf is 100.0000% match with .jpg extension. (160 out of 160)
--- file /home/tintinmcleod/Desktop/corpus/NEW-EXP-0/184216.pdf is 50.0000% match with .gif extension. (5 out of 10)
--- file /home/tintinmcleod/Desktop/corpus/NEW-EXP-0/184216.pdf is 1.0995% match with .xlsx extension. (19 out of 1728)
--- file /home/tintinmcleod/Desktop/corpus/NEW-EXP-0/184216.pdf is 98.8728% match with .pdf extension. (614 out of 621)
--- file /home/tintinmcleod/Desktop/corpus/NEW-EXP-0/184216.pdf is 0.9823% match with .xls extension. (86 out of 8755)
--- file /home/tintinmcleod/Desktop/corpus/NEW-EXP-0/184216.pdf is 20.7856% match with .pptx extension. (7424 out of 35717)
Processing 59 of 90 file /home/tintinmcleod/Desktop/corpus/NEW-EXP-0/714314.pdf
--- file /home/tintinmcleod/Desktop/corpus/NEW-EXP-0/714314.pdf is 0.0000% match with .bmp extension. (0 out of 7)
--- file /home/tintinmcleod/Desktop/corpus/NEW-EXP-0/714314.pdf is 0.7978% match with .doc extension. (37 out of 4638)
--- file /home/tintinmcleod/Desktop/corpus/NEW-EXP-0/714314.pdf is 4.8077% match with .png extension. (10 out of 208)
--- file /home/tintinmcleod/Desktop/corpus/NEW-EXP-0/714314.pdf is 10.5461% match with .docx extension. (280 out of 2655)
--- file /home/tintinmcleod/Desktop/corpus/NEW-EXP-0/714314.pdf is 1.3400% match with .ppt extension. (108 out of 8060)
--- file /home/tintinmcleod/Desktop/corpus/NEW-EXP-0/714314.pdf is 1.2500% match with .jpg extension. (2 out of 160)
--- file /home/tintinmcleod/Desktop/corpus/NEW-EXP-0/714314.pdf is 20.0000% match with .gif extension. (2 out of 10)
--- file /home/tintinmcleod/Desktop/corpus/NEW-EXP-0/714314.pdf is 0.8102% match with .xlsx extension. (14 out of 1728)
--- file /home/tintinmcleod/Desktop/corpus/NEW-EXP-0/714314.pdf is 100.0000% match with .pdf extension. (621 out of 621)
--- file /home/tintinmcleod/Desktop/corpus/NEW-EXP-0/714314.pdf is 0.3883% match with .xls extension. (34 out of 8755)
--- file /home/tintinmcleod/Desktop/corpus/NEW-EXP-0/714314.pdf is 8.8977% match with .pptx extension. (3178 out of 35717)
```

Result: Aggregate Prediction



```
run:
Predicting file /home/tintinmcleod/Desktop/corpus/NEW-EXP-0/948335.ppt using its extension.
*** Result: .ppt
Predicting file /home/tintinmcleod/Desktop/corpus/NEW-EXP-0/948335.ppt using libmagic 'file' command...
*** Result: .ppt
Predicting file /home/tintinmcleod/Desktop/corpus/NEW-EXP-0/948335.ppt using metadata viewer 'exiftool' command...
*** Result: .ppt
Predicting file /home/tintinmcleod/Desktop/corpus/NEW-EXP-0/948335.ppt using n-gram Analysis...
--- file /home/tintinmcleod/Desktop/corpus/NEW-EXP-0/948335.ppt is 100.0000% match with .bmp extension. (7 out of 7)
--- file /home/tintinmcleod/Desktop/corpus/NEW-EXP-0/948335.ppt is 37.0418% match with .doc extension. (1718 out of 4638)
--- file /home/tintinmcleod/Desktop/corpus/NEW-EXP-0/948335.ppt is 100.0000% match with .png extension. (208 out of 208)
--- file /home/tintinmcleod/Desktop/corpus/NEW-EXP-0/948335.ppt is 32.9190% match with .docx extension. (874 out of 2655)
--- file /home/tintinmcleod/Desktop/corpus/NEW-EXP-0/948335.ppt is 98.6228% match with .ppt extension. (7949 out of 8060)
--- file /home/tintinmcleod/Desktop/corpus/NEW-EXP-0/948335.ppt is 100.0000% match with .jpg extension. (160 out of 160)
--- file /home/tintinmcleod/Desktop/corpus/NEW-EXP-0/948335.ppt is 80.0000% match with .gif extension. (8 out of 10)
--- file /home/tintinmcleod/Desktop/corpus/NEW-EXP-0/948335.ppt is 3.8194% match with .xlsx extension. (66 out of 1728)
--- file /home/tintinmcleod/Desktop/corpus/NEW-EXP-0/948335.ppt is 30.9179% match with .pdf extension. (192 out of 621)
--- file /home/tintinmcleod/Desktop/corpus/NEW-EXP-0/948335.ppt is 5.5283% match with .xls extension. (484 out of 8755)
--- file /home/tintinmcleod/Desktop/corpus/NEW-EXP-0/948335.ppt is 27.8803% match with .pptx extension. (9958 out of 35717)
*** Result:  .bmp  .png  .ppt  .jpg
*** Final Result: Consistent filetype of .ppt
Done.
BUILD SUCCESSFUL (total time: 7 minutes 38 seconds)
```

Analysis



Athlete Profile (National)

Amy Riter – Proof Positive!

Amy Riter feels the weight of her words as they roll off her tongue: "There's more to lif than just sitting in this chair," she says, allowing each syllable to rumble like a determined char that courses through her veins, flowing from her mouth and dissipating into the open air. She says these words with a quiver in her voice; she says these words with uncertainty, but also unrelenting hope. Riter longs to put substance behind her words, make them more than something she just says. Faced with a new reality of life as a paraplegic, she realizes her words require action. Now, her wheelchair will become much more than her fate; it will become a key component to her success as an athlete competing in the 28th National Veterans Wheelchair Games. The most important factor that will contribute to Riter's success at the Games, and in the future, is her strong resolve to embrace life and inspire others.

"Everything has changed," she says. "Every bit of my life is different. My six-year-old s doesn't understand. He still wants mommy to get up and play with him like before, and that's hard." Riter, 33, also has a two-year-old daughter. "I want to be an inspiration to my kids. I want them to know that this isn't the end of my life, it's just another chapter."

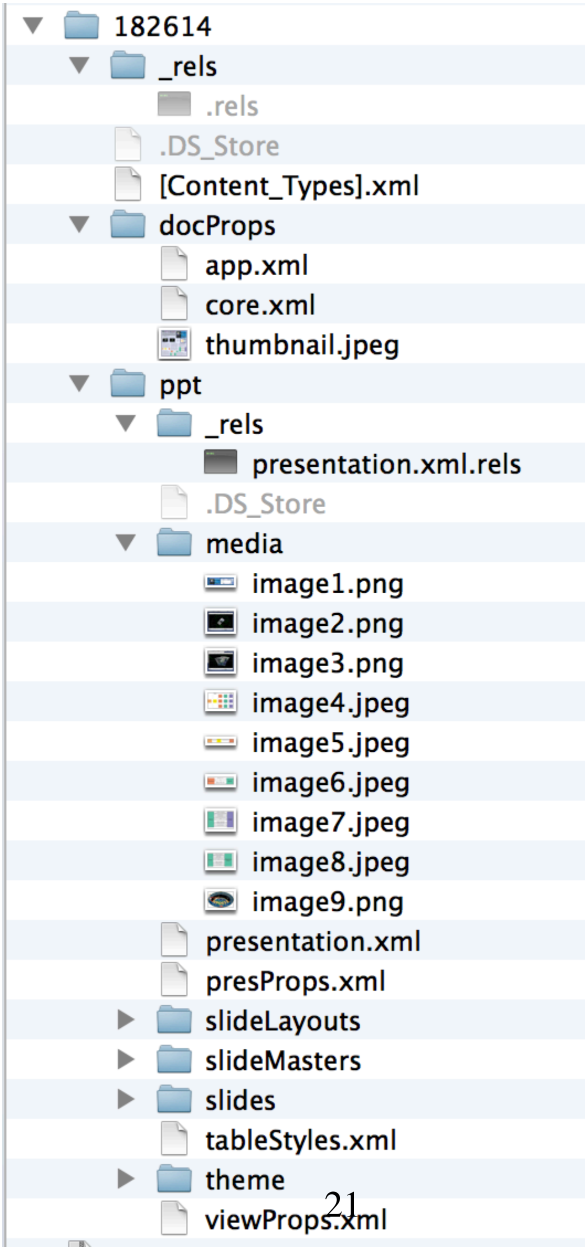
Riter, who lives in Canton, Ohio and receives treatment at the VA medical facility in Cleveland, credits her tough skin on her experience as a woman in the armed forces. She served in the Navy as a machinist mate (mechanic) during Desert Storm from 1994-96. "Being a woman in the military is a wonderful experience," she says. "It is really empowering to work with the best." At the Wheelchair Games, taking place in late July 2008 in Omaha, Nebraska, she will again be among the best. Last year, more than 500 veterans with disabilities participated in the event, the largest annual wheelchair sports competition in the world. Although she now looks forward to going for the gold in Omaha, she nearly passed up the opportunity initially.

"I was hesitant at first," she says. "I'm not around a lot of people in wheelchairs every day and I'm newly injured." Riter was introduced to the Games by a recreation therapy staffer at the Cleveland facility where she participates in therapy activities. She was also encouraged by Holly Koester, an Army veteran also from Ohio and a gold medal athlete who has an 18-year history at the National Veterans Wheelchair Games.

Riter is making a bold entrance into the world of sports. For her first wheelchair competition, has chosen weightlifting, bowling, air guns, and two field events, shot-put and discus. For her, the camaraderie is equally important to the competition. "I never really did sports before my injury, so we'll see," she says. "You can either get out there and do something in your wheelchair, or you can succumb to it – and that's never an option." A few more words Amy Riter intends to prove at the National Veterans Wheelchair Games in Omaha this summer.

PRESENTED BY THE PARALYZED VETERANS OF AMERICA AND THE DEPARTMENT OF VETERANS AFFAIRS

- A PDF file classified as PNG?
- This PDF contains several image element
- File Container and file inside container detected.



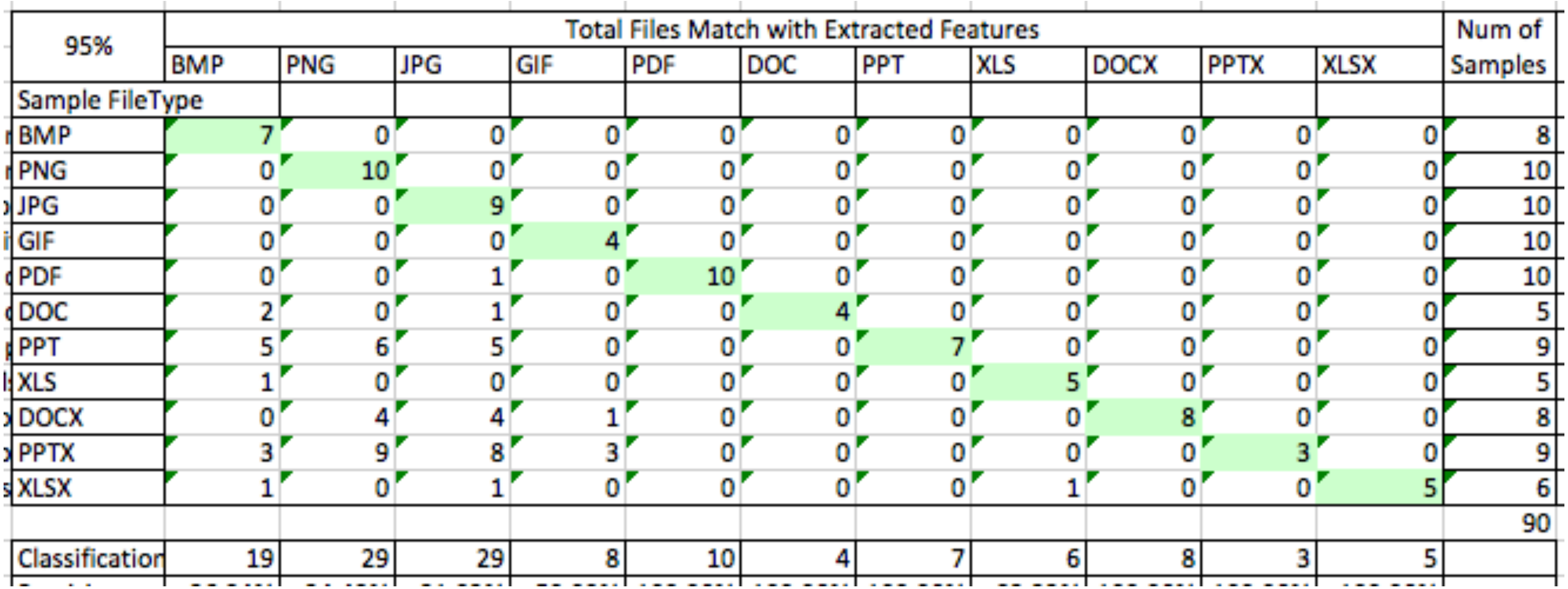
```
▼ 182614
    ▼ _rels
        .rels
      .DS_Store
      [Content_Types].xml
    ▼ docProps
        app.xml
        core.xml
        thumbnail.jpeg
    ▼ ppt
        ▼ _rels
            presentation.xml.rels
          .DS_Store
        ▼ media
            image1.png
            image2.png
            image3.png
            image4.jpeg
            image5.jpeg
            image6.jpeg
            image7.jpeg
            image8.jpeg
            image9.png
          presentation.xml
          presProps.xml
        ▶ slideLayouts
        ▶ slideMasters
        ▶ slides
          tableStyles.xml
        ▶ theme
          viewProps.xml
```

Analysis

- A DOCX file classified as PNG, JPG, and GIF?
- This DOCX contains several image element, too.
- File insides this container is not encoded

Analysis

Confusion Matrix for Testing Result

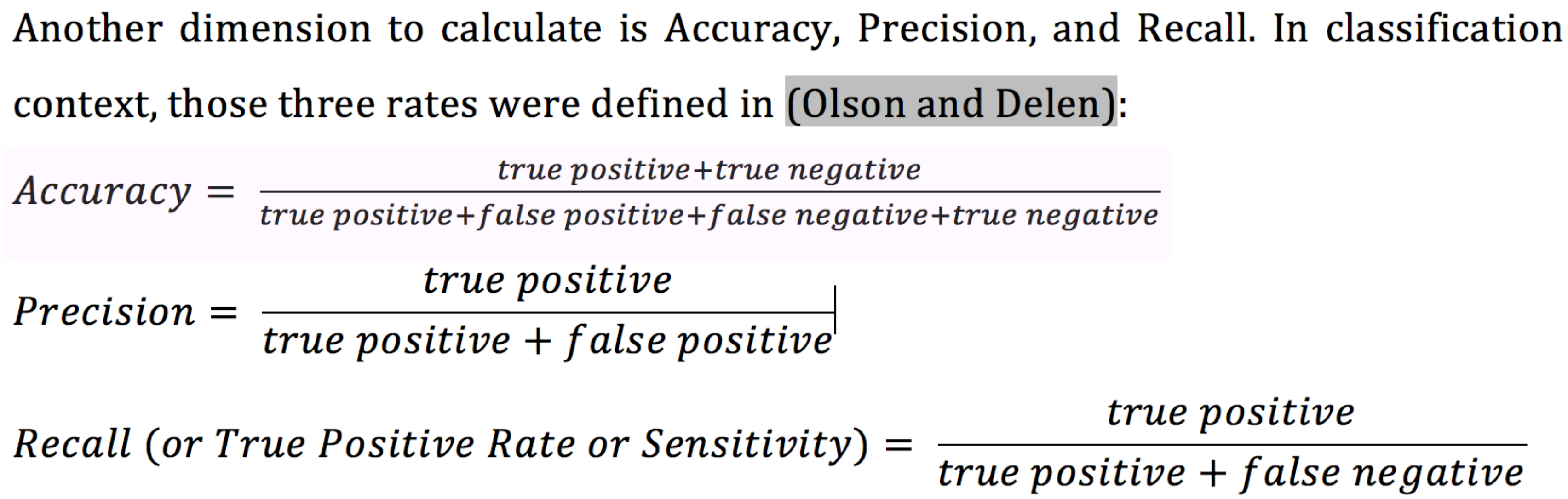


95%	Total Files Match with Extracted Features											Num of Samples
	BMP	PNG	JPG	GIF	PDF	DOC	PPT	XLS	DOCX	PPTX	XLSX	
Sample FileType												
BMP	7	0	0	0	0	0	0	0	0	0	0	8
PNG	0	10	0	0	0	0	0	0	0	0	0	10
JPG	0	0	9	0	0	0	0	0	0	0	0	10
GIF	0	0	0	4	0	0	0	0	0	0	0	10
PDF	0	0	1	0	10	0	0	0	0	0	0	10
DOC	2	0	1	0	0	4	0	0	0	0	0	5
PPT	5	6	5	0	0	0	7	0	0	0	0	9
XLS	1	0	0	0	0	0	0	5	0	0	0	5
DOCX	0	4	4	1	0	0	0	0	8	0	0	8
PPTX	3	9	8	3	0	0	0	0	0	3	0	9
XLSX	1	0	1	0	0	0	0	1	0	0	5	6
												90
Classification	19	29	29	8	10	4	7	6	8	3	5	

Num of Samples: Given fact that shows number of file samples being tested for a specific file type

Classification: Testing result shows number of files that have more than 95% matching rates for each features extracted for file type

Analysis



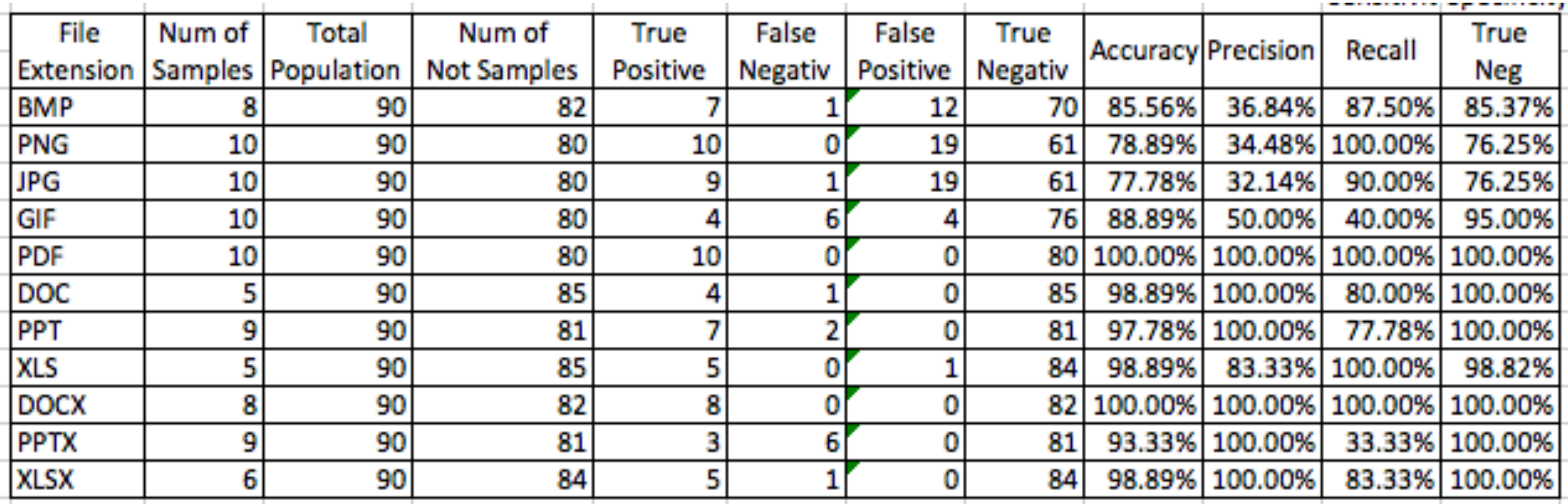
Another dimension to calculate is Accuracy, Precision, and Recall. In classification context, those three rates were defined in (Olson and Delen):

$$Accuracy = \frac{true\ positive + true\ negative}{true\ positive + false\ positive + false\ negative + true\ negative}$$

$$Precision = \frac{true\ positive}{true\ positive + false\ positive}$$

$$Recall\ (or\ True\ Positive\ Rate\ or\ Sensitivity) = \frac{true\ positive}{true\ positive + false\ negative}$$

Analysis



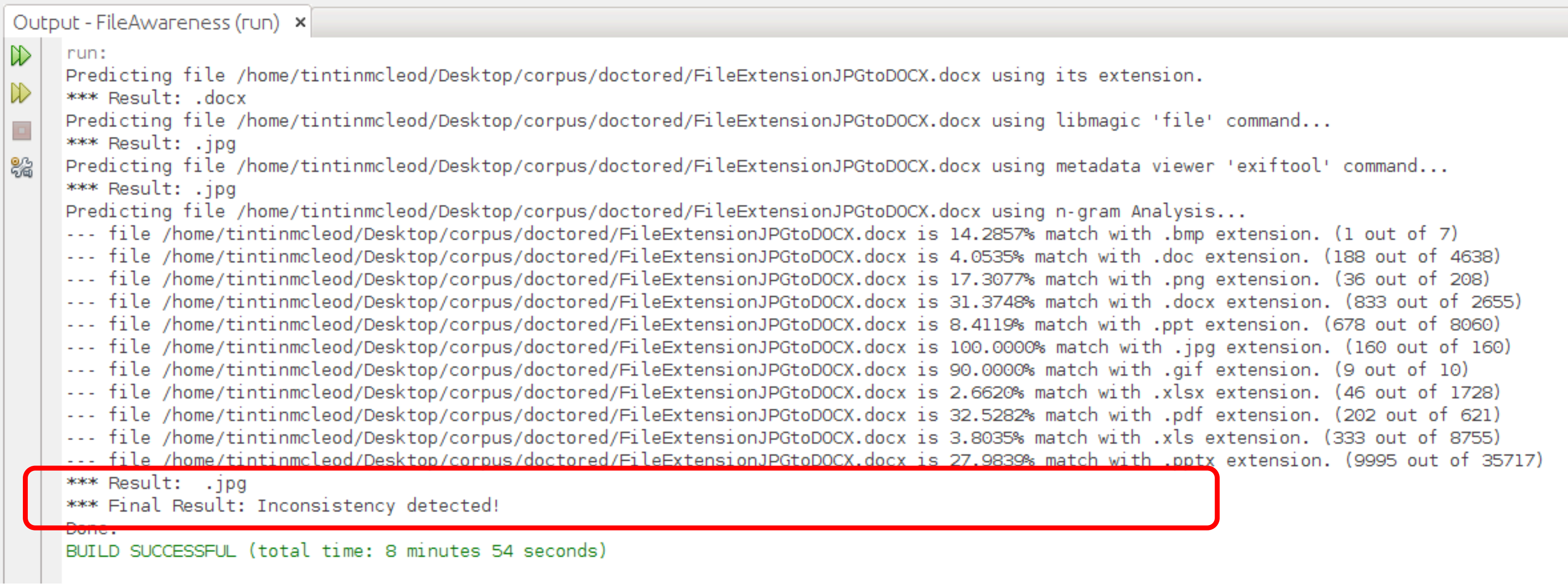
File Extension	Num of Samples	Total Population	Num of Not Samples	True Positive	False Negativ	False Positive	True Negativ	Accuracy	Precision	Recall	True Neg
BMP	8	90	82	7	1	12	70	85.56%	36.84%	87.50%	85.37%
PNG	10	90	80	10	0	19	61	78.89%	34.48%	100.00%	76.25%
JPG	10	90	80	9	1	19	61	77.78%	32.14%	90.00%	76.25%
GIF	10	90	80	4	6	4	76	88.89%	50.00%	40.00%	95.00%
PDF	10	90	80	10	0	0	80	100.00%	100.00%	100.00%	100.00%
DOC	5	90	85	4	1	0	85	98.89%	100.00%	80.00%	100.00%
PPT	9	90	81	7	2	0	81	97.78%	100.00%	77.78%	100.00%
XLS	5	90	85	5	0	1	84	98.89%	83.33%	100.00%	98.82%
DOCX	8	90	82	8	0	0	82	100.00%	100.00%	100.00%	100.00%
PPTX	9	90	81	3	6	0	81	93.33%	100.00%	33.33%	100.00%
XLSX	6	90	84	5	1	0	84	98.89%	100.00%	83.33%	100.00%

- ‘Accuracy’: 80% vs 49%
- Average Accuracy: 92.63% vs 96.33%
- Average Precision: 76.07% vs 51%
- Average Recall: 81.09% vs 48%

What if it is tested on …

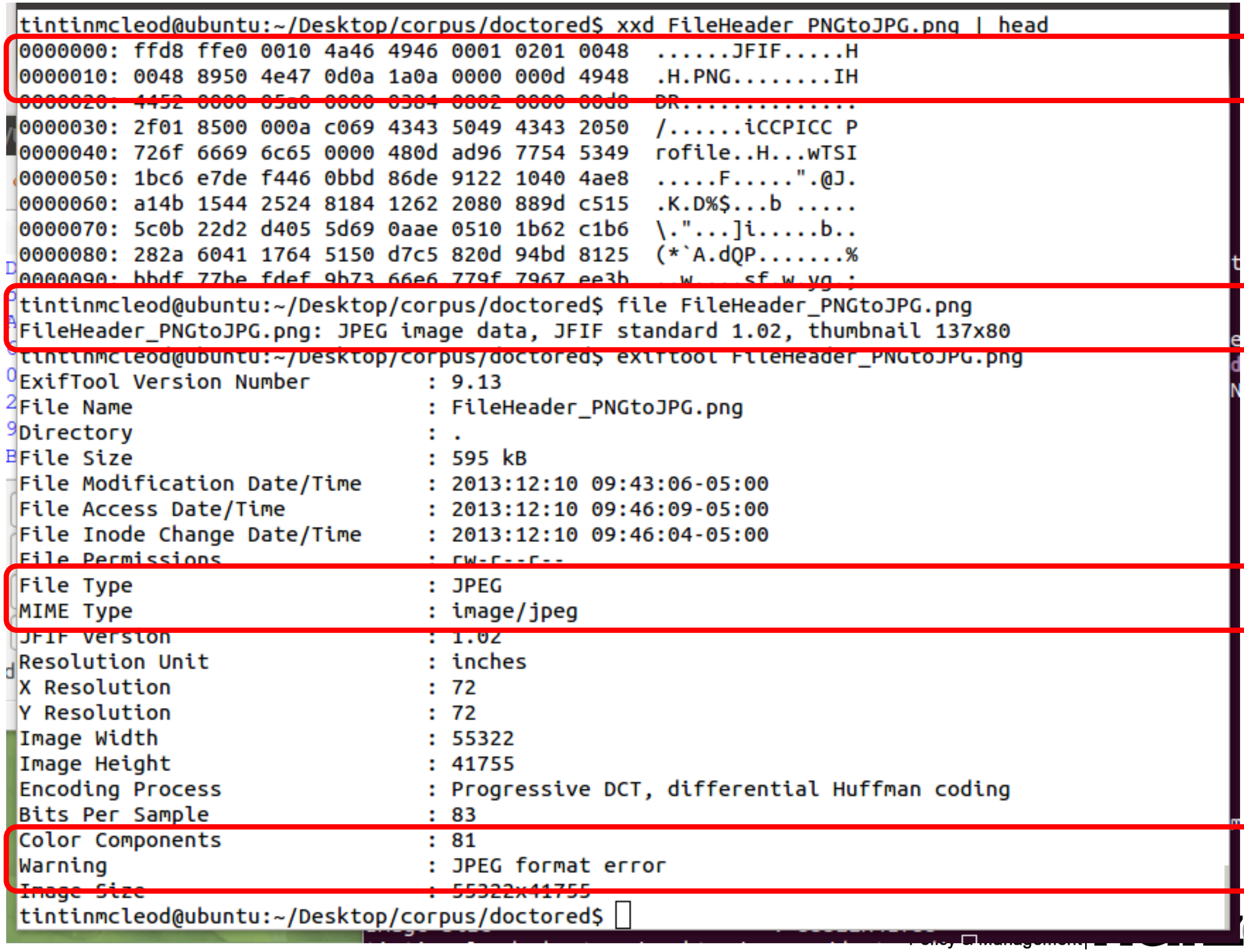
- Altered File Extension?
- Altered Libmagic in File Header?
- Altered File Header?
- Random bytes of data?

Altered File Extension



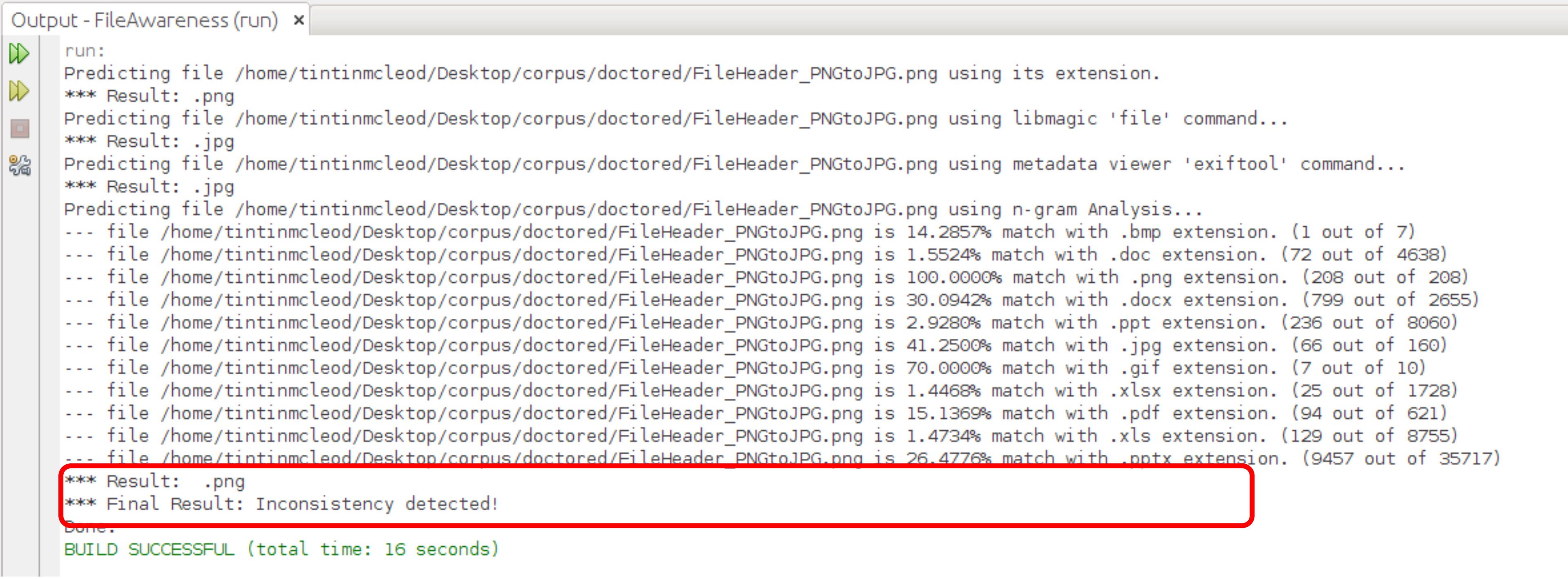
```
run:
Predicting file /home/tintinmcleod/Desktop/corpus/doctored/FileExtensionJPGtoDOCX.docx using its extension.
*** Result: .docx
Predicting file /home/tintinmcleod/Desktop/corpus/doctored/FileExtensionJPGtoDOCX.docx using libmagic 'file' command...
*** Result: .jpg
Predicting file /home/tintinmcleod/Desktop/corpus/doctored/FileExtensionJPGtoDOCX.docx using metadata viewer 'exiftool' command...
*** Result: .jpg
Predicting file /home/tintinmcleod/Desktop/corpus/doctored/FileExtensionJPGtoDOCX.docx using n-gram Analysis...
--- file /home/tintinmcleod/Desktop/corpus/doctored/FileExtensionJPGtoDOCX.docx is 14.2857% match with .bmp extension. (1 out of 7)
--- file /home/tintinmcleod/Desktop/corpus/doctored/FileExtensionJPGtoDOCX.docx is 4.0535% match with .doc extension. (188 out of 4638)
--- file /home/tintinmcleod/Desktop/corpus/doctored/FileExtensionJPGtoDOCX.docx is 17.3077% match with .png extension. (36 out of 208)
--- file /home/tintinmcleod/Desktop/corpus/doctored/FileExtensionJPGtoDOCX.docx is 31.3748% match with .docx extension. (833 out of 2655)
--- file /home/tintinmcleod/Desktop/corpus/doctored/FileExtensionJPGtoDOCX.docx is 8.4119% match with .ppt extension. (678 out of 8060)
--- file /home/tintinmcleod/Desktop/corpus/doctored/FileExtensionJPGtoDOCX.docx is 100.0000% match with .jpg extension. (160 out of 160)
--- file /home/tintinmcleod/Desktop/corpus/doctored/FileExtensionJPGtoDOCX.docx is 90.0000% match with .gif extension. (9 out of 10)
--- file /home/tintinmcleod/Desktop/corpus/doctored/FileExtensionJPGtoDOCX.docx is 2.6620% match with .xlsx extension. (46 out of 1728)
--- file /home/tintinmcleod/Desktop/corpus/doctored/FileExtensionJPGtoDOCX.docx is 32.5282% match with .pdf extension. (202 out of 621)
--- file /home/tintinmcleod/Desktop/corpus/doctored/FileExtensionJPGtoDOCX.docx is 3.8035% match with .xls extension. (333 out of 8755)
--- file /home/tintinmcleod/Desktop/corpus/doctored/FileExtensionJPGtoDOCX.docx is 27.9839% match with .pptx extension. (9995 out of 35717)
*** Result:  .jpg
*** Final Result: Inconsistency detected!
Done.
BUILD SUCCESSFUL (total time: 8 minutes 54 seconds)
```

Altered File Header



```
tintinmcleod@ubuntu:~/Desktop/corpus/doctored$ xxd FileHeader_PNGtoJPG.png | head
0000000: ffd8 ffe0 0010 4a46 4946 0001 0201 0048  ......JFIF.....H
0000010: 0048 8950 4e47 0d0a 1a0a 0000 000d 4948  .H.PNG........IH
0000020: 4452 0000 05a0 0000 0384 0802 0000 00d8  DR..............
0000030: 2f01 8500 000a c069 4343 5049 4343 2050  /......iCCPICC P
0000040: 726f 6669 6c65 0000 480d ad96 7754 5349  rofile..H...wTSI
0000050: 1bc6 e7de f446 0bbd 86de 9122 1040 4ae8  .....F.....".@J.
0000060: a14b 1544 2524 8184 1262 2080 889d c515  .K.D%$...b .....
0000070: 5c0b 22d2 d405 5d69 0aae 0510 1b62 c1b6  \."...]i.....b..
0000080: 282a 6041 1764 5150 d7c5 820d 94bd 8125  (*`A.dQP.......%
0000090: bbdf 77be fdef 9b73 66e6 779f 7967 ee3b  ..w....sf.w.yg.;
tintinmcleod@ubuntu:~/Desktop/corpus/doctored$ file FileHeader_PNGtoJPG.png
FileHeader_PNGtoJPG.png: JPEG image data, JFIF standard 1.02, thumbnail 137x80
tintinmcleod@ubuntu:~/Desktop/corpus/doctored$ exiftool FileHeader_PNGtoJPG.png
ExifTool Version Number         : 9.13
File Name                       : FileHeader_PNGtoJPG.png
Directory                       : .
File Size                       : 595 kB
File Modification Date/Time     : 2013:12:10 09:43:06-05:00
File Access Date/Time           : 2013:12:10 09:46:09-05:00
File Inode Change Date/Time     : 2013:12:10 09:46:04-05:00
File Permissions                : rw-r--r--
File Type                       : JPEG
MIME Type                       : image/jpeg
JFIF Version                    : 1.02
Resolution Unit                 : inches
X Resolution                    : 72
Y Resolution                    : 72
Image Width                     : 55322
Image Height                    : 41755
Encoding Process                : Progressive DCT, differential Huffman coding
Bits Per Sample                 : 83
Color Components                : 81
Warning                         : JPEG format error
Image Size                      : 55322x41755
tintinmcleod@ubuntu:~/Desktop/corpus/doctored$ 
```

Altered File Header



```
run:
Predicting file /home/tintinmcleod/Desktop/corpus/doctored/FileHeader_PNGtoJPG.png using its extension.
*** Result: .png
Predicting file /home/tintinmcleod/Desktop/corpus/doctored/FileHeader_PNGtoJPG.png using libmagic 'file' command...
*** Result: .jpg
Predicting file /home/tintinmcleod/Desktop/corpus/doctored/FileHeader_PNGtoJPG.png using metadata viewer 'exiftool' command...
*** Result: .jpg
Predicting file /home/tintinmcleod/Desktop/corpus/doctored/FileHeader_PNGtoJPG.png using n-gram Analysis...
--- file /home/tintinmcleod/Desktop/corpus/doctored/FileHeader_PNGtoJPG.png is 14.2857% match with .bmp extension. (1 out of 7)
--- file /home/tintinmcleod/Desktop/corpus/doctored/FileHeader_PNGtoJPG.png is 1.5524% match with .doc extension. (72 out of 4638)
--- file /home/tintinmcleod/Desktop/corpus/doctored/FileHeader_PNGtoJPG.png is 100.0000% match with .png extension. (208 out of 208)
--- file /home/tintinmcleod/Desktop/corpus/doctored/FileHeader_PNGtoJPG.png is 30.0942% match with .docx extension. (799 out of 2655)
--- file /home/tintinmcleod/Desktop/corpus/doctored/FileHeader_PNGtoJPG.png is 2.9280% match with .ppt extension. (236 out of 8060)
--- file /home/tintinmcleod/Desktop/corpus/doctored/FileHeader_PNGtoJPG.png is 41.2500% match with .jpg extension. (66 out of 160)
--- file /home/tintinmcleod/Desktop/corpus/doctored/FileHeader_PNGtoJPG.png is 70.0000% match with .gif extension. (7 out of 10)
--- file /home/tintinmcleod/Desktop/corpus/doctored/FileHeader_PNGtoJPG.png is 1.4468% match with .xlsx extension. (25 out of 1728)
--- file /home/tintinmcleod/Desktop/corpus/doctored/FileHeader_PNGtoJPG.png is 15.1369% match with .pdf extension. (94 out of 621)
--- file /home/tintinmcleod/Desktop/corpus/doctored/FileHeader_PNGtoJPG.png is 1.4734% match with .xls extension. (129 out of 8755)
--- file /home/tintinmcleod/Desktop/corpus/doctored/FileHeader_PNGtoJPG.png is 26.4776% match with .pptx extension. (9457 out of 35717)
*** Result:  .png
*** Final Result: Inconsistency detected!
Done.
BUILD SUCCESSFUL (total time: 16 seconds)
```

Conclusion

- Different way to choose learning data set makes automated extraction of n-gram in files faster. The prediction result yields with high accuracy, precision, and recall rate
- Can detect file container and the file type inside the container

Further Development

- To reduce learning time, use multi-threaded algorithm and multi-processor
- Possibility to develop using GPU such nvidia CUDA to have faster calculation process using jCuda library
- Statistical approach to classification: Support Vector Machine, k-Nearest Neighbor for higher rate of accuracy, precision, and recall

Bibliography

- Darwin, F. Ian. Fine Free File Command. 9 9 2013 <http://www.darwinsys.com/file/>.
- McDaniel, Mason B. An Algorithm for Content-Based Automated File Type Recognition. Master Thesis. James Madison University. Harrisonburg, 2001.
- Damashek, Marc. "Gauging Similarity with n-Grams: Language-Independent Categorization of Text." Science 267.5199 (1995): 843-848.
- Gopal, Siddharth, et al. "Statistical Learning for File-Type Identification." International Conference on Machine Learning and Applications and Workshops (ICMLA) 10th. Honolulu: Software Engineering Institute, 2011.
- Hiller, Scot. The System Registry. 1996. Microsoft. 5 October 2013 <http://msdn.microsoft.com/en-us/library/ms970651.aspx>.
- Mayer, Ryan C. Filetype Identification Using Long, Summarized n-Grams. Monterey: Master Thesis at Naval Postgraduate School, 2011.
- Axelsson, Stefan. "The Normalised Compression Distance as a file fragment classifier." The International Journal of Digital Forensics & Incident Response 7.Supplement (2010): S24-S31.

Thank You