

Komprimering

Oppgave 8 – IDATT2101

Innlevering av:

Nicolai Thorer Sivesind
Erlend Rønning
Aleksander Brekke Røed

Kjøring av filer

```
(base) Nicolais-MacBook-Pro:src nicolaisivesind$ java Main.java c diverse-txt.txt
Komprimering fullført
(base) Nicolais-MacBook-Pro:src nicolaisivesind$ java Main.java d diverse-txt.txt
Utpakking fullført
```

Skriv «c <originalfilnavn>» for å komprimere og «d <originalfilnavn>» for å pakke ut.

Lempel-Ziv

Teori

Lempel-ziv bruker referanser til sekvenser den kjenner igjen fra tidligere til å komprimere filer. Hvis algoritmen oppdager at en sekvens med tre eller fler bytes repeterer seg fra tidligere i teksten den har analysert, kan den spare plass ved å bruke to bytes for å referere til denne sekvensen fremfor å lagre bytesene som de er. Den første byten for å fortelle hvor den refererte sekvensen starter og den andre for hvor lang den er. Vi kan øke dette til å bruke to bytes for avstand og en byte for lengde. Da kan vi øke avstanden algoritmen kan se bakover for å sjekke etter repetert sekvens til 2^{15} , sammenlignet med 2^7 ved kun en byte. Vi ønsker kun å referere til denne hvis sekvensen er på 3 bytes eller mer. Dette er fordi at vi bruker 2 bytes for å lagre referansen uansett, så en referanse til en sekvens som har lengde på 2 eller mindre vil føre til å at vi bruker like mye eller mer plass enn hva den opprinnelige filen gjorde. Hvis vi setter at algoritmen også skal referere til sekvenser med lengde 2, risikerer vi å splitte en sekvens med ukomprimerte bytes (som ellers hadde vært en lang sekvens) i to, og dermed totalt sett bruke flere bytes på å representere sekvenser med ukomprimerte bytes. Dette til tross for at en referanse til en sekvens på to tegn i seg selv verken sparer eller bruker flere tegn enn å ha dem i en sekvens med komprimerte tegn.




Tester av algoritmen

Ford disse filene komprimerer algoritmen vi har skrevet en god del. Under er skjermdumper fra originalfil, komprimert og utpakkert fil fra komprimert. Alle de utpakkede filene har vi sjekket er lik med originalfilen ved bruk av UNIX sin terminalkommando «diff»

.txt og .lyx

 diverse-txt.txt 16 KB Modified: Today, 22:09 Add Tags... <div> <div>General:</div> <div> Kind: Plain Text Document Size: 16 427 bytes (20 KB on disk) Where: iCloud Drive • Desktop • NTNU • Algoritmer og Datastrukturer • o8 • Komprimering Created: 10 November 2021 at 22:09 Modified: 10 November 2021 at 22:09 <input type="checkbox"/> Stationery pad <input type="checkbox"/> Locked </div> </div>	 compressed-diverse-tx... 11 KB Modified: Today, 23:31 Add Tags... <div> <div>General:</div> <div> Kind: unix compressed archive Size: 11 197 bytes (12 KB on disk) Where: iCloud Drive • Desktop • NTNU • Algoritmer og Datastrukturer • o8 • Komprimering Created: 10 November 2021 at 23:31 Modified: 10 November 2021 at 23:31 <input type="checkbox"/> Stationery pad <input type="checkbox"/> Locked </div> </div>	 uncompressed-diverse... 16 KB Modified: Today, 23:31 Add Tags... <div> <div>General:</div> <div> Kind: Plain Text Document Size: 16 427 bytes (20 KB on disk) Where: iCloud Drive • Desktop • NTNU • Algoritmer og Datastrukturer • o8 • Komprimering Created: 10 November 2021 at 23:31 Modified: 10 November 2021 at 23:31 <input type="checkbox"/> Stationery pad <input type="checkbox"/> Locked </div> </div>
--	---	--

Her ser vi at den komprimerte filen er ca. 5KB mindre enn den originale filen.

 diverse-lyx.lyx 180 KB Modified: Today, 22:18 Add Tags... <div> <div>General:</div> <div> Kind: Document Size: 180 042 bytes (180 KB on disk) Where: iCloud Drive • Desktop • NTNU • Algoritmer og Datastrukturer • o8 • Komprimering Created: 10 November 2021 at 22:18 Modified: 10 November 2021 at 22:18 <input type="checkbox"/> Stationery pad <input type="checkbox"/> Locked </div> </div>	 compressed-diverse-l... 109 KB Modified: Today, 23:31 Add Tags... <div> <div>General:</div> <div> Kind: unix compressed archive Size: 108 770 bytes (111 KB on disk) Where: iCloud Drive • Desktop • NTNU • Algoritmer og Datastrukturer • o8 • Komprimering Created: 10 November 2021 at 23:31 Modified: 10 November 2021 at 23:31 <input type="checkbox"/> Stationery pad <input type="checkbox"/> Locked </div> </div>	 uncompressed-divers... 180 KB Modified: Today, 23:31 Add Tags... <div> <div>General:</div> <div> Kind: Document Size: 180 042 bytes (180 KB on disk) Where: iCloud Drive • Desktop • NTNU • Algoritmer og Datastrukturer • o8 • Komprimering Created: 10 November 2021 at 23:31 Modified: 10 November 2021 at 23:31 <input type="checkbox"/> Stationery pad <input type="checkbox"/> Locked </div> </div>
--	---	---

Her ser vi at den komprimerte er ca. 79KB mindre enn originalen.

.pdf

 opg8.pdf 87 KB Modified: Yesterday, 19:57 Add Tags... <div> <div>General:</div> <div> Kind: PDF document Size: 87 400 bytes (90 KB on disk) Where: iCloud Drive • Desktop • NTNU • Algoritmer og Datastrukturer • o8 • Komprimering Created: 9 November 2021 at 19:57 Modified: 9 November 2021 at 19:57 <input type="checkbox"/> Stationery pad <input type="checkbox"/> Locked </div> </div>	 compressed-opg8.Z 88 KB Modified: Today, 23:31 Add Tags... <div> <div>General:</div> <div> Kind: unix compressed archive Size: 87 802 bytes (90 KB on disk) Where: iCloud Drive • Desktop • NTNU • Algoritmer og Datastrukturer • o8 • Komprimering Created: 10 November 2021 at 23:31 Modified: 10 November 2021 at 23:31 <input type="checkbox"/> Stationery pad <input type="checkbox"/> Locked </div> </div>	 uncompressed-opg8.pdf 87 KB Modified: Today, 23:31 Add Tags... <div> <div>General:</div> <div> Kind: PDF document Size: 87 400 bytes (90 KB on disk) Where: iCloud Drive • Desktop • NTNU • Algoritmer og Datastrukturer • o8 • Komprimering Created: 10 November 2021 at 23:31 Modified: 10 November 2021 at 23:31 <input type="checkbox"/> Stationery pad <input type="checkbox"/> Locked </div> </div>
--	--	---

Her ser vi at den komprimerte filen faktisk er litt større enn originalfilen. Dette er fordi PDF-filer allerede er komprimert, så å komprimere disse kan faktisk medføre at den komprimerte blir litt større. Grunnen til det er at hvis man får mange tilfeller der man har en sekvens med ukomprimerte bytes som har en lengde på 1, så bruker man altså 2 bytes for å representere 1 byte. En byte for å indikere at det kommer en sekvens med ukomprimerte bytes, og en for selve byten. Hvis den jevnlig alternerer mellom sekvensreferanser og ukomprimerte sekvenser på 1 kan dette dermed føre at filen blir litt større enn originalen. Vi kan forebygge

dette ved å øke hukommelsesavstanden til algoritmen. I algoritmen vår er avstandshukommelsen kun på 1 byte. (Vi har testet med to også, men kun sammen med Huffman og ikke alene. Her ble de komprimerte filene faktisk marginalt større enn ved en byte).

 diverse-pdf.pdf 244 KB Modified: Today, 22:14 Add Tags... ▼ General: Kind: PDF document Size: 244 176 bytes (246 KB on disk) Where: iCloud Drive • Desktop • NTNU • Algoritmer og Datastrukturer • o8 • Komprimering Created: 10 November 2021 at 22:14 Modified: 10 November 2021 at 22:14	 compressed-diverse-... 234 KB Modified: Today, 23:32 Add Tags... ▼ General: Kind: unix compressed archive Size: 233 791 bytes (238 KB on disk) Where: iCloud Drive • Desktop • NTNU • Algoritmer og Datastrukturer • o8 • Komprimering Created: 10 November 2021 at 23:32 Modified: 10 November 2021 at 23:32	 uncompressed-divers... 244 KB Modified: Today, 23:32 Add Tags... ▼ General: Kind: PDF document Size: 244 176 bytes (246 KB on disk) Where: iCloud Drive • Desktop • NTNU • Algoritmer og Datastrukturer • o8 • Komprimering Created: 10 November 2021 at 23:32 Modified: 10 November 2021 at 23:32
--	--	---

Ved denne PDF-filen er den komprimerte filen mindre igjen. Grunnen til at den klarer å komprimere er at denne PDF-en kan være det første bildet som inneholder en repetisjon av mange blå piksler, men siden filen er komprimert kan det også rett og slett være en tilfeldighet at den klarer å komprimere denne filen og ikke den første.

Huffman

Huffman-koding komprimerer filer ved å se på hvilke bytes som blir brukt oftest i en fil. Den bruker så et binærtrep for å produsere binære koder for alle bytesene som finnes i denne filen. Huffman-kodene er enten kortere eller like lange som den originale byen. Dette fungerer derimot kun dersom antall unike bytes i filen er mindre 256. Hvis alle 256 tegn skal ha en unik Huffman-kode, må Huffman-treet bruke 8 ledd på å komme seg til hver eneste løvnode. Da vil alle bytes uavhengig av frekvens ha en kode som er 8 bits lang. Altså er det likegyldig om man bruker standard binærkode for bytes eller om man bruker Huffman-koding i slike tilfeller der huffmann-treet er fylt opp. Ulempen ved å velge Huffman i et slikt tilfelle, er at denne krever en frekvenstabell på ca. 1KB (vi bruker $1024 \text{ kb} + 32 \text{ bit}$, siden vi inkluderer en ekstra int i frekvenstabellen som symboliserer når algoritmen skal stoppe.) Dette vil føre til at man legger til ca. 1KB ekstra i filen.

Hvis man skal få maksimalt ut av Huffman-koding og Lempel-Ziv kombinert, bør man derfor bruke Huffman på bytesekvenser som ikke kunne bli komprimert siden disse har større sannsynlighet for å ha færre unike tegn enn hele koden tilsammen. I et slikt tilfelle må man

lage en kombinert frekvenstabellen for alle delsekvensene med ukomprimerte bytes før man starter komprimeringen. Hvis man lager en frekvenstabell for hver eneste delsekvens med ukomprimerte bytes, vil man legge til 1KB ekstra for hver delsekvens som kan føre til at man øker plassen som blir brukt isteden for å redusere den. Hvis disse delsekvensene er lange nok og de samme tegnene ofte går igjen, kan man i teorien tjene på det til tross for at man har en egen frekvenstabell for hver delsekvens.

Siden vi skrev først Lempel-Ziv og deretter Huffman, har vi implementert det slik at vi først komprimerer med Lempel-Ziv, også komprimerer vi bytetabellen denne igjen returnerer med Huffman-koding før vi skriver den til komprimert fil. Omvendt rekkefølge for dekomprimering. Dette gir positive uttelling sammenlignet med kun å komprimere med Lempel-Ziv, men kun for lyx-filen. De andre filene får en liten økning.




De minste filene tjener ikke nok på Lempel-Ziv og deretter Huffman (i motsetning til kun Lempel-Ziv komprimering) til at det kompenserer for de ekstra 1056 bytesene vi bruker for frekvenstabellen. For de større filene får vi også en økning i størrelse på litt under 1KB ekstra med data i den komprimerte filen (sammenlignet med kun Lempel-Ziv). Det er som sagt grunnet at den bruker nesten alle 256 ulike bytes. Den mellomste filen diverse.lyx blir ca. 5KB mindre enn hvis man kun komprimerer med lempel-ziv, så her tjener man altså på å bruke begge komprimeringsalgoritmene. Alle de komprimerte filene uavhengig om de kun er komprimert med Lempel-Ziv eller Lempel-Ziv også Huffmann, er mindre enn originalfilen. Unntaket er oppgave-PDFen. Dette som tidligere nevnt grunnet at PDF allerede er komprimert.

Vi har ikke mer kapasitet å bruke på denne innleveringen slik at vi kunne fått implementert en gunstig kombinasjon av de to komprimeringsalgoritmene istedenfor å komprimere hver for seg.

Vi har unngått å koke denne øvingen helt og har skrevet alt fra egen kunnskap + kunnskap tilegnet oss under arbeid. Dette har vært veldig krevende og dermed medført at vi har bruk svært mange timer på denne øvingen. Til gjengjeld har vi lært masse om primitive typer (spesielt bytes), bits, bit-operatorer, og ikke minst debugging.

Screenshots av Lempel-Ziv + Huffman

.txt

 diverse-txt.txt 16 KB Modified: Today, 15:05	 compressed-diverse-tx... 12 KB Modified: Today, 17:54	 decompressed-diverse... 16 KB Modified: Today, 17:54
Add Tags...	Add Tags...	Add Tags...
General: Kind: Plain Text Document Size: 16 438 bytes (20 KB on disk) Where: iCloud Drive ▸ Desktop ▸ NTNU ▸ Algoritmer og Datastrukturer ▸ o8 ▸ Komprimering Created: 10 November 2021 at 22:09 Modified: 12 November 2021 at 15:05	General: Kind: unix compressed archive Size: 11 742 bytes (12 KB on disk) Where: iCloud Drive ▸ Desktop ▸ NTNU ▸ Algoritmer og Datastrukturer ▸ o8 ▸ Komprimering Created: 12 November 2021 at 17:54 Modified: 12 November 2021 at 17:54	General: Kind: Plain Text Document Size: 16 438 bytes (20 KB on disk) Where: iCloud Drive ▸ Desktop ▸ NTNU ▸ Algoritmer og Datastrukturer ▸ o8 ▸ Komprimering Created: 12 November 2021 at 17:54 Modified: 12 November 2021 at 17:54

.lyx

 diverse-lyx.lyx 180 KB Modified: Today, 15:05	 compressed-diverse-l... 105 KB Modified: Today, 17:54	 decompressed-divers... 180 KB Modified: Today, 17:54
Add Tags...	Add Tags...	Add Tags...
General: Kind: Document Size: 180 052 bytes (180 KB on disk) Where: iCloud Drive ▸ Desktop ▸ NTNU ▸ Algoritmer og Datastrukturer ▸ o8 ▸ Komprimering Created: 10 November 2021 at 22:18 Modified: 12 November 2021 at 15:05	General: Kind: unix compressed archive Size: 104 813 bytes (106 KB on disk) Where: iCloud Drive ▸ Desktop ▸ NTNU ▸ Algoritmer og Datastrukturer ▸ o8 ▸ Komprimering Created: 12 November 2021 at 17:54 Modified: 12 November 2021 at 17:54	General: Kind: Document Size: 180 052 bytes (180 KB on disk) Where: iCloud Drive ▸ Desktop ▸ NTNU ▸ Algoritmer og Datastrukturer ▸ o8 ▸ Komprimering Created: 12 November 2021 at 17:54 Modified: 12 November 2021 at 17:54

.pdf

 oppg8.pdf 87 KB Modified: 9 November 2021 at 19:57	 compressed-oppg8.Z 89 KB Modified: Today, 18:53	 decompressed-oppg8.pdf 87 KB Modified: Today, 18:53
Add Tags...	Add Tags...	Add Tags...
General: Kind: PDF document Size: 87 400 bytes (90 KB on disk) Where: iCloud Drive ▸ Desktop ▸ NTNU ▸ Algoritmer og Datastrukturer ▸ o8 ▸ Komprimering Created: 9 November 2021 at 19:57 Modified: 9 November 2021 at 19:57	General: Kind: unix compressed archive Size: 88 832 bytes (131 KB on disk) Where: iCloud Drive ▸ Desktop ▸ NTNU ▸ Algoritmer og Datastrukturer ▸ o8 ▸ Komprimering Created: 12 November 2021 at 17:55 Modified: 12 November 2021 at 18:53	General: Kind: PDF document Size: 87 400 bytes (131 KB on disk) Where: iCloud Drive ▸ Desktop ▸ NTNU ▸ Algoritmer og Datastrukturer ▸ o8 ▸ Komprimering Created: 12 November 2021 at 17:55 Modified: 12 November 2021 at 18:53

 diverse-pdf.pdf 244 KB Modified: 10 November 2021 at 22:14	 compressed-diverse-... 235 KB Modified: Today, 18:53	 decompressed-divers... 244 KB Modified: Today, 18:53
Add Tags...	Add Tags...	Add Tags...
General: Kind: PDF document Size: 244 176 bytes (246 KB on disk) Where: iCloud Drive ▸ Desktop ▸ NTNU ▸ Algoritmer og Datastrukturer ▸ o8 ▸ Komprimering Created: 10 November 2021 at 22:14 Modified: 10 November 2021 at 22:14	General: Kind: unix compressed archive Size: 234 821 bytes (262 KB on disk) Where: iCloud Drive ▸ Desktop ▸ NTNU ▸ Algoritmer og Datastrukturer ▸ o8 ▸ Komprimering Created: 12 November 2021 at 17:55 Modified: 12 November 2021 at 18:53	General: Kind: PDF document Size: 244 176 bytes (262 KB on disk) Where: iCloud Drive ▸ Desktop ▸ NTNU ▸ Algoritmer og Datastrukturer ▸ o8 ▸ Komprimering Created: 12 November 2021 at 17:55 Modified: 12 November 2021 at 18:53