# תיעוד העבודה:

שם: דפי גרובר

תעודת זהות: 205681836

## על הפרויקט:

בפרויקט זה יצרתי spam filter שבדוק עבור כל מייל האם נמצא בו איזשהו spam phrase בשורת הנושא של המייל או בתוכן המייל עצמו.

הכלי מקבל בתחילת התוכנית כארגומנט קובץ עם ביטויים רגולריים של spam phrases, על מנת שיוכל לקבוע האם מייל מסוים מכיל spam phrase.

בנוסף התוכנית מקבלת כארגומנטים גם את רשימת המוענים, רשימת הנמענים, שורת הנושא של המייל ותוכן המייל.

במהלך ריצת התוכנית נקבע כמה מיילים נרצה לשלוח ובעת שליחת המייל נקבע באופן אקראי מיהם המוען והנמען.

בנוסף, לפני שליחת המייל נקבע באופן רנדומי האם להוסיף spam phrase למייל ואם כן נבחר להוסיף ביטוי נבחר באופן רנדומי אחד מהביטויים שניתנו ונבחר באופן רנדומי האם להוסיף את הביטוי לנושא המייל או לתוכן.

לאחר שנבחר הביטוי שצריך להוסיף הוא נוסף לנושא או לתוכן המייל בהתאם.

הסיבה להסתברויות הכתובות לעיל הן שרוב המיילים שאנו מקבלים היום אינם ספאם ואם אנו כן מקבלים הודעות ספאם לרוב הביטוי נמצא בנושא המייל.

כאשר התוכנית עוברת על כל המיילים שכל משתמש קיבל במידה והוא מוצא מייל שהוא ספאם הוא מדליק דגל שמודיע שהמייל הזה הינו ספאם, זאת מפאת הגבלות טכנולוגיות של POP3.

בסוף ריצת התוכנית נוצר קובץ למשתמש אדמין ובו פירוט כמות המיילים שהם ספאם שכל משתמש קיבל במהלך ריצת התוכנית.

- **POP3:** POP3 הינו פרוטוקול לשליפה של הודעות דוא"ל משרת דואר מרוחק.
  זהו פרוטוקול שרת–לקוח בשכבת היישום של רשתות TCP/IP.

## בעיות במהלך הפרויקט:

במהלך כתיבת הפרויקט נתקלתי במספר בעיות:

1. חוסר יכולת להעביר את המייל לתיקיית הספאם: בתוכנית העבודה המוצאת שלי התכנון המקורי היה להעביר את המיילים שסווגו כספאם לתיקיית ספאם ייחודית על מנת לשמור עליהם במסודר במקום ייעודי.
   בעקבות מגבלות טכנולוגיות של POP3 (חוסר יכולת להעביר בין תיקיות מייל) יצרתי דגל ייחודי "SPAM" שהמייל מקבל ברגע שהוא מאומת כספאם.

2. <u>התחברות למייל ייעודי על מנת לשלוח/לקבל מיילים:</u> בעקבות השימוש במיילים של Gmail היה צורך לשנות חלק מן הגדרות האבטחה של החשבון על מנת לאפשר קבלה ושליחה של המיילים מהתוכנית.
השינויים שנדרשו הם:
   a. ביטול האימות הדו שלבי
   b. אפשור של "גישה לאפליקציות ברמת אבטחה נמוכה"
לאחר השינויים הללו הבעיה טופלה.

## חוזקות:

- <u>Regex:</u> הכלי משתמש בביטויים רגולריים לזיהוי spam phrase ולכן הוא יכול לזהות ביטוי שיש בו יותר מכמות הרווחים המשוערים או שיש לו במקום רווח טאב לדוגמא ולכן מאפשר גיוון רב יותר של ביטויים לעומת משפטים מובנים.
- <u>הסתברויות:</u> כל הבחירות בנוגע לspam phrase, בין אם זה להוסיפו ואם כן איזה והיכן, נעשות בצורה רנדומית ולכן מדמה בצורה אמיתית יותר את המציאות בה אנו חיים.
  - בחירת האם להוסיף spam phrase למייל: ההסתברות שלא נוסיף ביטוי כזה הינו 0.6, זאת כיוון שרוב המיילים שאנו מקבלים כיום הם לא ספאם.
  - בחירת הביטוי שנרצה להוסיף: נבחר באופן אקראי ושווה בין כל הביטויים שניתנו לנו ביטוי.
  - בחירת מיקום הוספת הביטוי: ההסתברות להוספת הביטוי לנושא המייל הינו 0.65, כיוון שרוב הסיכויים שהביטוי יהיה בנושא המייל.

  בנוסף, בעת שליחת מייל נבחר באופן אקראי ושווה כתובת המוען וכתובת הנמען ובכך מדמים מציאות בצורה טובה יותר.
- <u>דיווח סופי:</u> הכלי מדווח לאדמין בסוף ריצת התוכנית כמה מיילי ספאם כל משתמש קיבל.

## חולשות:

- <u>העברה בין תיקיות:</u> אי אפשר להעביר את המייל שסווג כספאם לתיקיית ספאם ייחודית.

## רכיבים עיקריים:

בפרויקט זה השתמשתי במספר כלים תכנותיים:

1. <u>Singleton:</u> על מנת לוודא שכל המחלקות יעבדו בצורה מתואמת מבלי לחשוש שיהיו הבדלים בריצות, המחלקות הוגדרו כsingleton על מנת לייעל את העבודה.
2. <u>Regex:</u> השימוש בביטוי רגולרי עזר מאוד על מנת לייעל את תהליך החיפוש ומציאת הביטויים הבעייתיים.

3. Pair: על מנת לייעל את העבודה של התוכנית ועל מנת לשמור על פשטות במימוש יצרתי מחלקה Pair שמשמשת לשמור זוג של סטרינג (שמייצג את הכתובת מייל) ומספר (שמייצג את כמות המיילי ספאם שהמשתמש קיבל).