Merkkijonomenetelmät (syksy 2008)

Vastaa kurssikyselyyn (`http://ilmo.cs.helsinki.fi/kurssit/servlet/Valinta`).

Tehtäviä viimeiselle viikolle

1. Muodosta jonon `mississippi` loppuosataulukko käyttäen luennoilla esitettyä osarekursiivista algoritmia.
2. Laske jonon `mississippi` erilaisten osajonojen lukumäärä.
3. Ratkaise harjoitusten 5 tehtävä 6 käyttäen loppuosa*taulukkoa*.

Tyypillisiä koetehtäviä

1. Mitä yhteyksiä/eroavaisuuksia on seuraavilla käsitepareilla?
   (a) Knuth-Morris-Pratt-algoritmi ja Aho-Corasick-algoritmi.
   (b) Shift-Or-algoritmi ja Myersin bittirinnakkainen algoritmi.
   (c) Karp-Rabin-algoritmi ja Karp-Miller-Rosenberg nimentätekniikka.
   (d) Editointietäisyyden laskeminen ja likimääräinen hahmonsovitus.
   (e) Merkkijonopikajärjestäminen ja ternääripuu.
   (f) Lopusta-alkuun ja alusta-loppuun kantalukujärjestäminen.

   Muutama lause riittää vastaukseksi kuhunkin kohtaan
2. Anna Aho-Corasick-automaatti hahmojoukolle $\{$`ARI`, `ARMAS`, `KARITA`, `MARKO`, `RITVA`$\}$. Simuloi automaatin käyttöä tekstillä $T =$ `SANKARITAR`.
3. Olkoon S n:n pituinen merkkijono ja R m:n merkin monijoukko vakioaakkostossa Σ. R:n esiintymä S:ssä on m:n pituinen osajono, joka koostuu täsmälleen R:n merkeistä. Esimerkiksi joukon $R = \{$`A`, `A`, `B`, `C`$\}$ ainoa esiintymä jonossa $S =$ `ABAHGCABAH` on `CABA`. Anna ajassa $\mathcal{O}(n+m)$ toimiva algoritmi, joka löytää kaikki R:n esiintymät S:ssä.
4. Merkkijonoja $A = a_1 a_2 \cdots a_m$ ja $B = b_1 b_2 \cdots b_n$ kutsutaan *numeerisiksi merkkijonoiksi*, jos aakkostona on reaaliluvut, eli $a_i, b_j \in \mathbb{R}$ kaikilla $1 \leq i \leq m$, $1 \leq j \leq n$. Tarkastellaan erästä numeeristen merkkijonojen editointietäisyyttä, $D_K(A,B)$, joka määritellään seuraavasti. Sallitaan kolme editointioperaatiota: korvaus, poisto ja lisäys. Korvausoperaation $a_i \to b_j$ hinta on $|a_i - b_j|$. Poistot $a_i \to \epsilon$ ja lisäykset $\epsilon \to b_j$ ovat *ilmaisia*, mutta niiden yhteismäärä saa olla korkeintaan K. Editointietäisyys $D_K(A,B)$ on siis pienin kokonaishinta editointijonolle, joka muuttaa A:n B:ksi ja sisältää enintään K lisäystä ja poistoa. Tehtävänä on kehittää etäisyyden $D_K(A,B)$ laskemiseksi algoritmi, joka toimii ajassa $\mathcal{O}(mnK)$.
5. Kuvaa periaatetasolla menetelmiä, joilla binäärihakua voidaan tehostaa, kun alkiot ovat merkkijonoja.
6. Olkoon $\mathcal{R} = \{S_1, S_2, \ldots S_n\}$ joukko merkkijonoja vakioaakkostossa. Jonon S_i *lyhin yksilöivä osajono* on lyhin jono, joka esiintyy S_i:ssä mutta ei muissa $\mathcal{R}$:n jonoissa. Hahmottele lineaarisessa ajassa toimiva algoritmi, joka löytää kaikkien $\mathcal{R}$:n jonojen lyhimmän yksilöivän osajonon.