

Øving 7 – Komprimering

Algoritmer og datastrukturer

Jeg har samarbeidet med Oline Amundsen. Vi har gjort alt sammen, samarbeidet har vært faglig diskusjon, ikke fordeling av arbeidsmengde.

Deloppgave Lempel-Ziv

Vi implementerte en versjon av Lempel-Ziv. Vi brukte en «short» til å se tilbake i fila med. Vi prøvde først med en «byte» bare, som var lettere å implementere, men dette resulterte i at nesten alle filene ble større enn originalen. Med en «short» fikk vi noen resultater som var greie og andre som var veldig bra.

Deretter laget vi dekomprimeringen i et annet prosjekt og sjekket at vi klarte å dekomprimere filene så de ble helt like originalene.

Vi valgte å ikke optimalisere metoden som leter etter en match til en sekvens med tegn, da programmet kjørte på en grei nok tid og (Forelesningen.pdf bruke litt lengere tid, men under et minutt) denne optimaliseringen ikke var et krav og med litt lite tid, så valgte vi heller droppe dette.

Deloppgave Huffmankoding

Huffmann komprimeringen implementerte vi så den gikk gjennom den allerede komprimerte (med LZ) fila og laget en frekvens tabell basert på de ukomprimerte delene av fila. Så bygde vi opp et Huffmann-tre av en simpel Node-klasse, med et venstre- og høyrebarn, og en forelder. Vi valgte å lagre alle Huffmannkodene i int's og ikke i long, da de 32 bits'ene i int en viste seg å være mer enn nok, da vi ikke fikk noen trær som trengte mer.

Vi skrev først inn frekvenstabellen, som en int[256] som da legger til 1kB i den komprimerte fila. Deretter gikk vi gjennom den LZ komprimerte fila på nytt å skrev de essensielle delene rett til fila når de dukket opp (da bakover referansene og angivelse av antall byte med ukomprimert data). Når ukomprimert data dukket opp skrev vi en ny byte[] laget av de bit-representasjonene vi fikk fra Huffmann-treet og skrev dette til fila.

Så laget vi dekomprimeringen i samme prosjekt som den andre dekomprimeringen, og gikk gjennom den komprimerte fila. Vi henter først ut frekvenstabbeløen og rekonstruerer Huffmann-treet basert på dette. Så leser vi av der det er Huffmann komprimert, da X antall bytes etter en positiv «short» fra LZ vi leser av en byte og navigerer oss gjennom Huffmann-treet til vi enten finner en blad-node eller går tom for bits i byten. Finner vi en blad-node skriver vi dens byte verdi til fila, om ikke så leser vi av en byte til og fortsetter der vi slapp. Når vi har skrevet X antall bytes til fila stopper vi å lete i treet og begynner prosessen på nytt ved å lese av en ny essensiell del av LZ komprimeringen.

Sluttprogrammet

Sluttprogrammet vårt tar en valgt fil i og komprimerer den først med LZ så med Huffman. Så kan vi gå i dekomprimeringen å velge en komprimert fil og dekomprimerer den, da først med Huffman og så med LZ.

Alle filene ligger i en mappe «Filer» i samme mappe som prosjektmappene ligger, dette ble gjort for lettere felles tilgang til filene mellom programmene.

```
diff Oppgavetekst.pdf DeKomp.Oppgavetekst.pdf
diff Forelesningen.pdf DeKomp.Forelesningen.pdf
diff Forelesningen.txt DeKomp.Forelesningen.txt
diff Forelesningen.lyx DeKomp.Forelesningen.lyx
```

Alle filene kunne komprimeres og dekomprimeres og være helt uendret.

	Original størrelse	Lempel-Ziv	Huffman (Begge)	Komprimering Lempel-Ziv	Komprimering Huffman (Begge)
Oppgavetekst.pdf	84 900 bytes	83 960 bytes	84 984 bytes	1.12%	-0.10%
Forelesningen.pdf	837 404 bytes	729 204 bytes	730 228 bytes	12.92%	12.80%
Forelesningen.txt	15 882 bytes	9 496 bytes	10 024 bytes	40.21%	36.89%
Forelesningen.lyx	178 871 bytes	15 149 bytes	15 066 bytes	91.53%	91.58%





Det var kun på LYX fila at Huffman ga noe forbedring på den totale komprimeringen. Men det fyller vel kravet. Alle filene blir totalt sett mindre enn original filene, med unntak av Oppgavetekst.pdf som med Huffman ble 0.10% større enn originalen. Vi hadde problemer med Huffman og PDF formatet, da Huffman-treet vårt alltid ble (eller nesten ble) et komplett binærtre, altså alle bitrepresentasjonene var 8 bit lange, så det ble ikke noe forkorting, og da når vi legger til 1kB (frekvenstabellen) så ble resultatet dårligere enn før Huffman. Etter en god del debugging konkluderte vi med at frekvensen mellom de forskjellige byte-verdiene i PDF'en ble alt for like, noe som var ganske merkelig, men så ut til å stemme, dette gjorde at alle hadde nesten lik frekvens og havnet på åttende nivå i binær treet (niende lag).

Se skjermbilder av dataen over på neste side.





Skjermbilder av dataen over:

(rekke følge på bilde: Original, Lempel-Ziv-komprimert, Huffman-Komprimert, dekomprimert)





Oppgaveteskt.pdf:

 Oppgaveteskt.pdf 85 KB Modified: 14 October 2020 at 12:12 Add Tags... ▼ General: Kind: PDF document Size: 84 900 bytes (86 KB on disk)	 Oppgaveteskt.pdf.lz 84 KB Modified: Today, 13:07 Add Tags... ▼ General: Kind: Document Size: 83 960 bytes (86 KB on disk)	 Oppgaveteskt.pdf.lz.hm 85 KB Modified: Today, 13:07 Add Tags... ▼ General: Kind: Document Size: 84 984 bytes (139 KB on disk)	 DeKomp.Oppgavetesks... 85 KB Modified: Today, 13:10 Add Tags... ▼ General: Kind: PDF document Size: 84 900 bytes (139 KB on disk)
--	---	---	---





Forelesningen.pdf:

 Forelesningen.pdf 837 KB Modified: 14 October 2020 at 12:13 Add Tags... ▼ General: Kind: PDF document Size: 837 404 bytes (840 KB on disk)	 Forelesningen.pdf.lz 729 KB Modified: Today, 13:09 Add Tags... ▼ General: Kind: Document Size: 729 204 bytes (733 KB on disk)	 Forelesningen.pdf.lz.hm 730 KB Modified: Today, 13:09 Add Tags... ▼ General: Kind: Document Size: 730 228 bytes (791 KB on disk)	 DeKomp.Forelesninge... 837 KB Modified: Today, 13:10 Add Tags... ▼ General: Kind: PDF document Size: 837 404 bytes (856 KB on disk)
--	---	--	---

Forelesningen.txt:

 Forelesningen.txt 16 KB Modified: 14 October 2020 at 12:14 Add Tags... ▼ General: Kind: Plain Text Document Size: 15 882 bytes (16 KB on disk)	 Forelesningen.txt.lz 9 KB Modified: Today, 13:09 Add Tags... ▼ General: Kind: Document Size: 9 496 bytes (12 KB on disk)	 Forelesningen.txt.lz.hm 10 KB Modified: Today, 13:09 Add Tags... ▼ General: Kind: Document Size: 10 024 bytes (12 KB on disk)	 DeKomp.Forelesningen.... 16 KB Modified: Today, 13:10 Add Tags... ▼ General: Kind: Plain Text Document Size: 15 882 bytes (16 KB on disk)
--	--	---	---

Forelesningen.lyx:

 Forelesningen.lyx 179 KB Modified: Today, 09:30 Add Tags... ▼ General: Kind: TextEdit Document Size: 178 871 bytes (180 KB on disk)	 Forelesningen.lyx.lz 15 KB Modified: Today, 13:09 Add Tags... ▼ General: Kind: Document Size: 15 149 bytes (16 KB on disk)	 Forelesningen.lyx.lz.hm 15 KB Modified: Today, 13:09 Add Tags... ▼ General: Kind: Document Size: 15 066 bytes (16 KB on disk)	 DeKomp.Forelesninge... 179 KB Modified: Today, 13:10 Add Tags... ▼ General: Kind: TextEdit Document Size: 178 871 bytes (201 KB on disk)
---	--	---	--