METİN SIKIŞTIRMA PROJESİ

Yağız Türer, Hüseyin Yılmaz
Bilgisayar Mühendisliği Bölümü
Kocaeli Üniversitesi
yagizturer@gmail.com, ylmzhsyn98@gmail.com

Özet

Bu projede bizden sıkıştırılmamış bir metin belgesini LZ77 ve Deflate algoritmalarını kullanarak sıkıştırmamız, iki algoritmayla da sıkıştırılmış yeni verileri farklı dosyalara kaydetmemiz ve iki algoritmanın da sıkıştırma oranlarını tespit edip kıyaslamamız istenmiştir.

Giriş

Çalışmada amaç LZ77 ve Deflate algoritmalarını kullanarak sıkıştırılmamış bir metin belgesini sıkıştırarak bit düzeyinde işlemleri öğrenmemizi sağlamaktı. Projede C dilini kullanmayı tercih ettik.

LZ77 algoritması, daha önceki geçen harf dizisi tekrarlarının konumu ve uzunluğunun kaydedilmesi yoluyla metni sıkıştırmayı amaçlayan bir algoritmadır. Bu algoritmada her bir girdi için sırasıyla en uzun eşleşmenin o anda bulunan harften kaç indis önce başladığını, ardından eşleşmenin uzunluğu, en son da sıradaki harf yazdırılmıştır. Bu algoritma tekrarlı metinlerde boyutu çok fazla düşürebildiği gibi, tekrarın az olduğu metinlerde boyutu arttırabilmektedir.

Deflate algoritması ise LZSS ve Huffman kodlamalarının birleşiminden oluşan bir algoritmadır. Bu algoritmada metin bloklara bölünmektedir. Her bir bloğun Huffman ağaçları birbirinden bağımsızdır. Biz klasik yöntemlerin aksine, doğrudan sayılara Huffman değer vermek yerine, rakamlara Huffman değeri atamayı tercih ettik. Bizim yazdığımız algoritmada her bir rakam için bir Huffman kodu bulunuyor, bu rakamlar sayının basamakları olarak birleştiğinde sayıya ulaşılmış oluyor. Deflate algoritmasında, LZ77'nin aksine sıradaki harfi yazdırmadık, böylece birçok durumda sıkıştırma oranı artıyor. Ayrıca tüm Huffman ağaçlarında aynı rakamlara değer atandığı için bu ağaçların kodlanması da daha kolay bir hâl alıyor. Her bloğun başında o bloğun son blok olup olmadığı bilgisi saklanıyor, böylece decoder kolaylıkla metnin nerede bittiğini anlıyor.

Projede bizden yalnızca sıkıştırma yapmamız istenmişti, sıkıştırılmış dosyaları çözmemiz istenmemişti ama hem yazdığımız programı test etmek hem de kendi yazdığımız algoritmanın çalıştığını kanıtlamak için hem LZ77 hem de Deflate algoritmaları için birer çözücü ekledik.

Program çalıştığında "metin.txt" belgesi açılıyor, önce LZ77, ardından Deflate ile encode ediliyor, iki algoritmanın da sıkıştırma oranları yazdırılıyor, ardından iki çıktı dosyası da kendi decoderi ile çözülüp program sonlandırılıyor.

Yöntem

Programdaki tüm karakter dizilerinin boyutu dinamik olarak tahsis edilmektedir, böylece gereksiz bellek kullanımından kaçınıldı ve girdi boyutları ne kadar yüksek olursa olsun programın kayıpsız sıkıştırma yapması sağlandı. Dinamik bellek tahsisi için mallocla önce boyut 0 olarak atandı, daha sonra her gelen harfte realloc kullanılarak boyut 1 arttırıldı.

LZ77'yi encode etmek için kullandığımız fonksiyon, arama ve ileri tampon boyutlarını dışarıdan almaktadır. Boyutu küçültmek için eşleşme uzunluğu unsigned char olarak tanımlandı, eşleşme konumu ise short int olarak tanımlandı. Yani ileri tampon uzunluğu en fazla 255 olabiliyorken, arama tamponu uzunluğu ise en fazla 32767 olmaktadır. İmleç başta 0. indise götürülüyor, daha sonra en uzun eşleşmeden dönen sayının bir fazlası kadar ilerletiliyor. En uzun eşleşmeyi bulan fonksiyonun çalışma mantığı ise, önce tampon başlangıcının tespit edilmesi, daha sonra ileri tampon başlangıcının imleç yeri olarak alınması ve iki indisteki metin değerlerinin karşılaştırılması şeklindedir. Arama tamponu ve ileri tampon için her seferinde ayrı bir string oluşturulduğunda çalışma süresi çok fazla uzadığı için böyle bir yönteme başvuruldu. Bu işlem sonucunda en fazla eşleşme bulunan indis ve eşleşmenin uzunluğu pointer

olarak döndürülmektedir. Daha sonra binary yazma modunda açtığımız dosyaya sırasıyla 2 bayt olarak eşleşme konumu, 1 bayt olarak eşleşme uzunluğu, 1 bayt olarak da sıradaki harf yazdırılıyor. Her bir yazdırma sonrasında toplam boyuta 4 ekleniyor ve en sonunda sıkıştırma oranı hesaplanıyor.

LZ77 decode edilirken de yine boyutu dinamik olarak tahsis edilen bir metin stringi oluşturuluyor, daha sonra dosyadaki girdiler bitene kadar 2 bayt eşleşme konumu, 1 bayt eşleşme uzunluğu, 1 bayt da sıradaki harf okunuyor, bu okunan bilgiler stringe ekleniyor ve en sonunda oluşan bu metin bir dosyaya yazdırılıyor.

Deflate algoritmasında bitlerle işlem yapıldığı için dosyaya yazdırma işlemi ayrı bir fonksiyonla sağlandı. Bu fonksiyon dışarıdan o anki indisi almaktadır. Bir baytın ilk indisindeki değer 2^7'ye denk geldiği için indis 7'den başlıyor ve 0'a kadar gidiyor. İndis 0 olduğunda baytın tamamı dolduğundan dolayı, o bayt dosyaya yazdırılıyor ve bayt sıfırlanıyor, indis de tekrar 7'ye eşitleniyor, böylece bit yazdırma yaparken başka bir işleme gerek kalmıyor.

Deflate algoritmasında kullanmak üzere huffman nodelarını tutacak bir struct tanımladık. Bu struct, node'un ağırlığını, tuttuğu veriyi ve sağ ve sol dallarını kapsıyor. Huffman nodeları 0'dan 9'a rakamları tuttuğu için iki huffman dizisinin de node'u 10 olarak atandı. Başta her rakamın ağaırlığı 0 olarak belirleniyor. Daha sonra karşılaşılan sayıların basamaklarından rakamlara denk gelen node'un ağırlığı arttırılıyor. Blok bittiğinde de her seferinde nodelar sıralanıyor ve ağırlığı en küçük olan iki node birleştiriliyor. Geriye son bir node kaldığında bu node ağacın başlangıcı oluyor. İki huffman ağacı için de bu işlemler tamamlandıktan sonra blok dosyaya yazdırılmaya başlanıyor.

Her bloğun ilk biti, bu bloğun son blok olup olmadığı bilgisini tutuyor, ilk bit 0'sa bundan başka bloklar var, 1'se bu blok son blok demektir. Daha sonra da bloğun dinamik Huffman kodlamasıyla sıkıştırıldığını belirtmek için sırasıyla 1 ve 0 bitleri yazdırılıyor. Sonrasında ise 10 bit boyunca blokta kaç adet girdi olduğu bilgisi kaydediliyor. Bu yüzden yazdığımız algoritmada blok uzunluğu en fazla 1023 olmaktadır. Daha sonra ise oluşturulan Huffman ağaçları yazdırılıyor. Burada önce birinci ağaç için daha sonra da ikinci ağaç için 0'dan 9'a her sayının sırasıyla 4 bit olarak kod uzunluğu ardından da kodu yazdırılmaktadır.

Daha sonra girdiler yazdırılıyor. Eğer ki girdinin eşleşme uzunluğu 3'ten küçükse harf olduğunu belirtmek için 1 yazılıyor. Daha sonra da char değeri basamaklarına ayrılıp, her bir basamağın huffman kodu yazdırılıyor. Sayının kaç basamaklı olduğu bilinmediği için her basamağın başına 0 veya 1 konuyor, eğer 0 konmuşsa son basamak değil demektir.

Eşleşme uzunluğunun 3'ten büyük olduğu durumlarda ise önce birinci Huffman ağacındaki kodlar kullanılarak eşleşme uzunluğu yazdırılıyor, daha sonra da ikinci huffman ağacındaki kodlarla eşleşme konumu yazdırılıyor.

Bloktaki son girdi de dosyaya yazdırıldıktan sonra içinde bulunduğumuz byte'ın geri kalan tüm bitleri 0 ile dolduruluyor, böylece sıradaki blok yeni bir byte'dan başlıyor.

En son da yazdırılan dosyadaki sıkıştırma oranı ekrana yazdırılıyor.

Deflate dosyası decode edilirken de yazdırma sırasına göre bitler okunmakta, her bir huffman ağacı oluşturularak çözülmektedir.

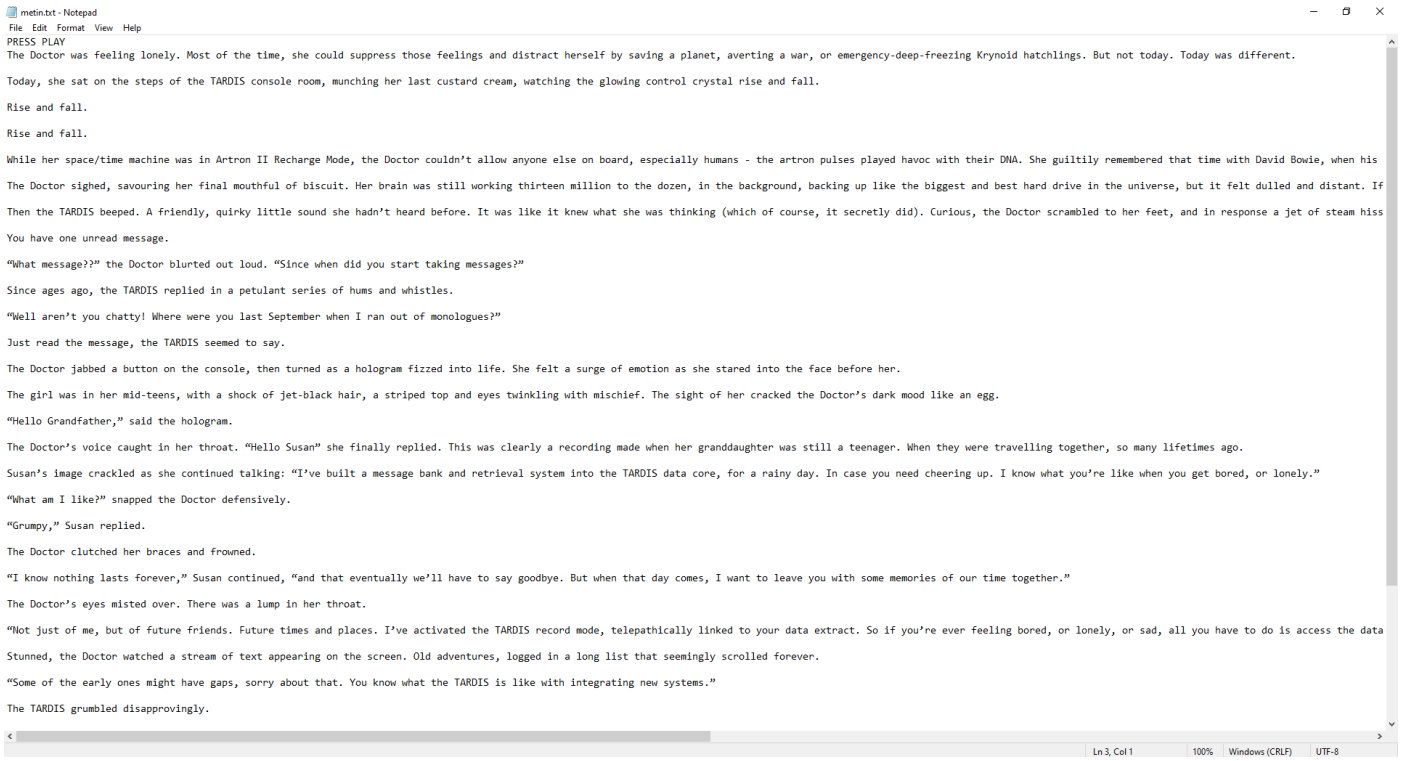
Sonuçlar

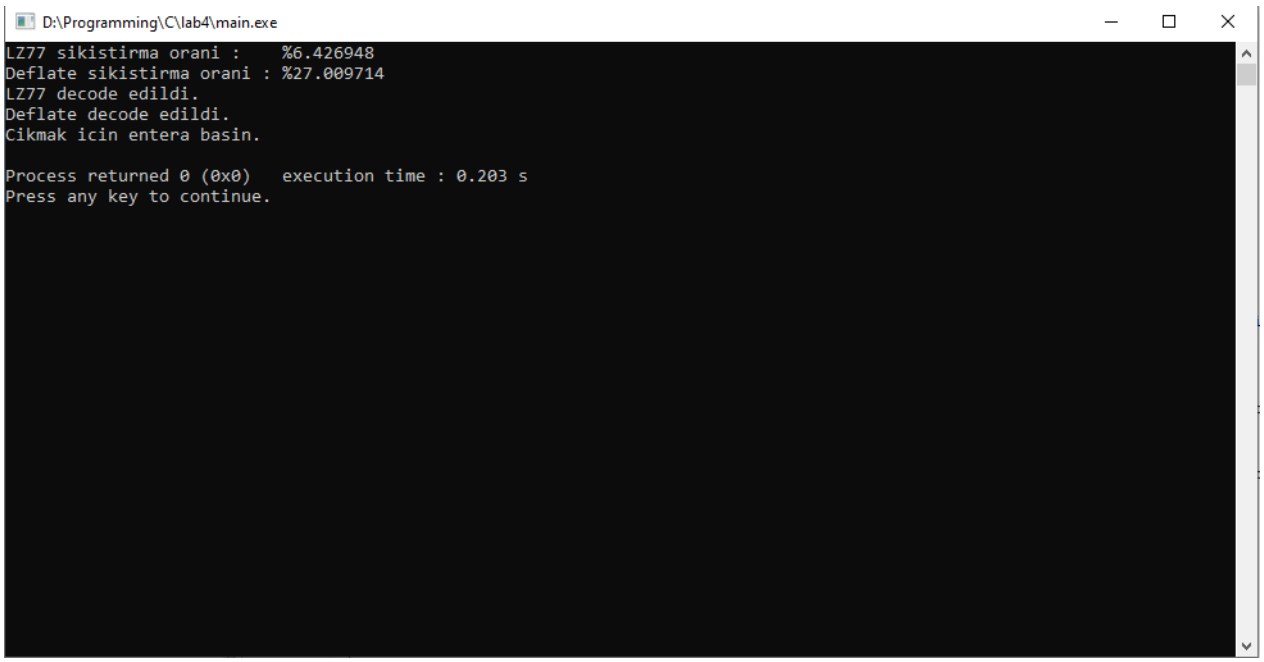
Tüm bu işlemler sonucunda verilen metin belgesi hem LZ77 hem de bizim uyarladığımız Deflate algoritmasıyla sıkıştırılmış ve sıkıştırma oranları yazdırılmıştır. Ve sıkıştırılan dosyalar decode edilerek sıkıştırma işleminin düzgün yapıldığı doğrulanmıştır.

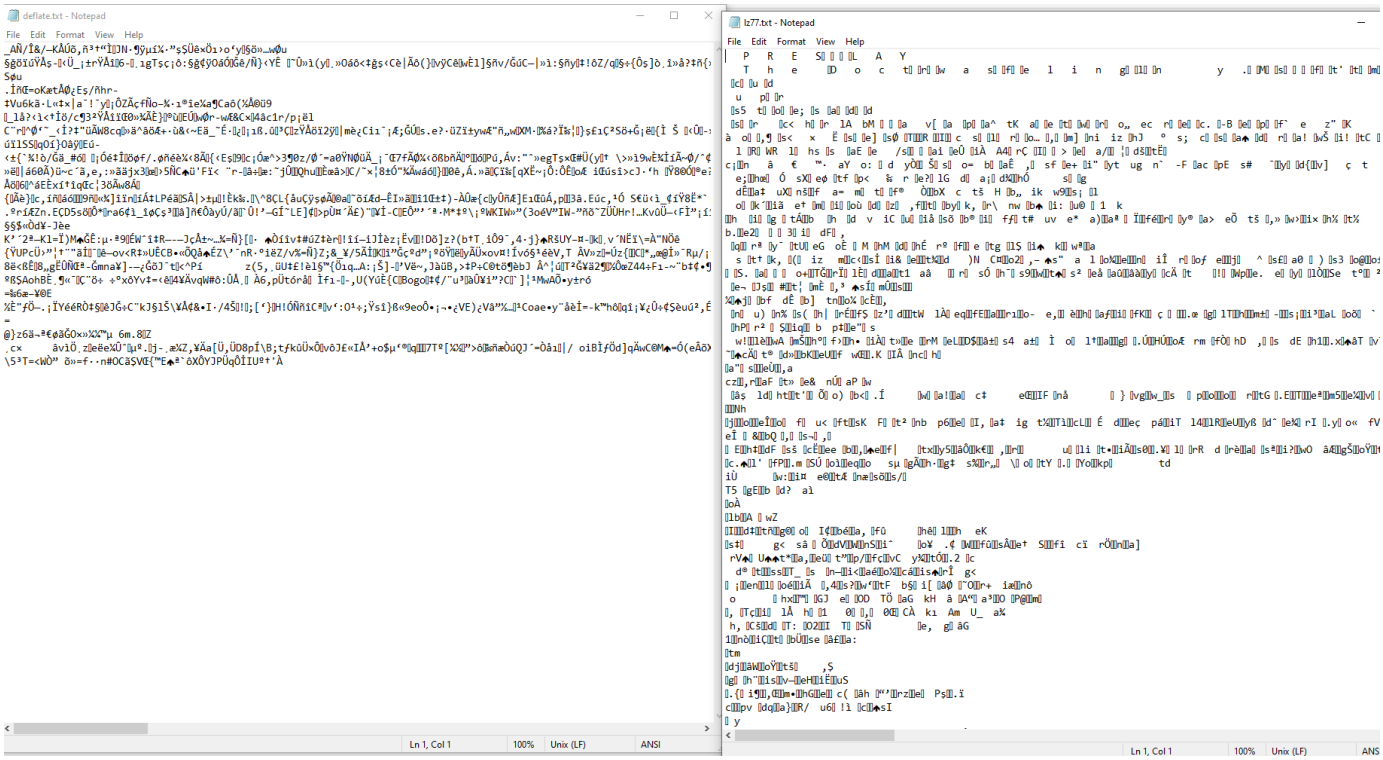
Program hem Windows hem de Linux işletim sistemlerinde test edildi ve metin belgesinin uzunluğu ne olursa olsun sorunsuz çalıştığı tespit edildi.

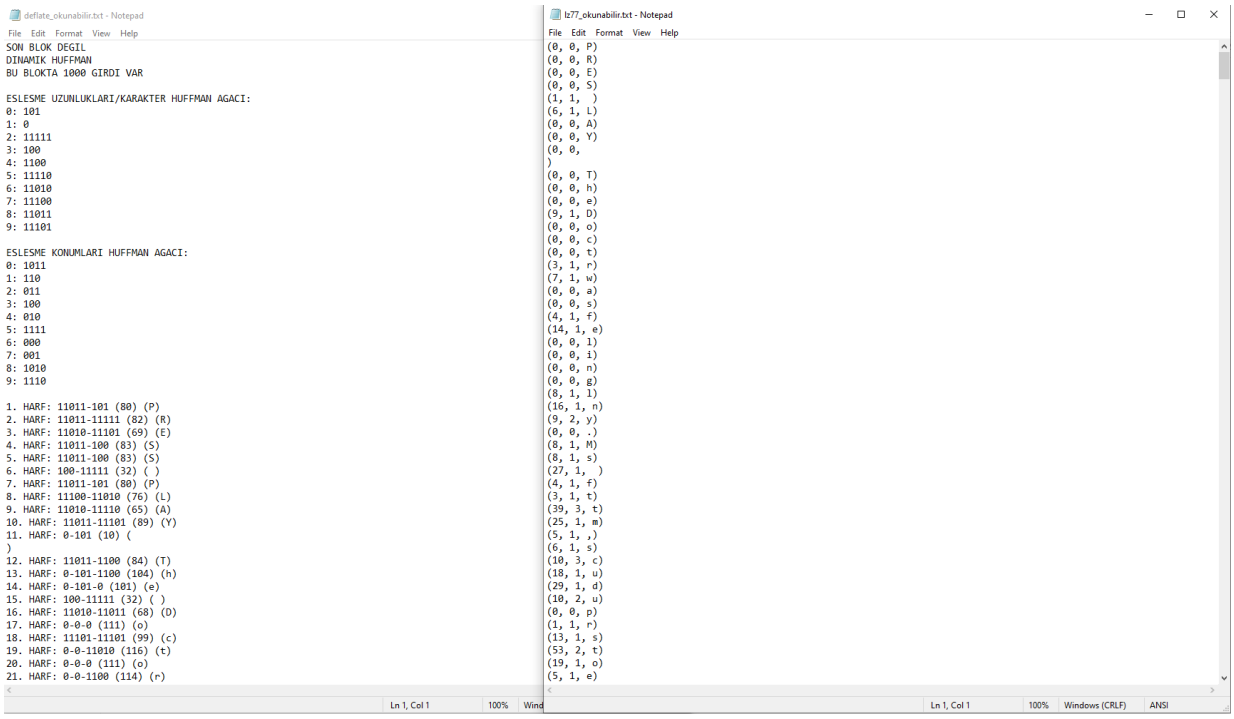
Decode edilmiş çıktılar orijinal metinle kıyaslandı ve arada bir fark olmadığı görüldü.

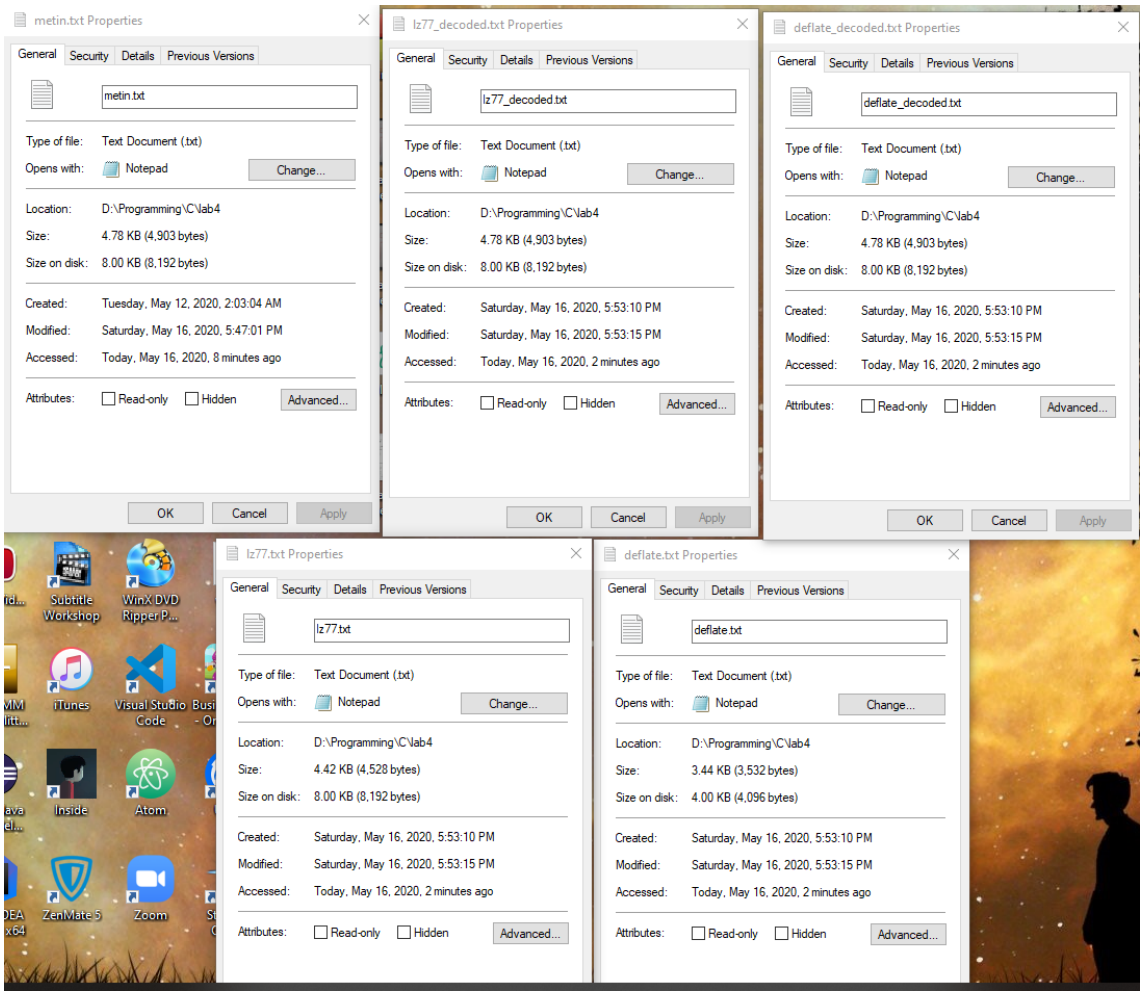
Çıktılar


metin.txt - Notepad
File Edit Format View Help
PRESS PLAY
The Doctor was feeling lonely. Most of the time, she could suppress those feelings and distract herself by saving a planet, averting a war, or emergency-deep-freezing Krynoid hatchlings. But not today. Today was different.
Today, she sat on the steps of the TARDIS console room, munching her last custard cream, watching the glowing control crystal rise and fall.
Rise and fall.
Rise and fall.
While her space/time machine was in Artron II Recharge Mode, the Doctor couldn't allow anyone else on board, especially humans - the artron pulses played havoc with their DNA. She guiltily remembered that time with David Bowie, when his
The Doctor sighed, savouring her final mouthful of biscuit. Her brain was still working thirteen million to the dozen, in the background, backing up like the biggest and best hard drive in the universe, but it felt dulled and distant. If
Then the TARDIS beeped. A friendly, quirky little sound she hadn't heard before. It was like it knew what she was thinking (which of course, it secretly did). Curious, the Doctor scrambled to her feet, and in response a jet of steam hiss
You have one unread message.
"What message??" the Doctor blurted out loud. "Since when did you start taking messages?"
Since ages ago, the TARDIS replied in a petulant series of hums and whistles.
"Well aren't you chatty! Where were you last September when I ran out of monologues?"
Just read the message, the TARDIS seemed to say.
The Doctor jabbed a button on the console, then turned as a hologram fizzed into life. She felt a surge of emotion as she stared into the face before her.
The girl was in her mid-teens, with a shock of jet-black hair, a striped top and eyes twinkling with mischief. The sight of her cracked the Doctor's dark mood like an egg.
"Hello Grandfather," said the hologram.
The Doctor's voice caught in her throat. "Hello Susan" she finally replied. This was clearly a recording made when her granddaughter was still a teenager. When they were travelling together, so many lifetimes ago.
Susan's image crackled as she continued talking: "I've built a message bank and retrieval system into the TARDIS data core, for a rainy day. In case you need cheering up. I know what you're like when you get bored, or lonely."
"What am I like?" snapped the Doctor defensively.
"Grumpy," Susan replied.
The Doctor clutched her braces and frowned.
"I know nothing lasts forever," Susan continued, "and that eventually we'll have to say goodbye. But when that day comes, I want to leave you with some memories of our time together."
The Doctor's eyes misted over. There was a lump in her throat.
"Not just of me, but of future friends. Future times and places. I've activated the TARDIS record mode, telepathically linked to your data extract. So if you're ever feeling bored, or lonely, or sad, all you have to do is access the data
Stunned, the Doctor watched a stream of text appearing on the screen. Old adventures, logged in a long list that seemingly scrolled forever.
"Some of the early ones might have gaps, sorry about that. You know what the TARDIS is like with integrating new systems."
The TARDIS grumbled disapprovingly.
Ln 3, Col 1
100%
Windows (CRLF)
UTF-8

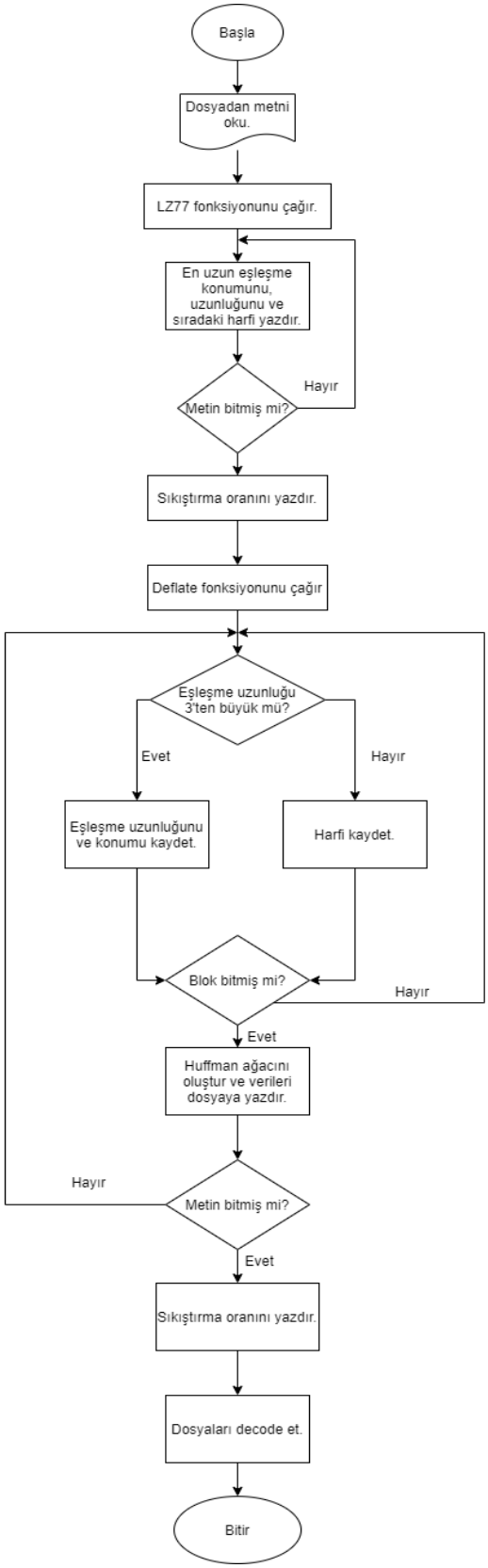

D:\Programming\C\lab4\main.exe
LZ77 sikistirma orani : %6.426948
Deflate sikistirma orani : %27.009714
LZ77 decode edildi.
Deflate decode edildi.
Cikmak icin entera basin.
Process returned 0 (0x0) execution time : 0.203 s
Press any key to continue.


deflate.txt - Notepad
File Edit Format View Help
lz77.txt - Notepad
File Edit Format View Help
lz77_decoded.txt - Notepad
File Edit Format View Help
PRESS PLAY
The Doctor was feeling lonely. Most of the time, she could suppress those feelings and distract herself by saving a p
Today, she sat on the steps of the TARDIS console room, munching her last custard cream, watching the glowing control
Rise and fall.
Rise and fall.
While her space/time machine was in Artron II Recharge Mode, the Doctor couldn't allow anyone else on board, especial
The Doctor sighed, savouring her final mouthful of biscuit. Her brain was still working thirteen million to the dozen
Then the TARDIS beeped. A friendly, quirky little sound she hadn't heard before. It was like it knew what she was thi
You have one unread message.
"What message??" the Doctor blurted out loud. "Since when did you start taking messages?"
Since ages ago, the TARDIS replied in a petulant series of hums and whistles.
"Well aren't you chatty! Where were you last September when I ran out of monologues?"
Just read the message, the TARDIS seemed to say.
The Doctor jabbed a button on the console, then turned as a hologram fizzed into life. She felt a surge of emotion as
The girl was in her mid-teens, with a shock of jet-black hair, a striped top and eyes twinkling with mischief. The si
"Hello Grandfather," said the hologram.
The Doctor's voice caught in her throat. "Hello Susan" she finally replied. This was clearly a recording made when he
Susan's image crackled as she continued talking: "I've built a message bank and retrieval system into the TARDIS data
"What am I like?" snapped the Doctor defensively.
"Grumpy," Susan replied.
The Doctor clutched her braces and frowned.
"I know nothing lasts forever," Susan continued, "and that eventually we'll have to say goodbye. But when that day co
The Doctor's eyes misted over. There was a lump in her throat.
"Not just of me, but of future friends. Future times and places. I've activated the TARDIS record mode, telepathicall
Stunned, the Doctor watched a stream of text appearing on the screen. Old adventures, logged in a long list that seem
"Some of the early ones might have gaps, sorry about that. You know what the TARDIS is like with integrating new syst
Ln 1, Col 1
100%
Windows (CRLF)
UTF-8
deflate_decoded.txt - Notepad
File Edit Format View Help
PRESS PLAY
The Doctor was feeling lonely. Most of the time, she could suppress those feelings and distract herself by savi
Today, she sat on the steps of the TARDIS console room, munching her last custard cream, watching the glowing c
Rise and fall.
Rise and fall.
While her space/time machine was in Artron II Recharge Mode, the Doctor couldn't allow anyone else on board, es
The Doctor sighed, savouring her final mouthful of biscuit. Her brain was still working thirteen million to the
Then the TARDIS beeped. A friendly, quirky little sound she hadn't heard before. It was like it knew what she w
You have one unread message.
"What message??" the Doctor blurted out loud. "Since when did you start taking messages?"
Since ages ago, the TARDIS replied in a petulant series of hums and whistles.
"Well aren't you chatty! Where were you last September when I ran out of monologues?"
Just read the message, the TARDIS seemed to say.
The Doctor jabbed a button on the console, then turned as a hologram fizzed into life. She felt a surge of emot
The girl was in her mid-teens, with a shock of jet-black hair, a striped top and eyes twinkling with mischief.
"Hello Grandfather," said the hologram.
The Doctor's voice caught in her throat. "Hello Susan" she finally replied. This was clearly a recording made w
Susan's image crackled as she continued talking: "I've built a message bank and retrieval system into the TARDI
"What am I like?" snapped the Doctor defensively.
"Grumpy," Susan replied.
The Doctor clutched her braces and frowned.
"I know nothing lasts forever," Susan continued, "and that eventually we'll have to say goodbye. But when that
The Doctor's eyes misted over. There was a lump in her throat.
"Not just of me, but of future friends. Future times and places. I've activated the TARDIS record mode, telepat
Stunned, the Doctor watched a stream of text appearing on the screen. Old adventures, logged in a long list tha
"Some of the early ones might have gaps, sorry about that. You know what the TARDIS is like with integrating ne
Ln 1, Col 1
100%
Windows (CRLF)
UTF-8


deflate_okunabilir.txt - Notepad
File Edit Format View Help
SON BLOK DEGIL
DINAMIK HUFFMAN
BU BLOKTA 1000 GIRDI VAR
ESLESME UZUNLUKLARI/KARAKTER HUFFMAN AGACI:
0: 101
1: 0
2: 11111
3: 100
4: 1100
5: 11110
6: 11010
7: 11100
8: 11011
9: 11101
ESLESME KONUMLARI HUFFMAN AGACI:
0: 1011
1: 110
2: 011
3: 100
4: 010
5: 1111
6: 000
7: 001
8: 1010
9: 1110
1. HARF: 11011-101 (80) (P)
2. HARF: 11011-11111 (82) (R)
3. HARF: 11010-11101 (69) (E)
4. HARF: 11011-100 (83) (S)
5. HARF: 11011-100 (83) (S)
6. HARF: 100-11111 (32) ( )
7. HARF: 11011-101 (80) (P)
8. HARF: 11100-11010 (76) (L)
9. HARF: 11010-11110 (65) (A)
10. HARF: 11011-11101 (89) (Y)
11. HARF: 0-101 (10) (
)
12. HARF: 11011-1100 (84) (T)
13. HARF: 0-101-1100 (104) (h)
14. HARF: 0-101-0 (101) (e)
15. HARF: 100-11111 (32) ( )
16. HARF: 11010-11011 (68) (D)
17. HARF: 0-0-0 (111) (o)
18. HARF: 11101-11101 (99) (c)
19. HARF: 0-0-11010 (116) (t)
20. HARF: 0-0-0 (111) (o)
21. HARF: 0-0-1100 (114) (r)
Ln 1, Col 1
100%
lz77_okunabilir.txt - Notepad
File Edit Format View Help
(0, 0, P)
(0, 0, R)
(0, 0, E)
(0, 0, S)
(1, 1,)
(6, 1, L)
(0, 0, A)
(0, 0, Y)
(0, 0,
)
(0, 0, T)
(0, 0, h)
(0, 0, e)
(9, 1, D)
(0, 0, o)
(0, 0, c)
(0, 0, t)
(3, 1, r)
(7, 1, w)
(0, 0, a)
(0, 0, s)
(4, 1, f)
(14, 1, e)
(0, 0, l)
(0, 0, i)
(0, 0, n)
(0, 0, g)
(8, 1, l)
(16, 1, n)
(9, 2, y)
(0, 0, .)
(8, 1, M)
(8, 1, s)
(27, 1,)
(4, 1, f)
(3, 1, t)
(39, 3, t)
(25, 1, m)
(5, 1, ,)
(6, 1, s)
(10, 3, c)
(18, 1, u)
(29, 1, d)
(10, 2, u)
(0, 0, p)
(1, 1, r)
(13, 1, s)
(53, 2, t)
(19, 1, o)
(5, 1, e)
Ln 1, Col 1
100%
Windows (CRLF)
ANSI


metin.txt Properties
General Security Details Previous Versions
metin.txt
Type of file: Text Document (.txt)
Opens with: Notepad
Change...
Location: D:\Programming\C\lab4
Size: 4.78 KB (4,903 bytes)
Size on disk: 8.00 KB (8,192 bytes)
Created: Tuesday, May 12, 2020, 2:03:04 AM
Modified: Saturday, May 16, 2020, 5:47:01 PM
Accessed: Today, May 16, 2020, 8 minutes ago
Attributes: Read-only Hidden Advanced...
OK Cancel Apply
lz77_decoded.txt Properties
General Security Details Previous Versions
lz77_decoded.txt
Type of file: Text Document (.txt)
Opens with: Notepad
Change...
Location: D:\Programming\C\lab4
Size: 4.78 KB (4,903 bytes)
Size on disk: 8.00 KB (8,192 bytes)
Created: Saturday, May 16, 2020, 5:53:10 PM
Modified: Saturday, May 16, 2020, 5:53:15 PM
Accessed: Today, May 16, 2020, 2 minutes ago
Attributes: Read-only Hidden Advanced...
OK Cancel Apply
deflate_decoded.txt Properties
General Security Details Previous Versions
deflate_decoded.txt
Type of file: Text Document (.txt)
Opens with: Notepad
Change...
Location: D:\Programming\C\lab4
Size: 4.78 KB (4,903 bytes)
Size on disk: 8.00 KB (8,192 bytes)
Created: Saturday, May 16, 2020, 5:53:10 PM
Modified: Saturday, May 16, 2020, 5:53:15 PM
Accessed: Today, May 16, 2020, 2 minutes ago
Attributes: Read-only Hidden Advanced...
OK Cancel Apply
Subtitle Workshop
WinX DVD Ripper P...
iTunes
Visual Studio Code
Inside
Atom
ZenMate 5
Zoom
lz77.txt Properties
General Security Details Previous Versions
lz77.txt
Type of file: Text Document (.txt)
Opens with: Notepad
Change...
Location: D:\Programming\C\lab4
Size: 4.42 KB (4,528 bytes)
Size on disk: 8.00 KB (8,192 bytes)
Created: Saturday, May 16, 2020, 5:53:10 PM
Modified: Saturday, May 16, 2020, 5:53:15 PM
Accessed: Today, May 16, 2020, 2 minutes ago
Attributes: Read-only Hidden Advanced...
deflate.txt Properties
General Security Details Previous Versions
deflate.txt
Type of file: Text Document (.txt)
Opens with: Notepad
Change...
Location: D:\Programming\C\lab4
Size: 3.44 KB (3,532 bytes)
Size on disk: 4.00 KB (4,096 bytes)
Created: Saturday, May 16, 2020, 5:53:10 PM
Modified: Saturday, May 16, 2020, 5:53:15 PM
Accessed: Today, May 16, 2020, 2 minutes ago
Attributes: Read-only Hidden Advanced...

Akış Şemaları:



Şema çizimleri draw.io üzerinden yapılmıştır.

Kaynakça:

https://www.tutorialspoint.com/c_standard_library/string_h.htm

https://www.tutorialspoint.com/c_standard_library/c_function_sprintf.htm

https://www.geeksforgeeks.org/bitwise-operators-in-c-cpp/

https://www.tutorialspoint.com/cprogramming/c_bitwise_operators.htm

https://www.tutorialspoint.com/c_standard_library/c_function_getc.htm

Handbook of Data Compression - David Salomon and Giovanni Motta

https://tools.ietf.org/html/rfc1951

https://zlib.net/feldspar.html

https://en.wikipedia.org/wiki/LZ77_and_LZ78

https://towardsdatascience.com/how-data-compression-works-exploring-lz77-3a2c2e06c097

https://cs.stanford.edu/people/eroberts/courses/soco/projects/data-compression/lossless/lz77/example.htm

https://www.geeksforgeeks.org/huffman-coding-greedy-algo-3/