Needle Toolkit - Project Documentation

(**Needle:** PDF Malware Analysis and Risk Assessment Framework)

Author: Om Fulsundar
Version: 1.0

Contents :

1. Abstract :

Needle is a modular command line tool designed specifically with the intent of analysing PDF files for possible malicious indicators. The tool uses techniques like metadata inspection, object detection, keyword scanning, extensive parsing of streams, and IOC detection before assigning a heuristic calculated score. The results are reflected on the screen as well as in a text file.

The toolkit is created with specific educational and defensive considerations, which allow students, interns, and analysts to gain first-hand experience of PDF malware analysis techniques. By analysing changes in metadata, embedded objects, suspicious keywords, and indicators of compromise, Needle explains how attackers use PDF documents for malicious activities.

This project emphasizes the significance of multi-layered security measures because Needle, while offering static analysis, also emphasizes the vulnerability of heuristic scoring alone, especially in addition to dynamic sandboxing. This toolkit provides an interface between theoretical knowledge and practical testing, thus reiterating skills in the field of malware report writing and assessment.

---

2. Keywords :

PDF malware, static analysis, metadata anomalies, IOC extraction, risk scoring, report generation, cybersecurity toolkit, CLI framework.

---

3. Introduction and Motivation :

PDFs are one of the most exploited document formats due to their ability to embed scripts, objects, and external references. Attackers leverage these features for phishing, exploits, and payload delivery. Needle was developed to provide a lightweight, modular framework for static PDF analysis.

Motivation:

- **Defensive perspective:** Help analysts and students understand how malicious PDFs are structured and why static checks are important for triage.
- **Educational value:** Provide hands-on experience with metadata inspection, IOC detection, and risk scoring.
- **Practical use:** Generate reproducible reports that can be archived and compared across multiple samples.

---

4. Design and Architecture :

Components

- **main.py** — CLI driver orchestrating analysis.
- **metadata_extractor.py** — Extracts author, creator, producer, and anomalies.
- **object_enumerator.py** — Counts embedded objects and types.
- **keyword_scanner.py** — Detects suspicious keywords.
- **deep_parser.py** — Extracts payload snippets and streams.
- **ioc_extractor.py** — Finds domains, IPs, file paths, registry keys.
- **risk_engine.py** — Aggregates findings and assigns risk score.
- **report.py** — Generates console and saved reports.

Design Principles

- **Modularity:** Each function separated into its own module.
- **Reproducibility:** Reports saved in results/ for later review.
- **Clarity:** Console output mirrors saved report for transparency.
- **Extensibility:** Easy to add new detection modules or scoring rules.

---

5. Implementation Details :

- **Metadata Analysis:** Extracts PDF metadata and flags anomalies.
- **Object Enumeration:** Identifies embedded objects and suspicious structures.
- **Keyword Scanning:** Searches for attack-related keywords (/JavaScript, /OpenAction, /EmbeddedFile).
- **Deep Parsing:** Extracts and decodes streams, highlighting suspicious payloads.
- **IOC Extraction:** Regex-based detection of URLs, IPs, file paths, and suspicious strings.
- **Risk Engine:** Assigns severity based on weighted indicators.
- **Report Generation:** Produces structured text reports with findings and risk assessment.

---

6. Workflow and Diagrams :

Workflow Steps:

1. Load PDF file.
2. Extract metadata.
3. Enumerate objects.
4. Scan for suspicious keywords.
5. Parse streams and extract payloads.
6. Identify IOCs.
7. Compute risk score.
8. Generate final report.

(Flowchart and workflow diagrams included separately in visuals file.)

---

7. Requirement Libraries :

To run Needle, the following libraries are required:

- **Python 3.x** (runtime environment)
- **PyPDF2** (basic PDF parsing)
- **re** (regex for IOC extraction)
- **hashlib** (hashing payloads for integrity checks)
- **os / sys** (file and system utilities)
- **datetime** (timestamps for reports)

---

8. Usage Overview :

Example execution:

```
python3 main.py data/malware_samples/sample.pdf
```

- Console output shows analysis summary.
- Detailed report saved in results/sample_report.txt.

---

9. Reporting and Outputs :

Reports include:

- Metadata section with anomalies.
- Object and keyword findings.
- Extracted IOCs.
- Risk score and severity level.
- Reasons for scoring.

Screenshot explanations (separate file):

- Terminal execution of sample PDF.
- Saved report in results/.
- Results directory overview.
- Test PDF sample used for analysis.

---

10. Evaluation and Limitations :

Strengths:

- Modular design, clear reporting, reproducible outputs.
- Effective for static triage of suspicious PDFs.

Limitations:

- Regex-based IOC detection may miss obfuscation.
- Cannot analyze encrypted/password-protected PDFs.
- Risk scoring is heuristic, not definitive.
- No dynamic execution of JavaScript.

---

11. Future Work :

- Integrate threat intelligence feeds for IOC validation.
- Add visualization (graphs of IOC counts, risk distribution).
- Export reports to PDF/HTML.
- Enhance JavaScript parsing for obfuscation detection.
- Support batch analysis of multiple PDFs.

---

12. Conclusion :

Needle's capability in modular static analysis gives fast insight into the probably malicious PDFs with metadata inspection, object enumeration, keyword scanning, and IOC extraction and risk scoring to produce an actionable report for analysts and students.

This toolkit considers the real-world relevance of PDF malware analysis: attackers often arm PDFs in phishing campaigns and exploit kits, and defenders need to be able to effectively triage them. Needle covers the gap between theoretical studies and practical applications, thus reinforcing skills in malware detection, reporting, and risk assessment.

While currently limited to static checks, Needle considers the concept of layered defences important and provides a basic framework that could be extended in the future by means of dynamic sandboxing and threat intelligence. It is modularly designed so that it will be able to evolve with emerging threats, making it also a very useful educational and research tool.

---

13. References :

- Sample malicious PDF used from GitHub repository (for testing).
- Python standard library documentation (re, hashlib, datetime, os, sys).
- Open-source PDF analysis tools (pdfid, pdf-parser.py, qpdf) consulted for methodology.
