Steps:

~~1. Create Ubuntu virtual machine- use it for your research~~

2. Start with a minimal dataset of blacks and whites.

נקבל מרן בהמשך, בינתיים נעבוד על pdfים שניצור בעצמנו.

3. Over the dataset:

3.1. Extract using Didier tools:

First phase:

|  |  |
| --- | --- |
| PDFID indicators from the pdf | Alexey |
| Use OCR to extract text from images (open source tools) | Alexey |
| Extract all URI | Shir |
| Extract all readable text from the pdf | Shir |
| Save the thumbnail picture using - PIL for example PIL import Image, ImageChops | Alexey |

Second phase:

* Read how to create an efficient image similarity engine - using approximate k- nearest neighbors (K-ANN) – שלב 2
  + How to represent an image so we can search efficiently – שלב 2
  + Which algorithm to use in order to search efficiently – שלב 2
  + constraints: easy to maintain, fast and simple to add new images. – שלב 2

Third phase:

* Develop text engine using word2vec or tfidf – שלב 3
* Create a text engine that detects malware based on the text. – שלב 3

Timeline:

1. extraction of data: 3 weeks [yellow]

2. use similarity to detect images [blue] 3 weeks

3. text engine - 3 weeks [orange]

4. Combine all to one system [6 weeks]

איך אנחנו עובדים על הדברים

* + - מכונה וירטואלית ubuntu
    - להוציא את ה-urlים
    - הדגימות יגיעו מרן – בשלב התחלתי לעבוד על דברים משלנו
    - היום אנטיוירוסים ורוב המוצרים עובדים על בסיס md5ים – ברגע שיש שינוי כלשהו בקובץ זה משנה את ה-md5
    - Uri בדגימה היה באובייקטים של ה-pdf ומתורגם לפעולה כשפותחים אותו.
    - מהקליק הגענו ל-url שזה downloader שכשלוחצים עליו מורד קובץ הרצה.
    - אחרי שיש את כל ה-uri, לשמור את העמוד הראשון כתמונה
    - ספריה בשם PIL שהיא מוציאה images, מוציאה עמוד ראשון ב-pdf וכו'. בעזרתה מוציאים את ה-thumbnail/preview של הקובץ.
    - לדעת לעשות דמיון מהיר ויעיל בין תמונות (image בתוך הקובץ)

קבצים רבים יש להם נתוני מסגרת שונים ליצירת md5ים לא חתומים, ולכן לחפש גם לפי תוכן ולא רק נתוני מסגרת, ולדעת לחפש בצורה יעילה בין תכני תמונה

K-ANN – חיפוש וקטורי

* + - להרחיב קצת על עולם הבעיה