

תאריך הגשה: 23:55 8/05/2016

תרגיל תכנות 1- קידוד סמבולים

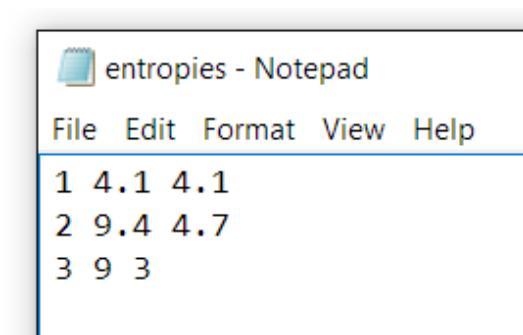
בתרגיל זה אתם תממשו קידוד ופענוח סמבולים. תקבלו קובץ עם טקסט אנגלי (orig.txt) המכיל רק אותיות קטנות a-z. עליכם לדחוס את הקובץ בעזרת קידוד הפמן, כאשר גודל של בלוק יהיה 1, 2 ו-3, וגם בעזרת קוד שנון עם גודל בלוק 1. הוראות הגשה:

1. את הקוד יש לכתוב ב-Mathematica ועליכם להגיש – קובץ עם קוד שכתבתם (קבצי ההרצה המקוריים אשר ניתן להריץ על מנת לבדוק את הקוד), דוח עד 2 עמודים שיכיל סיכום התוצאות (וטבלת הסיכום והסבר מפורט שלה), ואת קבצי הפלט שהתוכנה שכתבתם מייצרת. יש לכתוב הערות בגוף הקוד אשר מסבירות מה תפקיד של כל חלק בו.

2. בתרגיל זה, ההגשה היא בזוגות, כאשר החלוקה לזוגות היא באחריותכם.

עליכם לכתוב 4 פונקציות הבאות:

1. `CalculateEntropies(input_file_name)` – הפונקציה תקבל שם של קובץ המקורי וכפלט תיצור קובץ `entropies.txt` אשר יכיל חישוב של אנטרופיות אמפיריות (במקום הסתברות לוקחים תדירות של הופעת הסמבול (בלוק) בקובץ) עבור שלושת המקרים הבאים: סמבול = אות אחת, סמבול = 2 אותיות, סמבול = 3 אותיות. האנטרופיה תוצג בשתי דרכים: כביטים לסמבול (בלוק) וכביטים לאות (a-z). למשל: (האטרופיות בדוגמא לא בהכרח נכונות)

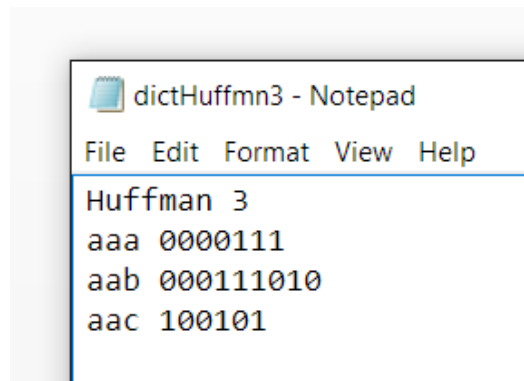


2. `CreateDict(input_file_name)` – הפונקציה תקבל שם של קובץ המקורי וכפלט תיצור 4 קבצים:

- `dictHuffman1.txt` – קובץ המכיל רשימה של מילות קוד הפמן עבור כל סמבול (סמבול = אות אחת).
- `dictHuffman2.txt` – קובץ המכיל רשימה של מילות קוד הפמן עבור כל סמבול (סמבול = 2 אותיות).
- `dictHuffman3.txt` – קובץ המכיל רשימה של מילות קוד הפמן עבור כל סמבול (סמבול = 3 אותיות).

- dictShannon.txt – קובץ המכיל רשימה של מילות קוד שנון עבור כל סמבול (סמבול= אות אחת).

כל קובץ מילון יכיל שתי עמודות. שורה ראשונה: עמודה ראשונה - שם של קידוד: Huffman או Shannon, ועמודה שנייה - גודל הבולוק: 1, 2, או 3. עבור כל שאר השורות: עמודה ראשונה – סמבול ועמודה שנייה - קוד בינארי מתאים. למשל: (הקידוד לא בהכרח נכון בדוגמא)



3. `Compress(input_file_name, dictionary_file_name)` – הפונקציה תקבל שם של קובץ המקורי וקובץ המילון. הפלט של הפונקציה יהיה הקובץ הדחוס עם שם בהתאם למילון: `compressed.h1` / `compressed.h2` / `compressed.h3` / `compressed.sh`.

שימו לב, הקובץ הדחוס צריך להיות קובץ בינארי ולא קובץ טקסט. כלומר את הקוד הבינארי עליכם לרשום כמספרים ולא כאותיות ASCII, כי אחרת על כל ביט 0/1 נשקיע בעצם 8 ביטים של ייצוג ASCII ולא נקבל דחיסה.

4. `Decompress(compressed_file_name, dictionary_file_name)` – הפונקציה תקבל שם של קובץ הדחוס וקובץ המילון. הפלט של הפונקציה יהיה הקובץ המפוענח עם שם: `decompressed.txt`. עליכם לוודא שהקובץ יצא זהה לקובץ המקורי.

סיכום:

1. סכמו את התוצאות בטבלה הבאה:

זמן ריצה של כל אחת מהפונקציות בשניות				אנטרופיה אמפירית (מס' ביטים לאות)	מס' ביטים ממוצע לאות בקובץ הדחוס	גודל של קובץ דחוס (ביטים)	מס' ביטים ממוצע לאות בקובץ מקורי	גודל של קובץ מקורי (ביטים)	שיטת קידוד
ENT	DIC	COM	DEC						
				סמבול = אות אחת					Huffman-1

				סמבול = 2 אותיות					Huffman-2
				סמבול = 3 אותיות					Huffman-3
				סמבול = אות אחת					Shannon
זמן פענוח		זמן קידוד		לא אקטואלי					7-Zip תוכנת

2. הסבירו באופן מפורט את כל התוצאות בטבלה.

בהצלחה!