Tiralabra 2015/periodi 3: Viikkoraportti 1

Kertauksena: Toteutan harjoitustyönä pakkausalgoritmia. Ensimmäisessä vaiheessa toteutan tekstitiedostojen Huffman-koodauksen, minkä jälkeen projektia on helppo laajentaa muilla pakkaustyökaluilla, joita voidaan soveltaa yhdessä Huffmanin kanssa.

Viime viikolla sain tosiaan jo merkkikohtaisten Huffman-koodisanojen muodostamisen kuntoon. Tämän viikon tärkein tehtävä oli viime viikolla laimilyödyn testaamisen ja dokumentoinnin laittaminen kuntoon – nyt koodissa on javadocit paikallaan ja jUnit-testit ainakin varsin kattavasti.

Tällä viikolla varsinainen pakkausohjelmakin eteni hieman. Tutustuin bittioperaatioiden pyöritttelemiseen Javalla, ja sainkin aikaiseksi jo ainakin jonkinlaisen toteutuksen bittien kirjoittamiseksi tiedostoon. BitWriter-luokan writeBits-metodi ottaa parametrina kirjoitettavien bittien merkkijonoesityksen, siis esimerkiksi ”1010010101”, ja kirjoittaa sen yksittäisiä bittejä käyttäen tavuittain tiedostoon. Tällä hetkellä ohjelma kutsuu metodia jokaiselle pakattavan tekstin merkille. Yksittäisen bitin koodaus aiheuttaa vakiomäärän työtä, joten operaation asymptoottinen aikavaativuus on luokkaa O(l*n), missä l on pisimmän yhden merkin koodisanan pituus, ja n merkkien määrä lähdetekstissä.

Testatakseni koodauksen toimimista ja etenkin nopeutta käytännön tilanteessa latasin Project Gutenbergista kirjan Liisa Ihmemaassa (alice.txt), jota mainin suorittaminen tällä hetkellä pakkaa. Pakkaus vähentää tiedoston koon tällä hetkellä noin 168 kilotavusta noin 92 kilotavuun, ja kestää noin sekunnin. Koon muutos on toki merkittävä, mutta edelleen kaukana 'oikeassa käytössä' olevien pakkausmetodien tehosta (esim bzip2 pakkaa saman noin 49 kilotavuun). Lisäksi on otettava huomioon, että tämän hetkinen pakkaus ei sisällä Huffman-koodisanoja purkua varten. Olettaen, että koodisanat ovat pituudeltaan keskimäärin tavun, ja että eri koodisanoja erottamaan luodaan yksi ylimääräinen, noin tavun pituinen koodisana, tuo koodaus pitäisi saada mahtumaan noin 500 tavuun.

Toinen huomio tämänhetkisen algoritmin puutteista on rajoitus pakattavien tiedostojen tyyppiin. Toteutin näin aluksi ohjelman niin, että pakkaus onnistuu vain 8-bit ASCII-merkeistä koostuville tekstitiedostoille. Jos tekstitiedostossa on muita merkkejä, ne jäävät pakkauksessa kokonaan pois (Alicessa yksi tällainen merkki). Nyt kun bittien pyörittely on jo hieman tutumpaa, uskoisin onnistuvani muokkaamaan tuon melko helposti niin, että 'charien' sijaan koodisanat annetaankin 'byteille', jolloin kaikkien tiedostojen luku ja koodaus tavuittain pitäisi onnistua.

Jo mainittujen jatkoideoiden lisäksi luvassa on tietenkin myös pakatun tiedoston purku. Tämän jälkeen, kun Huffman-koodaus toimii kaikkineen ja sen vaatimat tietorakenteet toteutettu omina toteutuksina, suunnitelin seuraavaksi vaiheeksi esimerkiksi 'run-length'-koodausta ja Burrows-Wheeler -muunnosta (nämä ja Huffman ovat mm. jo aiemmin mainitun bzip2:n käyttämiä operaatioita). Nämä tosin jäänevät vielä seuraavaa viikkoa pidemmälle.