Merkkijonomenetelmät (kevät 2005)

Harjoitus 4

Jarkko Toivonen pe 16.2. klo 10.15–12 CK107

1. Merkkijonon *q-gramma* on sen q:n pituinen osajono. Esim. jonon `ararat` 3-grammat ovat `ara`, `rar`, `ara` ja `rat`.

   (a) Osoita, että jos A ja B ovat merkkijonoja, joiden editointietäisyys $D(A, B)$ on enintään k, niin niillä on ainakin $|A| - q + 1 - kq$ yhteistä q-grammaa.

   (b) Hahmottele (a)-kohdan tulokseen perustuva seulontamenetelmä likimääräiseen hahmonsovitukseen.

2. Olkoon R monijoukko, jonka koko on n ja jossa on $k < n$ eri arvoa. Osoita, että ternääripikajärjestäminen järjestää joukon R ajassa $\mathcal{O}(n \log k)$ olettaen, että jakoalkioksi valitaan aina mediaani. (Vihje: Käytä tasoitettua analyysiä, missä kunkin kolmeenjaon maksavat joukkojen $R_=$ ja R' alkiot, missä R' on joukoista $R_<$ ja $R_>$ se, joka sisältää vähemmän eri arvoja.)

3. Esitä luentojen Algoritmista 3.3.2 (lopusta alkuun kantalukujärjestäminen) yleistetty versio, joka toimii myös eripituisilla merkkijonoilla lauseen 3.3.4 mukaisessa ajassa $\mathcal{O}(||\mathcal{R}|| + m\sigma)$, missä $||\mathcal{R}||$ on jonojen kokonaispituus ja m on pisimmän jonon pituus.

4. Luentojen algoritmissa 3.3.5 (alusta loppuun kantalukujärjestäminen) käytetään merkkijonopikajärjestämistä pienille joukoille. Mikä olisi algoritmin aikavaatimus, jos näin ei tehtäisi, eli jos rivi 1 korvattaisiin rivillä

   (1) **if** $|\mathcal{R}| \leq 1$ **then return** $\mathcal{R}$;

   Perustele.

5. Luennolla luvun 3.4 lopussa kuvataan, kuinka suuren kokonaislukuaakkoston tapauksessa kukin merkki voidaan hajottaa useammaksi merkiksi pienmmässä aakkostossa. Pienellä kokonaislukuaakkostolla voidaan tehdä päinvastoin: yhdistetään useampi merkki yhdeksi *superaakkoston* merkiksi. Esimerkiksi yhdistämällä k merkkiä siirrytään aakkostosta $[0, \sigma)$ aakkostoon $[0, \sigma^k)$. Jos kukin superaakkoston merkki on talletettu yhteen tietokoneen sanaan (eli k alkuperäistä merkkiä on pakattu yhteen sanaan), normaalit alarajat algoritmien vaativuudelle eivät enää päde.

   (a) Analysoi kaksivaiheisen kantalukujärjestämisen aikavaativuutta tällaisessa tapauksessa.

   (b) Hyvä valinta k:ksi on $\lfloor \log_\sigma n \rfloor$. Miksi? Mikä tällöin on kaksivaiheisen kantalukujärjestämisen aikavaativuus?