ביולוגיה חישובית תרגיל 2
רעות לב 207385741 יעלה גרנות 209133107

הוראות הרצה לקוד :

1) הרצת הפקודה python main.py בשורת הפקודה ללא צורך בהתקנות נוספות.
2)הרצת קובץ הרצה בשם"main.exe" : פתיחת הקישור לgit המצורף להגשה והורדת קובץ הzip הכלול בתיקייה. לאחר מכן יש להריץ את קובץ הexe ע"י דאבל קליק. קבצי הפלט "plain.txt" ו"perm.txt" יופיעו בתיקייה של קובץ הexe יחד עם קבצי הקלט.
בתחילת התוכנית המשתמש מתבקש לבחור את סוג האלגוריתם הרצוי - אלגוריתם רגיל/אלגוריתם דרווין/אלגוריתם למארק.

```
Which program do you want to run?
For genetic_algorithm enter 1
For darwin_algorithm enter 2
For lamarck_algorithm enter 3:1
```

לנוחיות הבודק ניתן להריץ את האלגוריתם עם מספר ת'רדים במקביל בשביל לקבל תמונה מלאה יותר על הצלחת האלגוריתם ללא צורך בהפעלות חוזרות, לצורך כך יש להוריד את ההערות בשורות 437-442 ולשנות את המשתנה `num_executions` למס' ההרצות במקביל הרצוי.

חלק א' - אופן המימוש של האלגוריתם:

<u>ייצוג הפתרונות</u>

בחרנו לייצג פתרון כמחרוזת בגודל 26 תווים שמכילה פרמוטציה של האותיות האנגליות כך שהאינדקס של כל אות במחרוזת הוא המיקום של האות האנגלית המתאימה, למשל עבור הפתרון yxintozjhebldukvsmpqrcwgaf, האות הראשונה (y) היא הקידוד לאות האנגלית הראשונה (a) וכו'. פתרון תקני יכיל את כל האותיות באנגלית ללא חזרות.
כדי לאכוף את תקניות הפתרונות כתבנו פונקציית תיקון ייעודית (נרחיב בהמשך).

<u>פונקציית הערכה</u>

לכל פתרון פענחנו את הצופן לפי הקידוד של הפתרון, ספרנו את מספר המילים האמיתיות הנמצאות בטקסט המפוענח לפי קובץ המילון "dict.txt" שניתן לנו כקלט.
ככל שמספר המילים היה יותר גבוה פונקציית הכשירות הייתה יותר גבוהה.
בנוסף כדי לוודא שיש קשר לוגי בין המילים ביצענו השוואת התפלגויות בין ההסתברות של כל אות אנגלית להופיע בטקסט נורמלי לבין מספר הפעמים שהיא מופיעה בטקסט שקיבלנו.
את ההפרש בין מספר האותיות המצופה לבין מספר הפעמים שהתקבל בפועל בהופעת כל אות החסרנו מציון הכשירות ובכך הורדנו את הסיכויים של פתרונות חסרות משמעות להתפתח אבולוציונית.
לסיכום פונקציית הכשירות היא מדד של ספירת מילים והתפלגות תקינה של הופעת האותיות בטקסט.

<u>פעולות גנטיות ליצירת הדור הבא</u>

יצרנו אוכלוסייה בגודל 1000, שהוגרלה באופן רנדומלי לחלוטין.
לאחר חישוב הfitness לכל אחד מהפתרונות בחרנו בחירה משוחדת של 60% מהאוכלוסייה בהתאם לציוני הכשירות שהתקבלו לכל אחד, מתוכם הגרלנו זוגות של הורים באופן רנדומלי וביצענו crossover כאשר מיקום החתך של פתרונות ההורים הוא אקראי.
לאחר מכן כל פתרון חדש עובר בדיקה ייעודית שמוודאת את תקינותו כך שכל אות מוכפלת מוחלפת באות חסרה. כאשר הפתרון תקין הוא עובר 2 מוטציות, כאשר כל מוטציה היא החלפה בין 2 אותיות רנדומליות. בצורה הזו יצרנו 60% מאוכלוסיית הדור הבא.

בנוסף, בחרנו 20% מהאוכלוסייה שקיבלו את ציוני הfitness הגבוהים ביותר ופיצלנו אותם לשתי קבוצות :
קבוצה ראשונה - כללה את ה25% בעלי הציון הגבוה ביותר שהועברו ישירות לדור הבא , על מנת לאפשר לפונקציית הכשירות להיות מונוטונית עולה.
קבוצה שנייה - כללה את שאר ה75%, עליהם ביצענו גם 2 מוטציות, מתוך כוונה להגדיל את השונות ולהתקדם לעבר הפתרון הנכון. לבסוף הוספנו אותם לאוכלוסיית הדור הבא.

כדי להשלים את גודל האוכלוסייה לגודל הנקבע , יצרנו את 20% האחרונים של האוכלוסייה בצורה רנדומלית כאשר באופן גנטי ניתן להשוות זאת לתופעה של הגירה המגדילה את המגוון הגנטי.

<u>בעיית התכנסות מוקדמת</u>

כדי למנוע מצב של התכנסות מוקדמת ביצענו את הפעולות הבאות:

- הגדרת גודל האוכלוסייה להיות ערך גדול יחסית המאפשר מגוון פתרונות.
- בכל דור דאגנו להכניס 20% פתרונות חדשים לגמרי.
- ביצוע 2 מוטציות לכל פתרון חדש שנוצר מ crossover וכן לרוב הפתרונות שקיבלו את הציון הגבוה ביותר.
- האלגוריתם מגדיל את מספר המוטציות לאחר כמות מסוימת של דורות בהן לא נצפה שינוי, כדי להגדיל את הסיכויים שהאלגוריתם ייתקע בפתרון לוקאלי.

<u>עצירת ריצה</u>

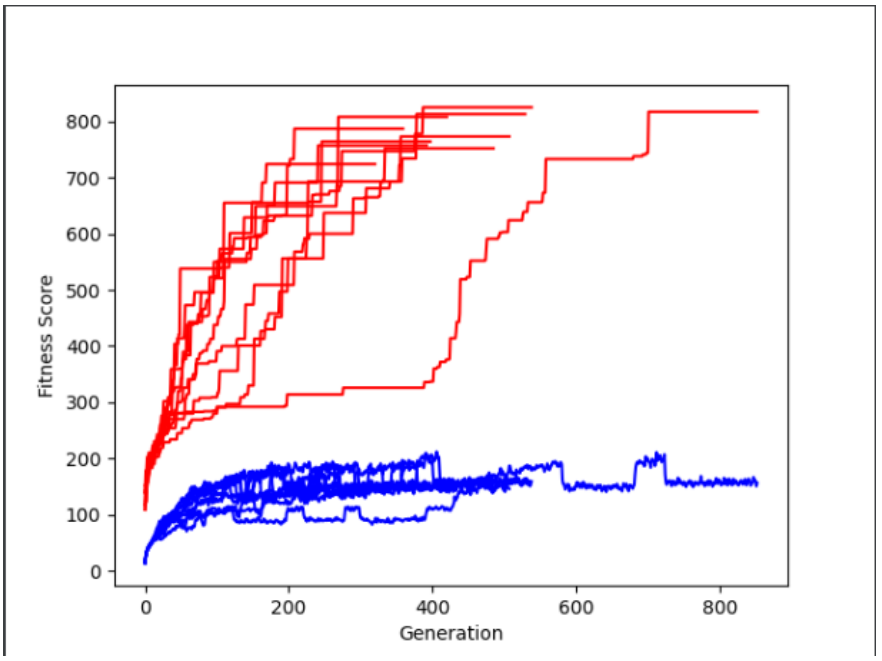
הגדרנו שעצירת התוכנית תקרה לאחר מספר מסוים של דורות בהן לא נצפה שינוי בציון פונקציית הfitness המקסימלי של האוכלוסייה. מצד אחד היה חשוב להגדיר ערך שהוא לא נמוך מדי כדי לשאוף להגיע לתוצאה הקרובה ביותר. ומצד שני שלא יהיה ערך גבוה מדי כדי לאפשר התקדמות ומזעור של זמן הריצה.

<u>התנהגות האלגוריתם</u>

כדי למדוד את ההתנהגות הרצנו את האלגוריתם בצורה מקבילית עם 10 ת'רדים וקיבלנו את הפתרונות הבאים בכל הרצה:

```
Best Solution: YPINTOQJCEXBDUKMSVLZRHWGAF
Best Solution: YJINTOXQCEZLDUGMSVPKRHWBAF
Best Solution: YQINTOBJCEVLDUPMSZXKRHWGAF
Best Solution: KVINTOWXCEBLDUZMSQPJRHYGAF
Best Solution: YJINTOKCXEBLDUQMSVPZRHWGAF
Best Solution: JQINTOPXCEBLDUZMSVKYRHWGAF
Best Solution: YJINTOXQCEBLDVKMSUZPRHWGAF
Best Solution: YVINTOJWCEKLDUXMSQPZRHBGAF
Best Solution: QXINTOYKCEBLDUVMSJPZRHWGAF
Best Solution: BQINTOKXCEYLDUZMSVPJRHWGAF
```

באופן ממוצע נכונות הדיוק של הפתרון עומד על 73.8% שזה כ18-20 אותיות נכונות בכל פתרון.
ממוצע הערך המקסימלי של הfitness הוא 782.
מספר הדורות הממוצע הוא 482.
תוצאות ההרצה באופן גרפי:



חלק ב' - אלגוריתמים גנטי של דרווין ולמארק

בחלק זה הוספנו לאלגוריתם הרגיל מימוש אבולוציוני ע"פ השיטה של דרווין וכן ע"פ השיטה של למארק:

1. <u>אלגוריתם Lamarck:</u>

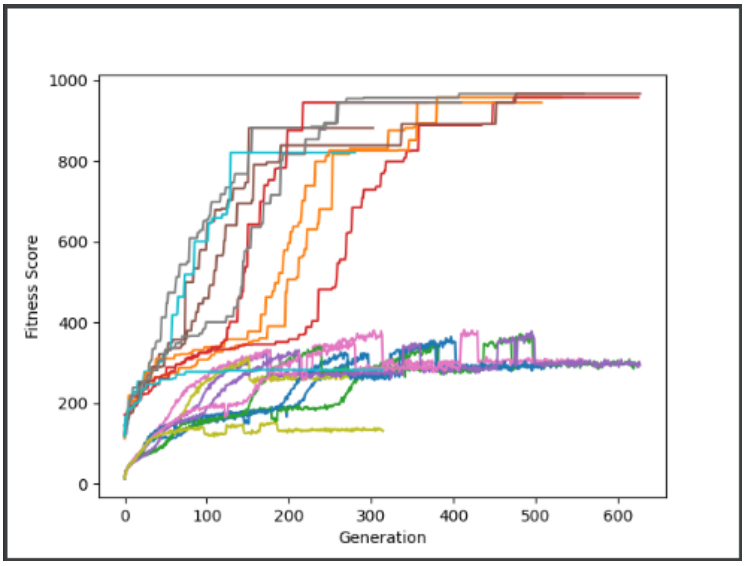
כל פתרון עבר 2 אופוזיציות לוקליות ובדיקה שאכן מדובר באופטימיזציה. במידה וכן, נשמרו השינויים עבור אותו פתרון, אחרת הפתרון חזר לגרסתו המקורית. לפי ציוני הfitness שניתנו לאותן הפתרונות הדור הבא נוצר מהדור הנוכחי <u>שלאחר האופטימיזציות</u> על ידי אותם הצעדים שנעשו באלגוריתם המקורי.

תוצאות -

```
Best Solution: YXINTOZJCEBLDUKMSVPQRHWGAF
Best Solution: YXINTOQJCEBLDUZMSVPKRHWGAF
Best Solution: YXINTOZJCEBLDUKMSVPQRHWGAF
Best Solution: YXINTOQJCEBLDUZMSVPKRHWGAF
Best Solution: YZINTOKQCEBLDUXMSVPJRHWGAF
Best Solution: YQINTOJKCEBLDUXMSVPZRHWGAF
Best Solution: YKINTOJZCEBLDUQMSVPXRHWGAF
Best Solution: YPINTOKZCEBLDUQMSVJXRHWGAF
Best Solution: YKINTOZQCEPLDUJMSVXBRHWGAF
```

באופן ממוצע נכונות הדיוק של הפתרון עומד על 86% שזה כ22.33 אותיות נכונות בממוצע בכל

פתרון.
ממוצע הערך המקסימלי של הfitness הוא 932.
מספר הדורות הממוצע הוא 469.
תוצאות ההרצה באופן גרפי:



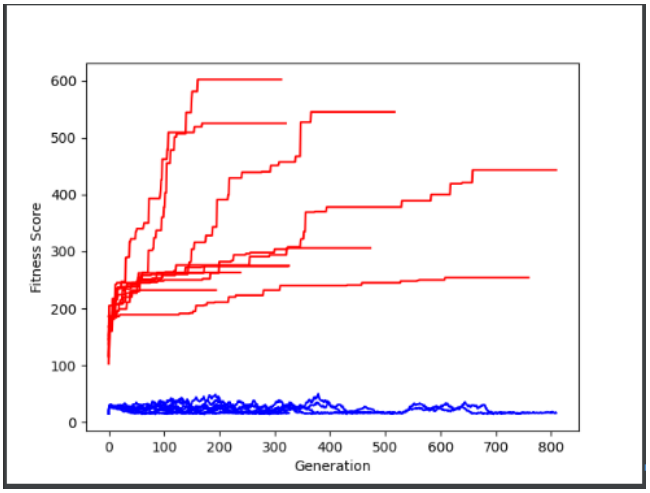
2.<u>אלגוריתם Darwin:</u>

גם פה ביצענו אופוזיציה לוקלית אחת ובדיקה שאכן מדובר באופטימיזציה. במידה וכן, נשמרו השינויים עבור אותו פתרון, אחרת הפתרון חזר לגרסתו המקורית. לפי ציוני הfitness שניתנו לאותן הפתרונות נוצר הדור הבא מהדור הנוכחי <u>שלפני האופטימיזציות</u> על ידי אותם הצעדים שנעשו באלגוריתם המקורי.
*בהתחלה ביצענו 2 אופוזיציות לוקאליות אך קיבלנו ביצועים לא טובים ולכן הקטנו לאופזיציה לוקאלית אחת בלבד. הרעיון היה להגדיל את הדימיון שבין הציון של הפתרון הנוכחי לבין הציון של הפתרון לאחר האופטימיזציה שכן הפתרון הנוכחי הוא זה שיקבע את הדור הבא בפועל והמוטציות לפיהם הציון נקבע לא בהכרח יופיעו בהמשך.
תוצאות -

כל פתרון עבר 2 אופוזיציות לוקליות ובדיקה שאכן מדובר באופטימיזציה. במידה וכן, נשמרו השינויים עבור אותו פתרון, אחרת הפתרון חזר לגרסתו המקורית. לפי ציוני הfitness שניתנו לאותן הפתרונות הדור הבא נוצר מהדור הנוכחי <u>שלאחר האופטימיזציות</u> על ידי אותם הצעדים שנעשו באלגוריתם המקורי.
תוצאות:

```
Best Solution: CWINETMFZRBDGYLKSVUXPAQJHO
Best Solution: YBITADXWPEMFHCVQSNRZLGKJUO
Best Solution: CBINTOYXJEZLDVUMSKPQRHWGAF
Best Solution: YUINTOPLMEBDQGZCSVXKRHWJAF
Best Solution: HSMRATLZXEBCPKUWNVJYFGDQIO
Best Solution: CVIFTOYGZXBEKWLQSMJPNARDUH
Best Solution: OKUPAMJZTEDGNBHLFSQXCRWVIY
Best Solution: YKINTOBGVEXLCUZMSJPQRHWDAF
Best Solution: CVIFOPXJHEBLDRAGSUZYKNWQTM
Best Solution: YXINTOGZBEKLDQJMSVUCRWHPAF
```

באופן ממוצע נכונות הדיוק של הפתרון עומד על 34.6% שזה כ9 אותיות נכונות בממוצע בכל פתרון כאשר השונות מאוד גדולה ונעה בין 3-17 אותיות נכונות.
ממוצע הערך המקסימלי של הfitness הוא 372.
מספר הדורות הממוצע הוא 429.
תוצאות ההרצה באופן גרפי:



<u>סיכום השוואה:</u>

	רגיל	דרווין	למארק
אחוז דיוק בממוצע	74%	34.60%	86%
מספר אותיות שפוענחו נכונה בממוצע	19.2	9	22.3
ערך מקסימלי בממוצע	782	372	932
מספר דורות בממוצע	482	429	469