Viikkoraportti 1

Tällä viikolla valitsin aiheeksi LZW-algoritmin. Tutkin algoritmin toteutustapaa pääosin helpoimman kautta eli YouYube-videoesitysten ja englanninkielisen Wikipedia-artikkelin avulla. CS Learning 101:n (http://www.cslearning101.com/) opetusvideo helpotti asian tajuamista eniten.

Viikkotyö koostui pääosin mietiskelystä ja raporttien kirjoittamisesta. Testasin myös tekstitiedostojen lukemista binaarimuotoon javalla ja tutkin, miten erilaiset tekstin koodausmuodot (ANSI, UTF-8 jne.) näkyvät tiedostossa. Eri tavoin koodatut tekstit näkyvät joko 8- tai 16-bittisinä eli ne mahtuvat yhteen tai kahteen tavuun. Pohdin sitä, että ei varmaankaan kannata ajatella pakkausta tekstimerkkeinä vaan tavumuodossa, joten yhden merkin voi pakata kahtena eri merkkinä, minkä ei pitäisi viedä enemmän tilaa kuin yhden merkin käsittely kaksitavuisena.Näin ollen jos merkistössä on jotain erikoisempia merkkejä, jotka eivät mahdu ensimmäiseen tavuun ja 256 merkkiin, ei kuitenkaan joudu alustamaan LZW-sanakirjaa, jossa olisi 256 potenssiin kaksi merkkiä. Riittää siis, kun alustaa 256 merkin sanakirjan ja käsittelee vain tavuja eikä kokonaisia merkkejä. Kun pakkaa vain näitä tavuja ja purkaa pakkauksen, java osaa kuitenkin tallentaa tavut tekstiksi uudelleen, niin että ne vastaavat alkuperäistä tekstiä. Testasin tätä satunnaisesti valitulla, UTF-8 merkistössä sijalla 50010 olleella merkillä, niin että tekstitiedosto oli tavuiksi lukemisen ja uudelleen tekstiksi kirjoittamisen jälkeen merkki merkiltä sama.