International Cyber Security and Machine Learning

AVIAD COHEN

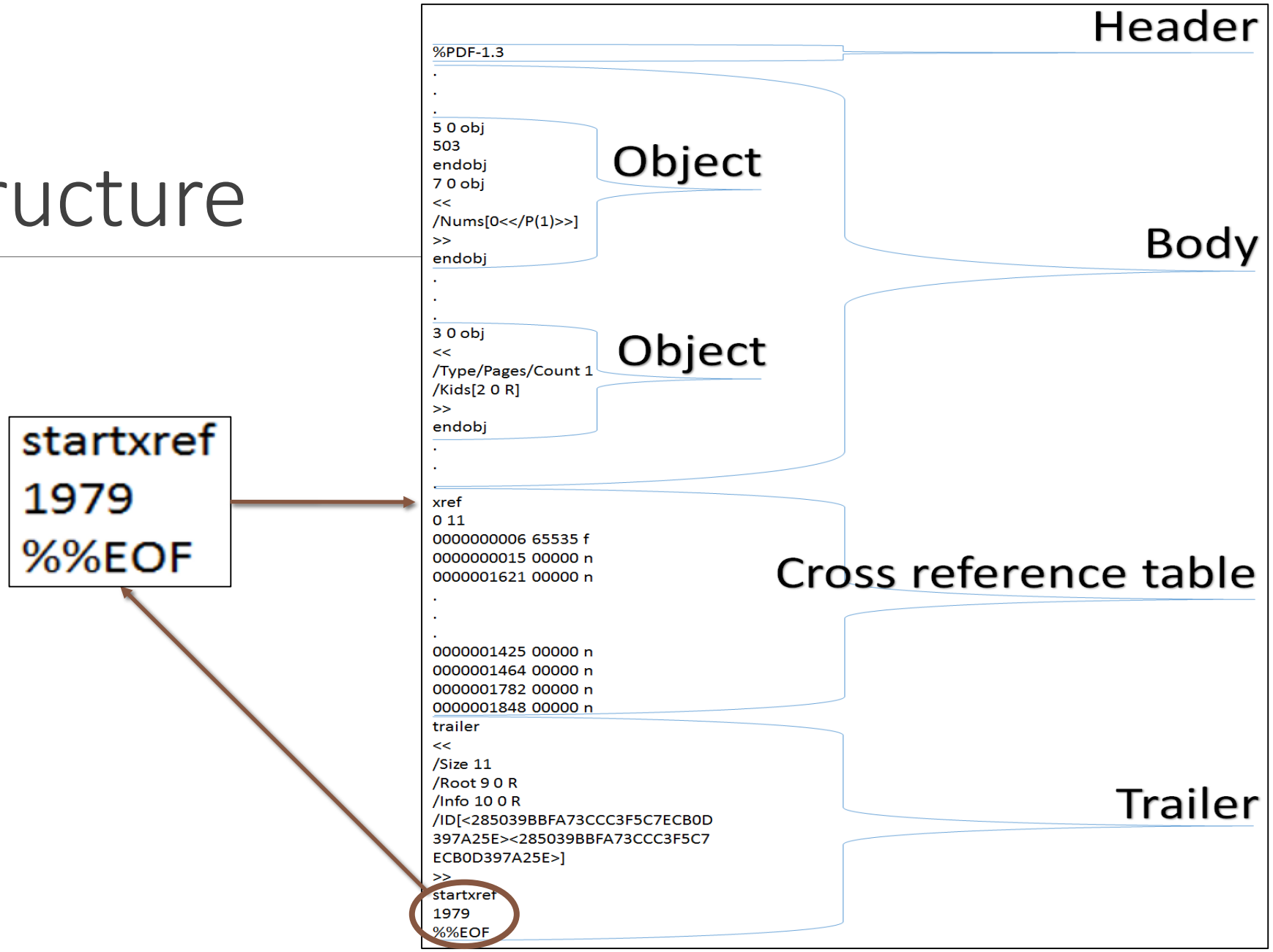
AVIADJO@GMAIL.COM

Lab: Feature Extraction, Selection and Dataset Creation

Outline

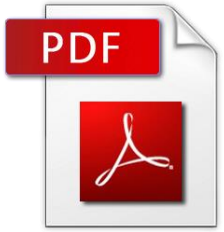
- Feature Extraction
  - Byte N-gram
  - String N-gram
  - PDF Keywords
- Dataset Creation
- Feature Selection In Excel
  - Fisher Score

PDF File Structure



PDF - Keywords Extraction using *PDFid*

- PDF Keywords feature extractor extracts only 14 features from each file.



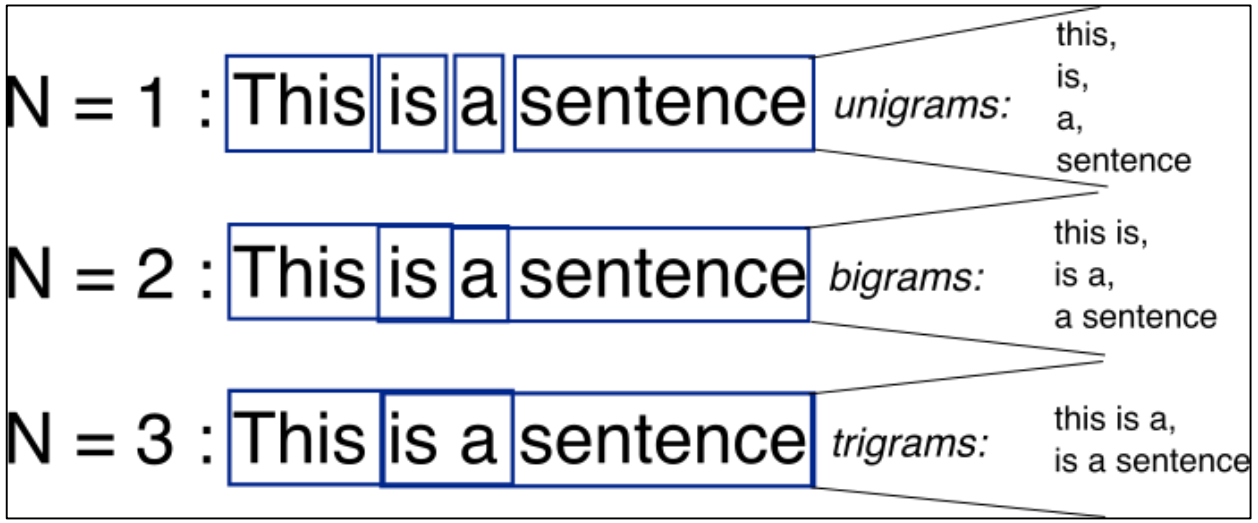
Keywords extraction

```
# pdfid.py test.pdf
PDFiD 0.0.2 test.pdf
 PDF Header: %PDF-1.1
 obj                    7
 endobj                 7
 stream                 1
 endstream              1
 xref                   1
 trailer                1
 startxref              1
 /Page                  1
 /Encrypt               0
 /JS                    1
 /JavaScript            1
 /AA                    0
 /OpenAction            1
 /JBIG2Decode           0
```

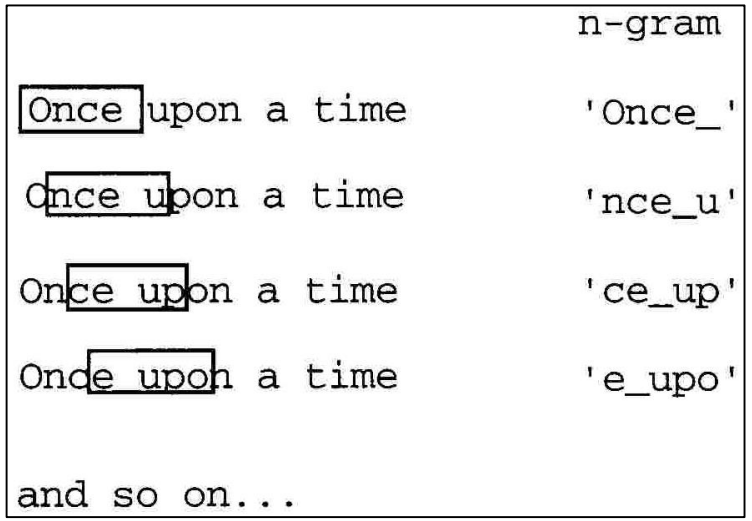
Feature Extraction – Example 2 – N-gram

- When using N-gram approach you should define what is a gram? Character/Word/Line/Byte
- When working on documents or code, a gram can be character/word/Line.
- It is common to use skip of 1 gram.

1-gram, 2-grams, 3-grams (gram=word)



5-grams (gram=character)



Feature Extraction – Example 2 – N-gram

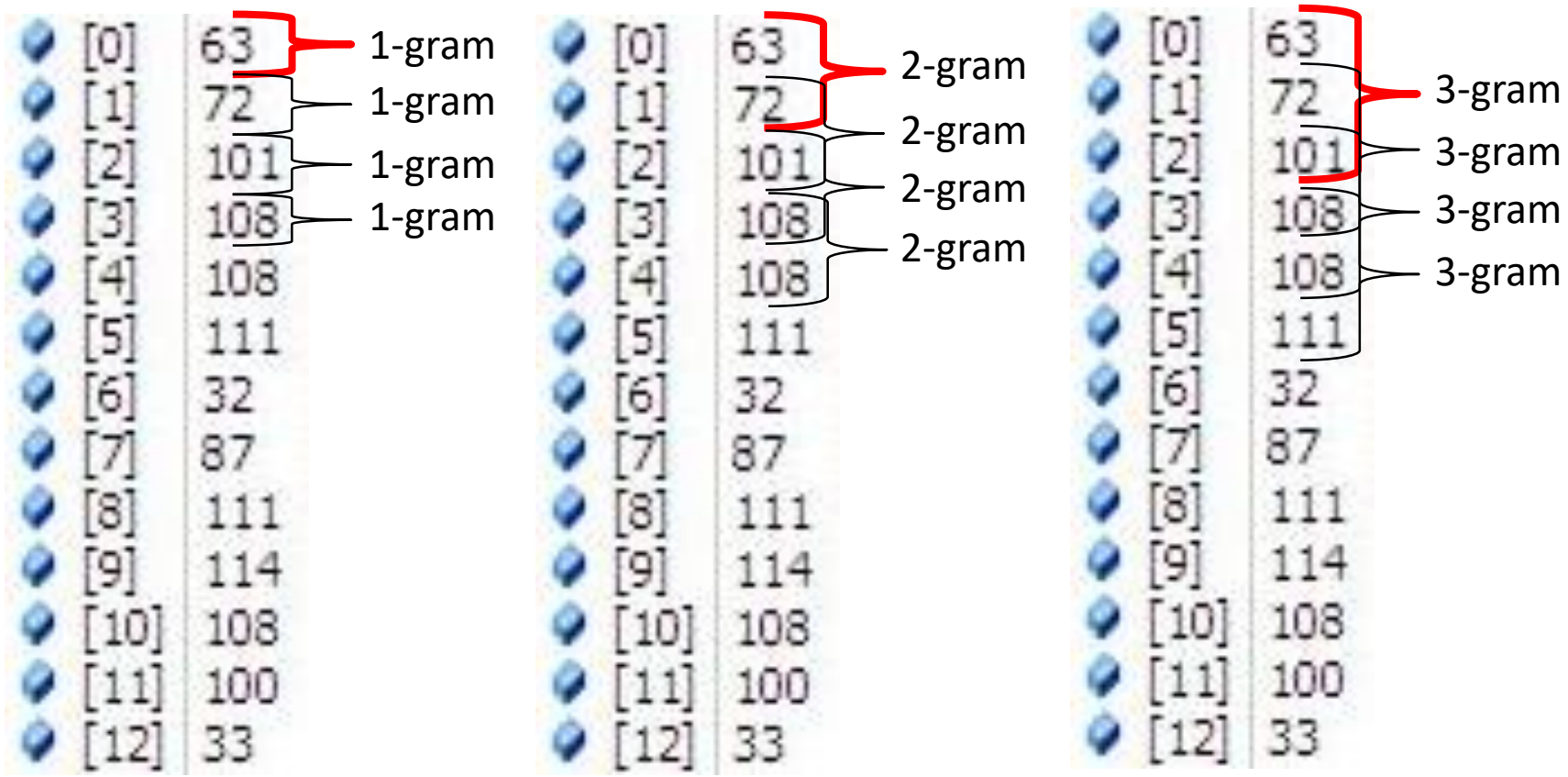
- When working on files without textual content (e.g., executable files, binary files) the preferred gram is byte (8 bits).
- Any file can be read as byte array (Byte[]).
- Byte value is between 0-255 ($2^8 = 256$).
- Byte can be represented with two hexadecimal (base 16) characters ($16 * 16 = 256$).

TestBinary.bson

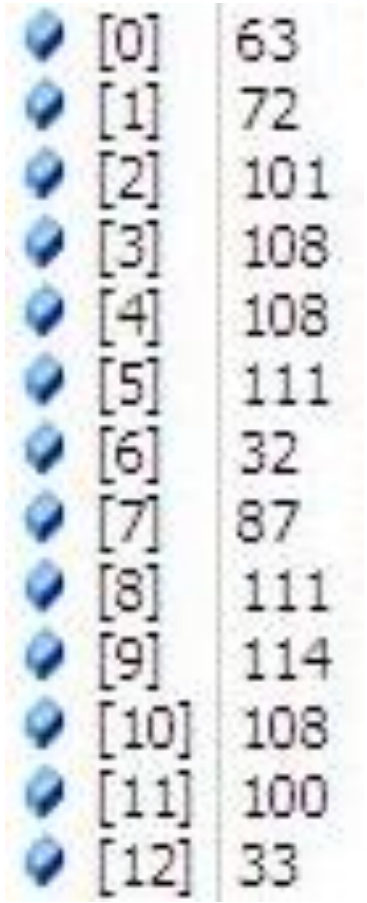
Offset(h)	00	01	02	03	04	05	06	07	08	09	0A	0B	0C	0D	0E	0F	
00000000	54	68	69	73	20	77	69	6C	6C	20	62	65	20	61	20	34	This will be a 4
00000010	30	47	42	20	62	79	74	65	20	73	74	72	65	61	6D	21	0GB byte stream!
00000020	64	00	00	00	03	48	65	61	64	65	72	00	4E	00	00	00	d....Header.N...
00000030	03	53	75	62	48	65	61	64	65	72	31	00	21	00	00	00	.SubHeader1.!...
00000040	02	4E	61	6D	65	00	05	00	00	00	42	6F	6E	64	00	10	.Name.....Bond..
00000050	4C	69	63	65	6E	73	65	00	07	00	00	00	00	03	53	75	License.......Su
00000060	62	48	65	61	64	65	72	32	00	10	00	00	00	08	49	73	bHeader2......Is
00000070	41	63	74	69	76	65	00	01	00	00	0A	50	61	79	6C	6F	Active.....Paylo
00000080	61	64	00	00													ad..

Feature Extraction – Example 2 – N-gram

Byte N-gram



Feature Extraction – Example 2 – N-gram



1-gram	Freq.
63	1
72	1
101	1
108	3
111	2
32	1
87	1
114	1
100	1
33	1

2-gram	Freq.
63,72	1
72,101	1
101,108	1
108,108	1
108,111	1
111,32	1
32,87	1
87,111	1
111,114	1
114,108	1
108,100	1
100,33	1

3-gram	Freq.
63,72,101	1
72,101,108	1
101,108,108	1
108,108,111	1
108,111,32	1
111,32,87	1
32,87,111	1
87,111,114	1
111,114,108	1
114,108,100	1
108,100,33	1

Feature Selection

Feature Selection – Filter Methods

Fisher Score

- Calculates the difference between positive and negative examples relative to certain feature, in terms of mean and standard deviation.
- Higher rank = higher contribution.
- $R_i = \frac{\left|\mu_{i,p} - \mu_{i,n}\right|}{\sigma_{i,p} + \sigma_{i,n}}$