

# Tietorakenteiden harjoitustyö: Huffman-koodaus

Tapio Ikkala

May 24, 2016

Tietorakenteiden harjoitustyön aiheenani on Huffman-pakkausalgoritmi tekstitiedostoille. Teen ohjelmani Java-kielellä.

Huffmanin algoritmissa pakattavan tiedoston merkit järjestetään yleisyysjärjestykseen, antaen yleisimmälle merkille lyhyin koodi. Tästä seuraa, että harvinaiset merkit saavat pidemmät koodit, mutta kunhan yleisten ja harvinaisten merkkien määrässä on riittävän suuri ero, pienenee tiedoston koko huomattavasti. Huffman-koodi antaa jokaiselle merkille niin sanotun etuliitekoodin (prefix code tai prefix-free code). Etuliitekoodissa minkään merkin koodi ei voi olla toisen merkin koodin alkuosa, joten koodi on aina yksiselitteinen, vaikka merkkien koodien pituudet vaihtelevatkin.

Toteutan algoritmia varten minimikeon, jonka avulla tiedostosta löytyneet merkit on helppo järjestää Huffman-puun rakentamista varten.

Ohjelma toimii siten, että sille syötetään tiedoston polku, uuden tiedoston nimi sekä valitaan halutaanko purkaa vai pakata.

Huffmanin algoritmin aikavaativuus on  $\theta(k \log k)$ , missä  $k$  on tiedoston eri merkkien lukumäärä (tekstitiedostoissa  $k \leq 256$ ). Pakkaamiseen ja purkamiseen kuluvaa aikaa hallitsee kuitenkin tiedoston lukeminen ja kirjoittaminen, joiden aikavaativuus on  $\theta(n)$ , missä  $n$  on tiedoston sisältämien merkkien lukumäärä.

Lähteenä olen käyttänyt Thomas H. Cormenin kirjaa “Introduction to Algorithms” sekä Wikipedian Huffman-koodista kertovaa [sivua](#).