מתן דנינו – 304802887

שיר אלבז – 204405690

גל ארוס - 204372619

תיאור פרוייקט דחיסת נתונים

הקדמה:

הפרוייקט הוא מימוש של אלגוריתם LZSS שהוא שדרוג של האלגוריתם LZ77, בעוד שני האלגוריתמים עובדים על אותו עיקרון של חלון קדמי ואחורי (ללא מילון), ההבדל העיקרי בינהם הוא האופן בו נכתבת הוראת הפיענוח. אם נסתכל על LZ77 מבנה ההוראה שהוא כותב נשאר קבוע לאורך כל הקידוד ונראה כך <index,length,char> כאשר כמות הביטים לייצוג ה-index תלויה בגודל החלון האחורי אך קבועה, כמות הביטים לייצוג ה-length תלויה בגודל החלון הקדמי אך קבועה וכמות הביטים לייצוג ה-char היא שמונה וגם קבועה, אם נסתכל על LZSS מבנה ההוראה שהוא כותב אינו נשאר קבוע כל הוראה מתחיל בביט שמייצג את סוג ההוראה שתבוא אחריו, אם הביט הוא אפס המשמעות היא שאחרי ביט ההוראה קיים תו לא מקודד ויש לקרא את שמונה הביטים הבאים כמו שהם, אם הביט הוא אחד המשמעות היא שאחרי ביט ההוראה קיים ייצוג של מחרוזת מהחלון האחורי ויש לקרא את האינדקס ואורך המחרוזת, ההוראות שלו ייראו כך: <0,char>,<1,index,len>. מה שגורם להבדל ביחס הדחיסה הוא שאלגוריתם LZ77 "מבזבז" הרבה ביטים כאשר הוא אינו מוצא מחרוזת ובמקרה זה ההוראה שלו נראת כך: <0,0,char> כאשר שני האפסים מבזבזים הרבה ביטים כדי לכתוב תו בודד שאינו מקודד וכמות הביטים המבוזבזים תלויה בגודל החלונות, הפתרון של LZSS לבעיה הוא הביט שמסמן את סוג ההוראה וכך כדי לכתוב תו בודד אינו מקודד הוא משתמש בגודל קבוע של תשעה ביטים. הבדל נוסף הוא ש- LZ77 תמיד מוסיף תו חדש בסוף ההוראה שלו ולפעמים פוגע במציאת התאמה טובה יותר, בעוד LZSS כותב תו בודד רק במקרה והתו מפריע לו למצוא התאמה מספיק טובה.

תיאור הפרוייקט:

בסיס הפרוייקט הוא אלגוריתם LZSS שעובד על כל סוגי הקלטים ויש צורך לפענח את הקובץ כדי לקרא אותו שוב, הפרוייקט מומש בשפת JAVA, הוספנו אפשרות לשנות את הפרמטרים של האלגוריתם בהתאם למידע אותו נרצה לדחוס. ברירת המחדל עבור גודל החלון האחורי היא: 2048Bytes וניתן לשנות אותו בטווח 32Bytes – 32Kb, ברירת המחדל עבור גודל החלון הקדמי שזה בעצם אורך המחרוזת הארוכה ביותר אותה נחפש היא: 32 תווים וניתן לשנות אותו בטווח 4 – 512 תווים, ברירת המחדל עבור אורך המחרוזת המינימלי היא: 3 תווים וניתם לשנות אותו בין 2 – 20 תווים. בנוסף ניתן לבחור להפעיל על הקובץ את האלגוריתמים Move To Front, Delta code, כדי לסדר את המידע באופן שונה ובכך לייעל את הדיחסה בקבצים מסויימים שהיו חלק מקבצי הבדיקה. בכל הפרמטרים ניתן לשלוט דרך ממשק המשתמש הגרפי שכולל שני מצבים: מצב אחד לדחיסה ואחד לפענוח ניתן להחליף בינהם באמצעות כפתורי הניווט העליונים. בתפריט הדחיסה יש צורך לבחור קובץ אותו נרצה לדחוס ונתיב בו נשמור את הקובץ הדחוס, ניתן לעשות זאת בשני דרכים: 1. להכניס נתיב מלא של קובץ המקור וקובץ היעד. 2. לבחור קובץ ממערכת הקבצים בעזרת הכפתור Add File, ברירת המחדל במצב זה לקובץ היעד היא באותה תקייה ושם הקובץ יתחיל ב _comp. בתפריט הפענוח יש צורך לבחור קובץ מקור ויעד ניתן לעשות זאת כמו בתפריט הדחיסה רק ששם הקובץ המפענח אם נבחר קובץ דרך הכפתור Add File יתחיל ב _dec.

הסבר על שלושת המחלקות:

1. **המחלקה Compressor** - היא המחלקה בה מימשנו את כל פונקציות הדחיסה והפענוח של האלגוריתמים LZSS, Delta Code. Move To Front code, המחלקה מכילה פונקציית main שמפעילה את כל אפשרויות המחלקה פירוט דרך הפעלת המain קיים בתיעוד המחלקה. בנוסף ניתן

להפעיל את המחלקה באופן ידני, קיים בנאי שמקבל את גודל החלון האחורי, הקדמי וההתאמה הקטנה ביותר וניתן להפעיל את הפונקציות הציבוריות הבאות:

1. LZSS_compress
2. LZSS_decompress
3. compressMoveToFront
4. decompressMoveToFront
5. compressDelta
6. decompressDelta

כל אחת מהפונקציות מקבלת מערך מחרוזות בשם input שבתא הראשון קיים הנתיב המלא לקובץ הקלט ומערך מחרוזות בשם output שבתא הראשון קיים הנתיב המלא לקובץ הפלט וקוראת לפונקציה הפרטית.

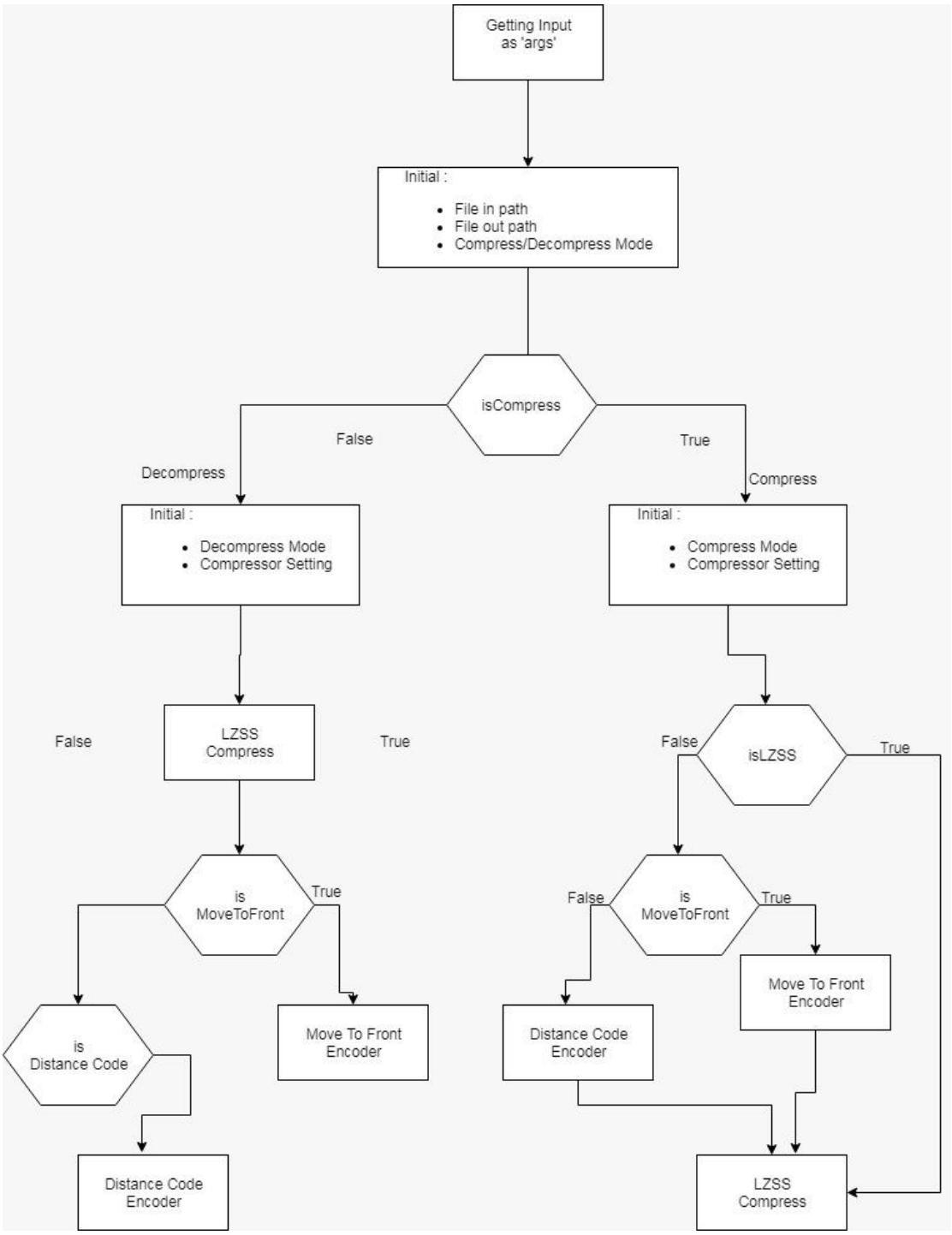
הפונקציות הפרטיות העיקריות במחלקה:

1. CompressLZSS(מצורף תרשים בהמשך)- הפונקציה כותבת את הפרמטרים בתחילת הקובץ, מאתחלת משתנים ומתחילה את הלולאה הראשית שממשיכה כל עוד יש עוד בתים לקרא. כדי למצוא את ההתאמה הטובה ביותר הפונקציה מחזיקה מחרוזת שהיא ההתאמה הטובה ביותר עד עכשיו וכל פעם מוסיפה לה תו עד שאינה מוצאת התאמה בחלון האחורי או שהמחרוזת בגודל המקסימלי, כאשר לא ניתן להגדיל את ההתאמה הפונקציה בודקת אם המחרוזת ארוכה מהאורך המינימלי שהוגדר ואם כן כותבת את ההוראה (1, index, length), אם המחרוזת קצרה מהאורך המינימלי כותבים את ההוראה (0, char) עבור התו הראשון במחרוזת.
2. DecompressLZSS(מצורף תרשים בהמשך)- הפונקציה קוראת מהקובץ את הפרמטרים לגבי גודל החלון ואורכי התאמות ומאתחלת משתנים בהתאם. הלולאה העיקרית מתחילה בקריאת ביט בודד אם ערכו אחד הפונקציה קוראת מהקובץ את האינדקס והאורך לפי כמות הביטים הנדרשת עבור כל אחד ומעתיקה את המחרוזת מהחלון האחורי, אם ערך הביט הוא אפס הפונקציה קוראת את שמונה הביטים הבאים וכותבת אותם

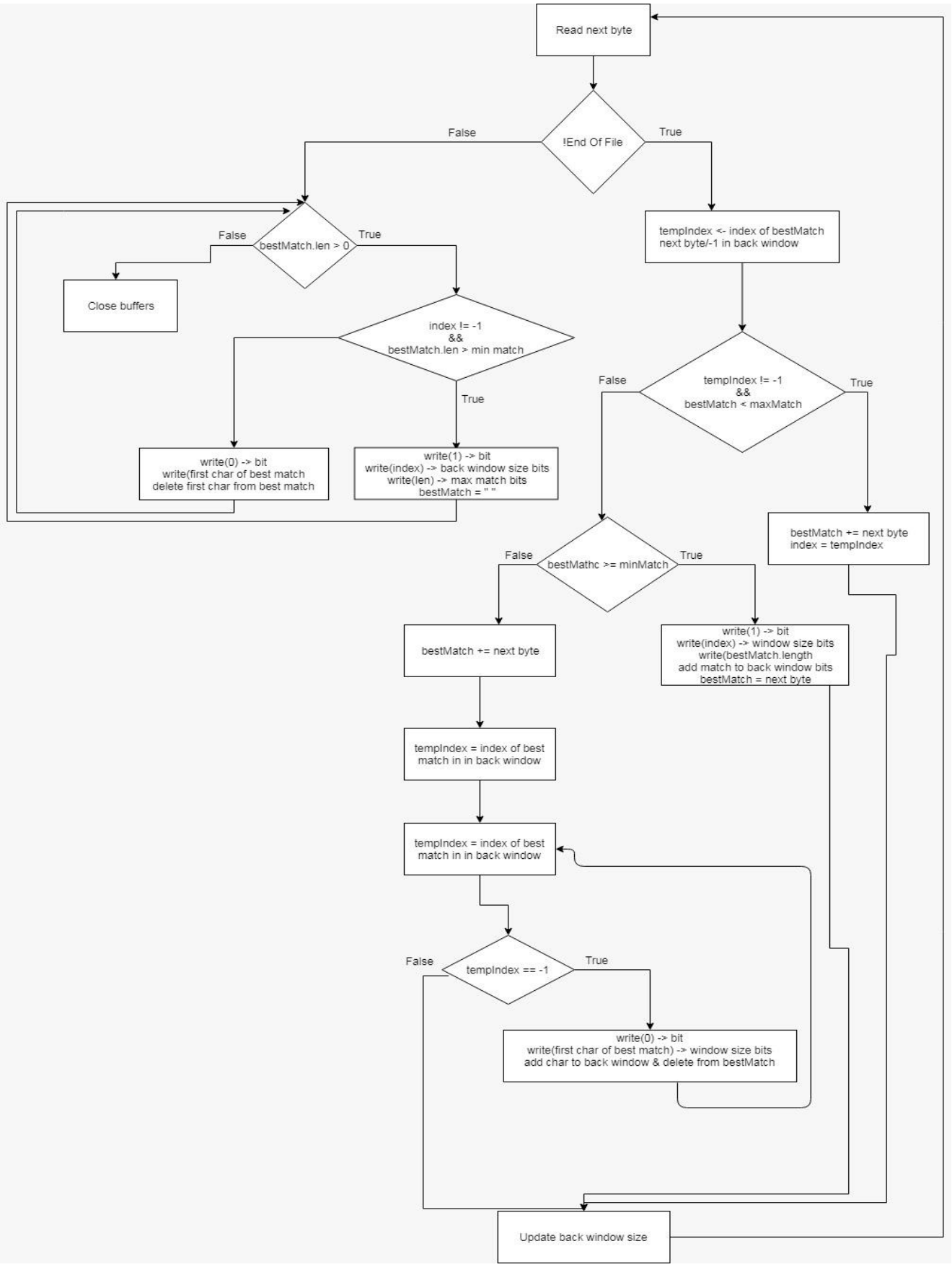
2. **המחלקה BitBuffer** – מימשנו לצורך כתיבה וקריאה נוחה של ביטים בדחיסה ובפענוח.

3. **המחלקה LZSS_Compressor** – מכילה את כל הגדרות ה- GUI

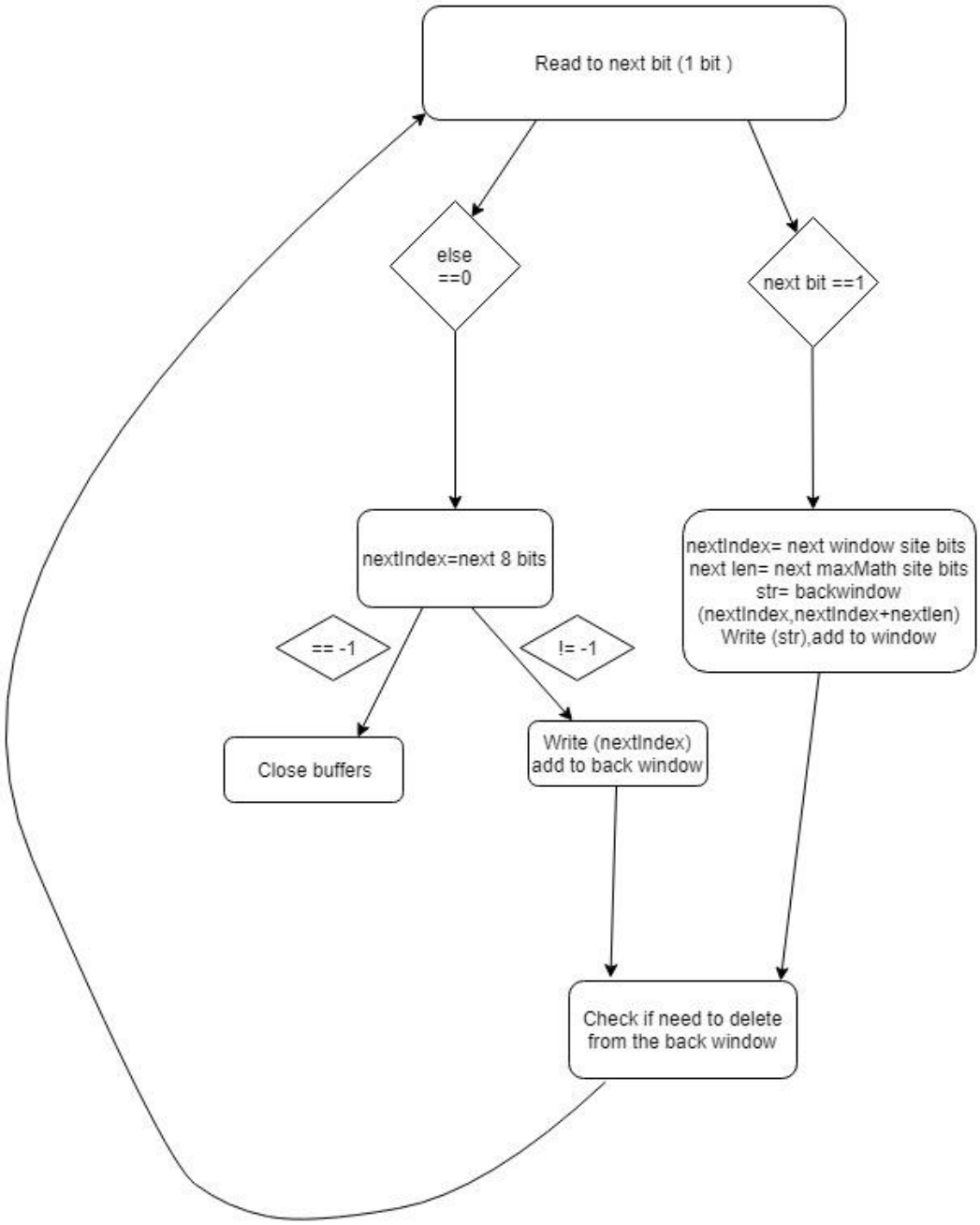
תיאור פונקציית ה- main בתרשים:



תיאור פונקציית הדחיסה של האלגורתים LZSS (LZSS_compress):



תיאור פונקציית הפענוח של אלגוריתם LZSS בתרשים:



בדיקת האלגוריתם על קבצים שונים:

לצורך בדיקת האלגוריתם השתמשתנו בארכיוני קבצים מוכרים אותם ניתן למצוא בקישורים הבאים:

1. http://corpus.canterbury.ac.nz/descriptions/#artificl
2. http://www.data-compression.info/Corpora/index.html

בנוסף לצורך השוואה למימוש שלנו השתמשנו במימוש מהאינטרנט לאלגוריתם LZ77 אותו ניתן למצוא בקישור הבא: https://gist.github.com/fogus/5404660

הבדיקות נעשו עם חלון אחורי בגודל 2048 בתים, חלון קדמי בגודל 32 ואורך התאמה מינימלי 3 (רק LZSS).

The Artificial Corpus

File Name	Size (Bytes)	LZ77	LZSS	LZSS + MTF	LZSS + Delta
a.txt	1	300.00%	600.00%	600.00%	600.00%
aaa.txt	100000	20.60%	6.87%	6.87%	6.87%
alphabet.txt	100000	23.53%	6.89%	6.90%	6.87%
random.txt	100000	100.00%	111.61%	111.54%	112.30%

The Canterbury Corpus

File Name	Size (Bytes)	LZ77	LZSS	LZSS + MTF	LZSS + Delta
alice29.txt	152089	92.34%	53.88%	98.63%	65.87%
asyoulik.txt	125179	94.69%	57.94%	102.44%	70.60%
cp.html	24603	75.00%	47.44%	99.48%	56.15%
fields.c	11150	74.67%	36.17%	87.63%	43.22%
grammar.lsp	3721	77.37%	40.61%	88.42%	46.71%
kennedy.xls	1029744	70.23%	27.95%	15.08%	37.10%
lcet10.txt	426754	91.04%	53.29%	97.13%	65.23%
plrabn12.txt	481861	98.13%	61.16%	99.66%	75.56%
ptt5	513216	36.17%	16.85%	21.42%	18.87%
sum	38240	83.49%	47.66%	71.28%	58.06%
xargs.1	4227	86.02%	52.16%	101.42%	62.17%

The Large Canterbury Corpus

File Name	Size (Bytes)	LZ77	LZSS	LZSS + MTF	LZSS + Delta
bible.txt	4047392	87.08%	45.53%	93.65%	54.75%
E.coli	4638690	97.50%	37.89%	39.01%	46.29%
world192.txt	2473400	91.05%	58.22%	99.59%	71.02%

The Larg Ecalgary Corpus

File Name	Size (Bytes)	LZ77	LZSS	LZSS + MTF	LZSS + Delta
bib	111261	86.64%	54.15%	102.85%	64.85%
book1	768771	97.56%	61.81%	100.04%	76.19%
book2	610856	90.10%	52.65%	97.79%	64.30%
geo	102400	98.92%	84.05%	88.76%	107.79%
news	377109	87.20%	57.43%	99.57%	68.68%
obj1	21504	78.99%	57.28%	85.52%	64.52%
obj2	246814	75.14%	43.77%	84.09%	53.18%
paper1	53161	88.76%	51.81%	99.43%	62.85%
paper2	82199	92.48%	54.26%	99.39%	66.03%
paper3	46526	93.61%	56.88%	100.14%	69.68%
paper4	13286	92.32%	55.60%	100.38%	66.87%
paper5	11954	90.91%	53.89%	99.74%	64.37%
paper6	38105	88.10%	50.78%	96.50%	61.00%
pic	513216	36.17%	16.85%	21.42%	18.87%
progc	39611	85.42%	48.60%	96.99%	58.38%
progl	71646	70.25%	34.42%	79.35%	41.04%

progp	49379	71.60%	34.95%	80.46%	42.15%
trans	93695	73.78%	42.38%	85.09%	49.82%

The Lukas Corpus 2d 8 tif

File Name	Size (Bytes)	LZ77	LZSS	LZSS + MTF	LZSS + Delta
lukas_2d_8_breast_0_t.tif	3752938	60.36%	50.41%	51.41%	45.56%
lukas_2d_8_breast_1_t.tif	3678612	64.88%	57.53%	58.22%	51.00%
lukas_2d_8_foot_0_t.tif	3125062	77.31%	57.16%	59.36%	39.99%
lukas_2d_8_foot_1_t.tif	2235408	95.36%	70.40%	78.82%	49.78%
lukas_2d_8_hand_0_t.tif	2500096	96.12%	63.59%	68.38%	48.63%
lukas_2d_8_hand_1_t.tif	2535246	93.23%	61.69%	66.82%	45.82%
lukas_2d_8_head_0_t.tif	2656678	60.25%	49.59%	51.46%	39.09%
lukas_2d_8_head_1_t.tif	2608068	66.50%	57.37%	59.17%	45.85%
lukas_2d_8_knee_0_t.tif	2701240	85.06%	66.09%	64.93%	44.32%
lukas_2d_8_knee_1_t.tif	2655704	92.04%	68.22%	71.08%	46.09%
lukas_2d_8_leg_0_t.tif	1728972	70.95%	61.10%	63.19%	51.90%
lukas_2d_8_leg_1_t.tif	1318720	81.58%	75.34%	79.56%	61.36%
lukas_2d_8_pelvis_0_t.tif	3124892	98.61%	97.34%	97.31%	79.28%
lukas_2d_8_pelvis_1_t.tif	3034932	95.97%	95.30%	94.62%	78.26%
lukas_2d_8_sinus_0_t.tif	2424218	90.68%	70.61%	78.17%	50.54%
lukas_2d_8_sinus_1_t.tif	2241804	89.41%	75.79%	82.81%	54.79%
lukas_2d_8_spine_0_t.tif	1759608	99.71%	93.43%	93.92%	78.73%
lukas_2d_8_spine_1_t.tif	1786082	99.84%	89.97%	94.35%	76.77%
lukas_2d_8_thorax_0_t.tif	3537852	94.98%	90.56%	88.66%	69.34%
lukas_2d_8_thorax_1_t.tif	2854408	87.08%	81.51%	83.11%	67.31%

The Lukas Corpus 2d 16 dicom

File Name	Size (Bytes)	LZ77	LZSS	LZSS + MTF	LZSS + Delta
lukas_2d_16_breast_0.dcm	8391656	59.03%	38.96%	38.89%	46.32%
lukas_2d_16_breast_1.dcm	8391656	63.65%	43.71%	43.26%	52.32%
lukas_2d_16_food_0.dcm	8391640	70.22%	51.20%	46.71%	58.74%
lukas_2d_16_food_1.dcm	8391638	75.12%	54.04%	48.92%	61.35%
lukas_2d_16_hand_0.dcm	8391632	71.70%	43.98%	41.48%	50.78%
lukas_2d_16_hand_1.dcm	8391636	74.99%	47.98%	43.47%	54.42%
lukas_2d_16_head_0.dcm	6230152	57.57%	39.66%	38.85%	47.21%
lukas_2d_16_head_1.dcm	6230150	70.50%	49.97%	49.00%	59.69%
lukas_2d_16_knee_0.dcm	8391660	71.22%	51.58%	44.78%	57.92%
lukas_2d_16_knee_1.dcm	8391654	71.74%	49.30%	43.70%	55.95%
lukas_2d_16_leg_0.dcm	7534686	48.01%	33.78%	32.68%	38.67%
lukas_2d_16_leg_1.dcm	7534676	45.62%	32.11%	30.91%	36.15%
lukas_2d_16_pelvis_0.dcm	7534634	87.47%	66.46%	65.03%	82.01%
lukas_2d_16_pelvis_1.dcm	7534634	91.10%	69.69%	66.43%	84.48%
lukas_2d_16_sinus_0.dcm	8391632	64.95%	41.97%	40.70%	48.57%
lukas_2d_16_sinus_1.dcm	8391636	63.54%	40.16%	39.02%	46.87%

lukas_2d_16_spine_0.dcm	3768218	95.99%	61.64%	64.77%	76.38%
lukas_2d_16_spine_1.dcm	3768220	98.85%	63.21%	69.39%	78.63%
lukas_2d_16_thorax_0.dcm	7534606	95.93%	70.51%	67.64%	84.55%
lukas_2d_16_thorax_1.dcm	6197038	84.80%	63.01%	61.63%	76.11%

The Protein Corpus

File Name	Size (Bytes)	LZ77	LZSS	LZSS + MTF	LZSS + Delta
hi	509519	99.92%	86.56%	90.59%	103.53%
hs	3295751	99.55%	84.94%	88.18%	101.44%
mj	448779	99.86%	83.72%	88.88%	101.49%
sc	2900352	99.43%	86.14%	89.60%	101.96%

The Miscellaneous Corpus

File Name	Size (Bytes)	LZ77	LZSS	LZSS + MTF	LZSS + Delta
pi.txt	1000000	100.00%	67.82%	67.81%	78.83%

The Silesia Corpus

File Name	Size (Bytes)	LZ77	LZSS	LZSS + MTF	LZSS + Delta
dickens	10192446	95.74%	57.90%	99.44%	71.43%
mozilla	51220480	77.22%	48.36%	73.18%	57.17%
mr	9970564	76.59%	49.18%	52.39%	58.21%
nci	33553445	46.70%	21.28%	31.42%	24.10%
ooffice	6152192	90.11%	64.64%	92.29%	76.20%
osdb	10085684	93.13%	84.86%	104.58%	89.23%
reymont	6627202	87.18%	42.37%	91.59%	51.21%
samba	21606400	67.00%	38.82%	75.56%	44.47%
sao	7251944	99.70%	86.79%	108.39%	98.50%
webster	41458703	77.79%	44.91%	94.79%	54.73%
xml	5345280	52.75%	25.09%	59.77%	29.05%
x-ray	8474240	99.95%	94.71%	84.86%	107.49%