# סיכום הפרויקט

**יצירת מילון של מחרוזות שתוקפים ירצו להתחזות אליהן – סאבסטרינגים פריוולנטיים**

1. להריץ FILE LISTER ולחלץ את מחרוזות הקבצים (ללא תיקיות)
2. להפריד כל נתיב על פי '\', ויוצאות הרבה מחרוזות בודדות.
3. לסדר אותם לפי האורך, נגיד שWINDOWS הוא הכי ארוך אז נקבע את N=7.
4. עכשיו מתוך המחרוזות המופרדות לחפש את הNGRAMS הנפוצים החל מ N=7
   1. ("C:", "WINDOWS", "SYSTEM", "SYSTEM32"..).
5. למיין על פי FREQ ולקחת TOP K=700 של הNGRAMS הנפוצים.
6. **נשארנו עם מילון של 700 הNGRAMS שהם הכי פריווילנטיים, ולכן גם יעדים להתחזות.**

**יצירת מילון של מתחזים ליוניגרם או בייגרמס מעניינים**

1. עבור כל מחרוזת בTOP 700 NGRAMS שיצרנו קודם:
2. ניקח לדוגמא את WINDOWS
3. ניקח יוניגרמס ובייגרמס , ונשמור עבורם את הפרמוטציות המתחזות
   1. ניקח את W ונוסיף לו את הפרמוטציות המתחזות
   2. ..mal\_uni\_bi\_grams[W].add(‘vv’)
   3. ..mal\_uni\_bi\_grams[W].add(‘w’)
   4. ניקח את WI ונוסיף לו:
      1. ..mal\_uni\_bi\_grams[WI].add(‘vv1’)
4. **ואז בעצם עבור כל מחרוזת במילון של סאבסטרינגים פריוולנטים, ועבור כל יוניגרם או בייגרם במחרוזת הזאת, יש לנו סט של מתחזים (שכל מחרוזת בסט היא גם יוניגרם או בייגרם בעצמה).**

**ואז האלגוריתם במוצר שלנו ישתמש בשני המילונים שלעיל כך:**

1. בהנתן מחרוזת TEST כמו **W1ndovvs** פרק ליוניגרם ובייגרמס
   1. (כלומר כך שw1ndovvs תתפרק ל W, W1, 1, 1N)
2. נחפש את ההומוגליפים לW1ndows, אבל **רק בתנאי שזה מקטין לי את המרחק האמינג** (כמות האובייקטים הלא חופפים) **לאיזשהו סטרינג פריוולנטי**.
   1. למשל, נוכל להחליף את W1 בWI . זה מקטין לנו את המרחק האמינג למחרוזת הפריוולנטית WINDOWS ולכן נמשיך עם זה.
      1. הבא שנחליף יהיה vv עם w . זה גם מקטין לנו את המרחק האמינג לWINDOWS ל0 ולכן נמשיך עם זה.
         1. **הגענו למרחק 0 עבור המחרוזת הפריוולנטית WINDOWS.**נוציא כפלט נסיון התחזות למחרוזת WINDOWS, עם 2 החלפות של בייגרמס.

**השאלה הרצויה: אם אני רואה עכשיו w1ndows אני רוצה לדעת האם זה הומוגליף של משהו פריוולנטי**

יש לי סאסטרינג פריוולנטים, ויש לי הומוגליף ליוני ובייגרמס.

אז צריך להחליף את היוני וביי בהומוגליפים שלהם,

.... עשינו מילון של כל ההומוגליפים לWINDOWS

אבל אני רוצה טבלה יותר קטנה מטבלת כל ההומוגליפים של וינדוס.

נחפש כל הומוגליף של w1ndows אבל רק בתנאי שזה מקטין לי את המרחק ג'קרד לאיזשהו סטרינג פריוולנטי. (מרחק האובייקטים הלא חופפים, נגיד האמינג).

Windows מביא אותי יותר קרוב לw1ndows אז אני שומר אותו.

**איך יוצרים מילון של התחזויות עבור יוניגרם או בייגרמס**

1. עבור כל מחרוזת בTOP 700 NGRAMS שיצרנו קודם:
2. ניקח לדוגמא את WINDOWS
3. במקום להשוות לכל פרמוטציה אפשרית (מה שמייצר הקוד הנוכחי), נרצה לעשות השוואה של כל היוניגרמס והבייגרמס אחת לשנייה (כלומר 27\*28):
   1. ניקח את W, ונחליף אותה בכל הסט עבור W. עבור כל אות בסט, נייצר השוואה אל מול W וניתן מידת מרחק (לפי הפיקסלים החופפים חלקי סך הפיקסלים)
   2. אחרי זה ניקח את WI ונחליף את W בכל הסט, ואת I עם כל הסט (כלומר פרמוטציות של WI)
4. על כל פרמוטציה שיצרנו ליוניגרם או בייגרם נהפוך לתמונה ונשווה את הפרמוטציות למקור, ונוציא ציון.
   1. למשל השוואת תמונה של w1 ל wi, או vv ל w..
5. נראה שהוצאנו ציונים גבוהים למקרים מעניין ולהיפך..
6. **ואז בעצם יצרנו מילון של סאבסטרינגים פריוולנטי יוניגרם או בייגרם יש לנו פרמוטציות מתחזות (שכל מחרוזת שם היא גם יוניגרם או בייגרם).**

איך לנקות הומוגליפים שלא מיוצגים בפונט שלי? לירון מציע לעבור בלולאה על כל הABC והתווים המיוחדים (יש 31 כאלו) ומייצר לכל אחד מהם תמונה. אם היא זהה לתמונה של מלבן (אני מייצר אחד כזה מראש) אז למחוק אותו מהמילון.

1. **להוציא את הNGRAMS הנפוצים**

המטרה היא לקחת את כל הקבצים (להגדיר breaking char כלומר \) ואז מתוך מה שנשאר מהמופרדים, לחפש את הNGRAMS הכי נפוצים.

הסרטינגים הכי נפוצים (מבין כל המחרוזות המופרדים ב\):

C: - this is the most common string

Windows (or users) – which has a lot of files within it

Basicely there wont be filename in the ngrams..

עכשיו בוא נסדר אותם לפי אורך, נגיד שWINDOWS הוא הכי ארוך שיש 7 תווים. אז אין טעם לבדוק NGRAMS שגדולים מ7. אז בוא נחפש מN=7, נגיד WINDOWS הוא מספיק פריוולנטי, אז בחרנו גם בו. אם WINDOWS הוא לא מספיק פריוולנטי אבל יש לו SUB STR אחד או יותר שהם מאוד פריוולנטים, אז נבחר בהם.

מציע להשתמש בספרייה של COLLECTION בCOUNTER וב ב ITER TOOLS CHAIN מאפשר לשרשר לג'נרטור אחד .. מה שלא מופיע בCOUNTER אפשר לעשות ADD למה שלא קיים ועדיין יוסיף אותו עם 1.

ואז לוקחים TOP 700 ומביא את ה700 הכי נפוצים.

1. **איך מוציאים דמיון בין תמונות – ללא הטבלה – כלומר למלא ולהוסיף לטבלה הנוכחית**

**נעשה את זה רק על המחרוזות הממש ממש פריוולנטים (כמו C: או WINDOWS) כי זה יקר , לצורך הדגמה**

בוא נניח שWINDOWS פריוולנטי. בוא נחפש דברים שנראים דומה לWINDOWS. נגיד w1ndows , כמה זה דומה?

כדי לעשות אוטומציה , קח כל NGRAM של 1 או 2 בסטרינג הזה ותנסה להחליף אותו עם כל NGRAM אפשרי של 1 או 2.

לוקחים את הסטרינג והופכים אותו לתמונה, ניקח את כל השינויים הקטנים עליו (כל יוניגרם ובייגרם אפשרי להחליף באחר), ולמדוד את המרחק ביניהם. ואם יש פונקציה שנותנים לה שני מחרוזות ומחזירה מרחק, אנחנו יכולים למלא את הטבלה.

אם אתם יודעים את רשימת הפונטים שמעניינים אתכם, אז אתם יכולים לבנות את הטבלה שלכם בעזרת הדמיון תמונה ^ .

1. **איך כותבים את הפונקציה שמקבלת שתי תמונות ונותנת מידת מרחק**

אנחנו רוצים לייצר השוואת תמונות, אבל אנחנו לא רוצים להשוות כל פרמוטציה אפשרית על WINDOWS, אלא רק לעשות השוואה של כל היוניגרמס והבייגרמס אחת לשנייה (כלומר 27\*28 אפשרויות) ניקח מקרים מעניינים ונראה שהוא נתן מרחק גבוה ביניהם, וההיפך.

איך עושים את זה? רוצים להשוות סטרינגים יחסית קצרים (יוניגרמס בייגרמס):

* מניחים שהופכים סטרינג לתמונה – הוא נשאר בשורה אחת
* מניח שהשורה תמיד באותו גובה.
* רק פיקסלים לבנים או שחורים.

בוא נעשה קורלציה ביניהם, נשים סטרינג אחד על השני, פיקסל הוא או זהה או שונה (שחור או לבן). נספור כמה פיקסלים זהים חלקי כמה פיקסלים יש. (אבל זה לא תמיד טוב, כי הרוחב של התמונה יכול להשתנות). אז בוא ניקח את הסטרינגים, נשים את הקצר יותר ראשון, עליו נניח את השני בנקודה הכי שמאלית שלו (V יחפוף בצד שמאל של W). ואז נזיז קצת ימינה ונצבע בשחור גם משמאל וגם מימין, כך עד שמגיע לקצה.  
אקח את הנקודה בה יש המספר הכי גדול של פיקסלים זהים, ומחלק בכמות הפיקסלים שיש לסטרינג הארוך יותר. זה אומר שאם אני משווה סטרינג לעצמו V לV אז יש לי 100% שוויון. זוהי מידת קרבה. המרחק זה 1 פחות הקרבה. זה סוג של COS SIGN SIMILIARITY.   
אם אני רוצה להשוות את V לWW , אשים בצד שמאל ואחלק ברוחב של WW ואקבל מידת קרבה של רבע או מידת מרחק של שלושת-רבעי.

איך מראים שהמידת מרחק באמת טובה?

**תיקח נגיד את rn ו m שנמצאים גם בטבלה. זו דוגמא לראות שזה נותן מרחק קטן. ולהיפך, סטרינגים רחוקים ניקח בחירה אקראית i ו- 00 .**

אחרי זה אנחנו יכולים להריץ ML על היוניגרם , .. האמת שאין צורך

אפשר לשחזר עם LOG REG ולראות עם איזה ערך של הNGRAMים הוא מצליח לשחזר, האם הסטרינג שקיבלנו הוא הומוגליף של סטרינג פריוולנטי.

**השאלה הרצויה: אם אני רואה עכשיו w1ndows אני רוצה לדעת האם זה הומוגליף של משהו פריוולנטי**

יש לי סאסטרינג פריוולנטים, ויש לי הומוגליף ליוני ובייגרמס.

אז צריך להחליף את היוני וביי בהומוגליפים שלהם,

.... עשינו מילון של כל ההומוגליפים לWINDOWS

אבל אני רוצה טבלה יותר קטנה מטבלת כל ההומוגליפים של וינדוס.

נחפש כל הומוגליף של w1ndows אבל רק בתנאי שזה מקטין לי את המרחק ג'קרד לאיזשהו סטרינג פריוולנטי. (מרחק האובייקטים הלא חופפים, נגיד האמינג).

Windows מביא אותי יותר קרוב לw1ndows אז אני שומר אותו.

**סעיף 2 – SYSCALLS**

האם אותו רצף של סיסקולס, אבל שמתחיל מאקספלויט שונה – גם מסווג כזדוני?

האם זה שהוא מנצל אספלויט חדש, מתחמק מגלאי שנועד

הדאטה סט האוסטרלי, הוא כדי להריץ את המודלים על דאטה סט קיים. כי כחלק מלזהות את החולשות של המודלים, צריך לממש את המודל.

בניית סט חדש:

הצענו נגיד לקחת את Metasploit ולעשות פרמוטציות כאלו:

4 סטינגרים שונים (כמו tcp, reverse\_tcp, http, https)

4 payloadים שונים

לקחת 3 עבור הTRAIN ו 1 עבור הTEST

ולהגיד כל מודל האם הוא יודע להתמודד עם מקרה שהוא לא ראה או כן ראה.