Tekstin Tiivistys - Huffman algoritmi
Määrittelydokumentti

v1.01

Tulen toteuttamaan Huffmannin algoritmin tekstitiedoston kompressointiin. Kompressoitu tiedosto tulee myös olla muotoa jonka pystyy ilman ongelmia saattamaan taas luettavaan muotoon. Tavoite on että syötteenä annettun tekstitiedoston saisi n.40-60% pienemmäksi kooltaan alkuperäiseen verrattuna.

Tiedossa olevat käyttöön tulevat algoritmit:

Huffmannin algoritmi:

Selkeä ja toimiva algoritmi teksitiedoston kompressointiin.

Puun esijärjestyksessä läpikäynti:

Meidän tarkoitukseen sopiva binäärisen hakupuun läpikäynti algoritmi.

Integer.toBinaryString(int i)

Helpoin tapa saattaa integer binääriksi.

Tiedossa olevat käyttöön tulevat tietorakenteet:

Binäärinen Hakupuu:

Helpoin tapa jakaa uudet binääriarvot kullekkin merkille riippuen niiden toistojen määrästä syötteessä.

PriorityQueue:

Tehokas tietorakenne puun rakentamiseen, sillä pienimpiä solmuja poistetaan ja uusia lisätään solmuja tietorakenteeseen useasti puun rakennuksen aikana.

HashMap:

Map jota käytetään pitämään muistissa kuinka monta kutakin merkkiä on syötteessä tähän mennessä tullut. Valitsin tietorakenteen sillä siitä on nopeaa tarkistaa onko tiettyä merkkiä jo lisätty tietorakenteeseen ja siinä on helppoa pitää muistissa sekä merkkiä että siihen liittyvää numeroa.

ArrayList:

Lista jokaisesta merkistä, ja sen uudesta bitti setistä. Valitsin tämän sillä siihen on helppo lisätä ja koska löydettyjen merkkejen määrä ei pakosti ole tiedossa niin lista kasvattaa kokoaan dynaamisesti.

Integer Array

Määritellyn (8) kokoinen lista tavujen kirjoittamisen helpottamiseksi. Koska lista on tietyn kokoinen, ja sen sisältö vaihtuu usein, on Arrayn käyttö tässä tapauksessa tehokkain ja aika ja tilavaatimuksiltaan.

Tilavaatimus:

Itse algoritmi tarvitsee tilaa vain tietylle määrälle ascii merkkejä ja niihin liittyviin numeroihin maksimissaan O(1), muttta tietenkin jos kompressoitu tiedosto lasketaan mukkaan niin tilavaatimus nousee O(0.4n-0.6n) eli O(n).

Aikavaatimus:

Kompressoinnissa, koska erilaisia merkkejä on vain teitty määrä, hakupuun ja priorityqueuen läpikäyntiin kuluu maksimissaan vain lineaarinen määrä aikaa O(1), eli meidän pitää välittää ainoastaan tekstitiedoston lukemisesta ja merkin tiedon lisäämisestä Arrayhin. Vaikka tämä joudutaan tekemään kahdesti, niin aikavaatimus ei ole O(2n) suurempi, eli riippuvainen syötteen (tiedoston) koosta. Koska kertoimet eivät tähän aikavaatimuksen merkintä tapaan vaikuta, on kompressoinnilla arvo O(n). Takaisin luettavaan muotoon saatettaessa taas käydään jokainen merkki kompressoidussa tiedostossa läpi, jokaista merkkiä kohden pitää käydä binäärihakupuuta läpi ja kompressoidussa muodossa oleva bittisetti vaihdetaan takaisin alkuperäiseen. Hakupuulla on maksimissaan 255 lehteä, eli emme ole puun läpikäynnistä tässä tilanteessa kiinostuneita. Eli aikavaatimus on yhä täysin riippuvainen syöteen koosta O(n).

Lähteet:

http://en.wikipedia.org/wiki/Huffman_coding
https://www.youtube.com/watch?v=ZdooBTdW5bM