

טופס דיווח חודשי

נתוני הפרויקט

כיווץ מחרוזות					שם הפרויקט
מיקה עמית וליאת רוזנברג					שם המנחה
adiper8@gmail.com	email	200477842	ת.ז.	עדי פרלוב	שם הסטודנט
mizreut@gmail.com	email	305160079	ת.ז.	רעות מזרחי	שם הסטודנט

סטטוס הפרוייקט לחודש אוגוסט

7/9/2016	תאריך
במהלך חודש אוגוסט הכנו את התוצרים הסופיים של הפרוייקט: ספר פרויקט, מצגת הפרוייקט ואתר לפרוייקט.	סטטוס (מה עשיתי החודש...)
/	בעיות מיוחדות
/	שינויי ל"ז
/	תאריך מפגש עתידי עם המנחה
/	תוכנית עבודה לחודש הקרוב
	הערות
<p><u>סיכום פגישות עם המנחה שנעשו במהלך החודש:</u></p> <p>/</p>	

סטטוס הפרוייקט לחודש יולי

5/8/2016	תאריך
לאחר שהצגנו את המימוש בפני מיקה וליאת, אנו עובדים על התוצרים הסופיים של הפרוייקט להגשה.	סטטוס (מה עשיתי החודש...)
/	בעיות מיוחדות
/	שינויי לו"ז
/	תאריך מפגש עתידי עם המנחה
/	תוכנית עבודה לחודש הקרוב
	הערות
<p><u>סיכום פגישות עם המנחה שנעשו במהלך החודש:</u></p> <p>/</p>	

סטטוס הפרוייקט לחודש יוני

7/7/2016	תאריך
במהלך חודש יוני מימשנו את חלקו השני והאחרון של האלגוריתם LZ77 עם טעויות, בחלק זה אנו מבצעים בדיקה של מספר הפעמים שהמחרוזת הנוכחית קיימת בהמשך ה-buffer, ומוספים לדירוג המחרוזת את מספר פעמים זה. הוספת בדיקה זו מאפשרת כיווץ יעיל יותר מאחר וההנחה היא שאם מחרוזת נמצאה כארוכה ביותר באיטרציה הנוכחית וגם נמצאת מספר פעמים בהמשך, אותה מחרוזת תהווה חלק מהמחרוזת הארוכה ביותר באיטרציות נוספות בהמשך ה-buffer. כמצופה, הוספת חלק זה שיפרה במקרים מסוימים את יחס הכיווץ.	סטטוס (מה עשיתי החודש...)
/	בעיות מיוחדות
/	שינויי ל"ז
/	תאריך מפגש עתידי עם המנחה
/	תוכנית עבודה לחודש הקרוב
	הערות
<p><u>סיכום פגישות עם המנחה שנעשו במהלך החודש:</u></p> <p>במהלך הפגישה עם מיקה וליאת הצגנו את המימוש לאלגוריתם LZ77 עם טעויות על כל חלקיו, הרצנו מספר דוגמאות יחד וניתחנו את אופן הכיווץ, את יעילותו ואת ההשפעה של כל חלק על ה-ratio compression.</p> <p>דנו על העדיפות בחיפוש מחרוזת bit by bit בהמשך ה-buffer.</p> <p>כמו כן, דנו על השיפור בביצועים של האלגוריתם LZ77 עם טעויות.</p>	

סטטוס הפרוייקט לחודש מאי

6/6/2016	תאריך
במהלך חודש מאי מימשנו את חלקו הראשון של האלגוריתם LZ77 עם טעויות, בחלק זה אנו מאפשרים עד K טעויות עבור כל התאמה, האלגוריתם בצעד ה- i מוצא את ההתאמה הארוכה ביותר, מסמן את התו האחרון כטעות, ומדרג את כדאיות המחרוזת הנוכחית כ- $length(longest\ match)/f + 1$ כך ש- f מציין את מספר הטעויות עד כה, על ידי מתן דירוג זה, אנו נותנים עדיפות להתאמות שהכפילו את אורכן מהטעות הקודמת. הכנסת חלק זה בלבד, כצפוי, שיפר את ה- ratio compression במגוון מחרוזות.	סטטוס (מה עשיתי החודש...)
/	בעיות מיוחדות
/	שינויי ל"ז
3/7/16	תאריך מפגש עתידי עם המנחה
בחודש הקרוב, נממש את שלב 2 של האלגוריתם LZ77 (עם טעויות) בו נבדוק את מספר הפעמים שהתאמה הארוכה ביותר קיימת בהמשך ה- buffer, כחלק מההנחה שהתאמה זו תהיה כחלק מהתאמות רבות יותר בהמשך המחרוזת.	תוכנית עבודה לחודש הקרוב
	הערות
<p><u>סיכום פגישות עם המנחה שנעשו במהלך החודש:</u></p> <p>במהלך הפגישה עם מיקה וליאת הצגנו את המימוש לחלקו הראשון של האלגוריתם LZ77 עם טעויות, ערכנו בדיקות על מחרוזות רנדומליות, ומחרוזות תבניות, וראינו את השיפור הצפוי ביחס הכיווץ.</p> <p>בנוסף, דיברנו על מימוש החלק השני ועל אופן חיפוש המחרוזת לקבלת תוצאות אופטימליות.</p>	

סטטוס הפרוייקט לחודש אפריל

6/5/2016	תאריך
<p>במהלך חודש אפריל מימשנו תכנית הפועלת ע"פ האלגוריתם Huffman Encoding, התוכנית מקבלת מחרוזת מא"ב כלשהו, מכווצת את המחרוזת ע"י יצירת עץ על פי האלגוריתם של Huffman ולבסוף מחזירה את מספר הכולל של ביטים בזיכרון על מנת לייצג את המחרוזת המכווצת.</p> <p>התחלנו לממש את האלגוריתם LZ77 (עם טעויות), המימוש יתחלק לשני שלבים:</p> <p>1. כאשר מזהים במחרוזת את המילה הארוכה ביותר (הנוכחית), נאפשר עד K שינויים של ביטים, וכל טעות תקבל דירוג באופן הבא:</p> $(n+1) / (\text{current longest match})$ <p>כאשר n מייצג את מספר הטעויות עד כה. האלגוריתם יקבל את הטעות עם הדירוג הגבוה ביותר – ובהתאם ישנה את buffer.</p> <p>2. בחלק זה, נבחן את האפשרות של שינוי מתן הדירוג באופן הבא:</p> $N + [(n+1) / (\text{current longest match})]$ <p>עד כה N מייצג את מספר הפעמים בו מופיעה המילה הארוכה ביותר עד לסיום המחרוזת.</p> <p>בעזרת שינויים אלה, נקבל ratio compression טוב יותר עבור מחרוזות.</p>	סטטוס (מה עשיתי החודש...)
/	בעיות מיוחדות
/	שינויי לוגיקה
5/6/16	תאריך מפגש עתידי עם המנחה
<p>בחודש הקרוב, נממש את שלב 1 של האלגוריתם LZ77 (עם טעויות) כפי שהוסבר קודם לכן.</p>	תוכנית עבודה לחודש הקרוב
	הערות
<p><u>סיכום פגישות עם המנחה שנעשו במהלך החודש:</u></p> <p>במהלך הפגישה עם מיקה וליאת הצגנו את המימוש לאלגוריתם Huffman Encoding, ודנו על האופן בו מתבצע הכיווץ.</p> <p>בנוסף, הצגנו את המימוש של האלגוריתם LZ77 עם טעויות עד כה, ודנו על התוצאות הצפויות ביחס הכיווץ לאחר שינוי זה.</p>	

סטטוס הפרוייקט לחודש מרץ

7/4/2016	תאריך
במהלך חודש מרץ קראנו את המאמר Huffman Encoding and Data Compression ע"פ בקשת מיקה וליאת, על מנת שנממש תוכנית המקבלת כקלט מחרוזת השייכת לא"ב כלשהו (לאו דווקא בינארית) וע"פ אלגוריתם הדחיסה של Huffman מחשבת את מספר הביטים בזיכרון שצריך על מנת לייצג את המחרוזת הדחוסה. בנוסף, סיימנו לממש את האלגוריתם LZ77 (ללא טעויות), והצגנו את הגרסה למיקה וליאת, עברנו על שלביו, דנו באלגוריתם עצמו ובשינויים הדרושים על מנת לממש את הגרסה הסופית של LZ77 (עם טעויות).	סטטוס (מה עשיתי החודש...)
/	בעיות מיוחדות
/	שינויי ל"ז
1/5/16	תאריך מפגש עתידי עם המנחה
בחודש הקרוב, נכתוב את המימוש לתוכנית שמחזירה כפלט את מספר הביטים לייצוג מחרוזת מכווצת ע"פ האלגוריתם של Huffman, כמו כן, נממש את האלגוריתם LZ77 (עם טעויות) על מנת להביא ליחס כיווץ טוב יותר.	תוכנית עבודה לחודש הקרוב
	הערות
<p><u>סיכום פגישות עם המנחה שנעשו במהלך החודש:</u></p> <p>במהלך הפגישה עם מיקה וליאת הצגנו את המימוש לאלגוריתם LZ77 (ללא טעויות), דנו בשיפור יחס הדחיסה הצפוי במחרוזות בינארית ע"י שימוש באלגוריתם LZ77 עם טעויות, כמו כן, קיבלנו הנחיות למימוש האלגוריתם, ודנו ברעיון בו נוכל להחליט בצורה מדויקת יותר האם תו מסוים במילה בתוך מחרוזת בינארית צריך להיות מסומן כטעות, ע"י כך שנבחן את כמות הפעמים שבו המילה מופיעה בהמשך המחרוזת.</p> <p>הגענו למסקנה שבדיקה כזו אכן תביא ליחס כיווץ טוב יותר, ולכן נוסיף פרמטר זה לשיקול האם לסמן תו כטעות.</p> <p>בנוסף, דנו בתוכנית אשר מחשבת את מספר הביטים לייצוג מחרוזת דחוסה ע"פ האלגוריתם של Huffman וקיבלנו הנחיות לכך.</p>	

סטטוס הפרוייקט לחודש פברואר

5/3/2016	תאריך
<p>במהלך חודש פברואר כתבנו את המימוש לאלגוריתם LZ77 (ללא טעויות) ע"פ ההנחיות שקיבלנו ממיקה וליאת.</p> <p>על מנת לממש את האלגוריתם חזרנו אל המאמר A Universal Algorithm for Sequential Data Compression והעמקנו את הבנתו בשלבי האלגוריתם. בנוסף, התוכנית שאנו כותבים תפיק output כנדרש של הנתונים הבאים:</p> <ol style="list-style-type: none"> 1. המילה הארוכה ביותר עד כה. 2. אורך המילה הארוכה ביותר. 3. ה- offset אל המילה הארוכה ביותר. 4. התו הבא במילה הארוכה ביותר. 5. אורך המחרוזת המכוסה (compressed string) בביטים. 6. Ratio compression <p>כאמור, התוכנית ממומשת בשפת C++ בסביבת הפיתוח Visual studio.</p>	סטטוס (מה עשיתי החודש...)
/	בעיות מיוחדות
/	שינויי ל"ז
30/3/16	תאריך מפגש עתידי עם המנחה
<p>בחודש הקרוב נמשיך לעבוד על מימוש האלגוריתם LZ77 (עם טעויות), במקביל נקרא את המאמר Huffman Encoding and Data Compression על מנת לממש תוכנית המוציאה כפלט את אורך המחרוזת המכוסה לפי האלגוריתם של Huffman</p>	תוכנית עבודה לחודש הקרוב
	הערות
<p style="text-align: center;"><u>סיכום פגישות עם המנחה שנעשו במהלך החודש:</u></p> <p><u>במהלך הפגישה עם מיקה וליאת הצגנו להן את ההתקדמות, ודנו בדרכים שונות למימוש כך שיתאים מבחינת סיבוכיות לאלגוריתם הסופי LZ77 עם טעויות, בנוסף התבקשנו לממש תוכנית המקבל מחרוזת בינארית אקראית ומוציאה כפלט את אורך המחרוזת המכוסה (בביטים) ע"פ Huffman Encoding.</u></p>	

סטטוס הפרוייקט לחודש ינואר

5/2/2016	תאריך
<p>במהלך חודש ינואר העמקנו את הידע בהוכחת אופטימליות קידוד האלגוריתם LZ77 שמתבסס בעיקר על כך שאם נסתכל על המקרה הרנדומלי כך שכל תו במחרוזת (מאלפבית כלשהו עם מספר בדיד של איברים) נבחר בהתפלגות אחידה, כלומר, התו ה-i נבחר בהסתברות p_i שווה לכולם, ומכך נראה שכיווץ האלגוריתם LZ77 נעשה באופן אופטימלי אסימפטומטית.</p> <p>הצגנו את ההוכחה למיקה וליאת בפגישה.</p> <p>התחלנו לכתוב את הגרסה הראשונה של אלגוריתם LZ77 ע"פ ההנחיות שקיבלנו, את התכנית אני כותבים בשפת ++C, סביבת העבודה היא Visual Studio, בגרסה זו הכיווץ הוא ללא טעויות.</p>	סטטוס (מה עשיתי החודש...)
/	בעיות מיוחדות
/	שינויי ל"ז
24/2/16	תאריך מפגש עתידי עם המנחה
בחודש הקרוב אנו נכתוב את המימוש לאלגוריתם LZ77 ללא טעויות ע"פ ההנחיות שניתנו ע"י מיקה וליאת.	תוכנית עבודה לחודש הקרוב
	הערות
<p><u>סיכום פגישות עם המנחה שנעשו במהלך החודש:</u></p> <p><u>במהלך המפגש עם מיקה וליאת הצגנו את הוכחת אופטימליות הכיווץ של האלגוריתם LZ77.</u></p> <p><u>דנו ב- worst case, ובמקרה הרנדומלי שהוזכר קודם וראינו מדוע הכיווץ נעשה באופן אופטימלי בצורה אסימפטומטית.</u></p> <p><u>כמו כן, דיברנו על שיטות שונות למימוש האלגוריתם LZ77, היתרונות והחסרונות של כל שיטה, סיבוכיות מימוש האלגוריתם, וקיבלנו הנחיות ודרישות לפלט הרצוי של הגרסה הראשונה לאלגוריתם LZ77 ללא טעויות.</u></p>	

סטטוס הפרוייקט לחודש דצמבר

3/1/2016	תאריך
<p>A Universal Algorithm for Sequential Data Compression, הבנו את המאמר לעומק, ועמדנו על המשמעות של כיווץ מחרוזות (Compression), וכיווץ מחרוזות בינאריות בפרט, והדרך לשחזר את המחרוזת המכוצת למחרוזת המקורית (Decompression), הבנו את האלגוריתם LZ77 אשר הוא הבסיס לאלגוריתם הסופי של כיווץ מחרוזות עם טעויות, העמקנו ביתרונותיו בהיותו אלגוריתם greedy, סיבוכיות האלגוריתם ובמימושיו השונים. בנוסף, נפגשנו עם מיקה וליאת על מנת לידע אותן על קצב ההתקדמות שלנו ועל כך שבחלק זה נעבוד על הצגת הוכחת אופטימליות הכיווץ של האלגוריתם LZ77.</p>	סטטוס (מה עשיתי החודש...)
/	בעיות מיוחדות
/	שינויי ל"ז
1/2/2016	תאריך מפגש עתידי עם המנחה
<p>בחודש הקרוב אנו נעמיק את הידע בהוכחת אופטימליות הכיווץ של האלגוריתם LZ77, ונציג את ההוכחה בפני מיקה וליאת.</p>	תוכנית עבודה לחודש הקרוב
	הערות
<p style="text-align: right;"><u>סיכום פגישות עם המנחה שנעשו במהלך החודש:</u></p> <p><u>במהלך המפגש עם מיקה וליאת דנו על שלבי ריצת האלגוריתם LZ77 ועל היותו אלגוריתם greedy, דיברנו על חישוב גודל המחרוזת המכוצת, סיבוכיות, ומימוש האלגוריתם.</u></p> <p><u>המשך העמקת הידע בהוכחת אופטימליות הכיווץ של האלגוריתם LZ77 והצגת ההוכחה במפגש הבא.</u></p>	