Requirements

- Python
  - Orange library
    - http://orange.biolab.si/

Files

- classyPDF.py
  - class to handle post-parsing
- judge.py
  - read single file or directory and classify the sample
- data/complete.tab
  - full tabular format of usable features based on the following datasets
    - malicious - ~15K
    - non-malicious - ~6K
    - targeted - 320
- peepdf_classy
  - modified version of peepdf parser

Building Orange on Ubuntu

1. svn co http://orange.biolab.si/svn/orange/trunk/orange/
2. svn co http://orange.biolab.si/svn/orange/trunk/source
3. make
4. follow the rest of the directions here:
   a. http://orange.biolab.si/svn/orange/trunk/orange/doc/INSTALL.linux.txt

General notes

Features

Ordered by relevance to the default dataset.

linearized	boolean int
js	count
javascript	count
names	count
importData	count

aa	count
encrypted	boolean int
sElements (suspicious elements)	count
ccittfaxDecode	count
acroform	count
flash	count
dctDecode	count
flateDecode	count
sActions (suspicious actions)	count
sEvents (suspicious events)	count
asciiHexDecode	count
nLrgObjs (large objects)	count
openAction	count
nVersions	count
ascii85Decode	count
pageCount	count
filesize	int
nObjs	count
jbig2decode	count
nSmlObjs (small objects)	count
nStreams (steams)	count
runLengthDecode	count
lzwDecode	count
launch	count
embeddedFiles	count
submitForm	count
encrypt	count

jpxDecode	count
crypt	count

Swapping classifiers

Classifiers can be adjusted within the judge.py file towards the top where the standard KNN filter has been defined. See documentation on the orange website on other available classifiers. It is possible to combine multiple classifiers together for better results.

False positives

Apart from feature extraction, not much has done in regards to tuning the classifier or feature pruning. The best way to deal with false positives is to define an acceptable threshold for the classifier decision. Workflows should be put in place so that properly or manually classified documents are appended to the end of the complete dataset. This will improve the classifier over time and ensure it continues functioning as intended.

Feature relevance

The current classifier data includes 35 total features to represent the PDF. Feature relevance can be calculated against the dataset to identify features that may provide little to no value. It is recommended to adjust these based on your working needs. Note, this should be done using the Orange library and not deleting from the actual data tab file.

Testing accuracy

Several tests for accuracy exist within the Orange framework. It is recommended that these tests be performed on the data during any major or minor change. This testing ensures that no newly introduced data severely influences the outcome of the classifier.