Merkkijonomenetelmät (syksy 2003)

Harjoitus 1

1. Analysoi triviaalialgoritmin 1.0.1 keskimääräistä merkkivertailujen lukumäärää. Oletetaan, että aakkoston Σ kukin merkki esiintyy samalla todennäköisyydellä $1/\sigma$ jokaisessa P:n ja S:n positiossa.

2. (a) Laske KMP-algoritmin $fail$-taulukko hahmolle `IKONITIKITIKO`.
   (b) Esitä tulos myös jononsovitusautomaattina.
   (c) Simuloi hahmontunnistuksen etenemistä itse valitsemassasi tekstissä.

3. (a) Modifioi KMP:n esiprosessointia niin, että uusi $fail$-taulukko $fail'$ toteuttaa seuraavan:
      – Korjaussiirtymän jälkeen hahmosta luettava merkki on erilainen kuin se merkki, joka olisi voitu lukea ennen korjaussiirtymää (eli jos $fail'[j] = k > 0$, niin $p_k \neq p_j$).
   (b) Sovella algoritmiasi tehtävän 2 hahmoon ja totea erot.

4. Luennolla esitetty KMP-algoritmi pysähtyy hahmon ensimmäiseen esiintymään. Muuta algoritmia niin, että se tulostaa *kaikki* esiintymät siirtymättä taaksepäin tekstissä.

5. KMP-algoritmi on *aakkostoriippumaton*, eli sen aikavaatimus ei riipu käytössä olevasta aakkostosta. Monissa muissa merkkijonoalgoritmeissa aakkoston tyyppi vaikuttaa aikavaativuuteen. Usein oletetaan, että käytössä on *vakioaakkosto* $\Sigma = \{1, \ldots, \sigma\}$, jolloin voidaan vakioajassa liittää/etsiä merkkiin $a \in \Sigma$ liittyvää lisäinformaatiota kokoa σ olevasta taulukosta. Pienempi oletus on, että aakkosto on *järjestetty*, jolloin kahdesta aakkoston merkistä a ja b voidaan sanoa onko $a < b$, $a > b$, vai $a = b$ (esim. $\Sigma \subset \mathbb{R}$). Tällöin taulukon käyttö ei onnistu suoraan. Tämän kiertämiseksi:

   - Olkoon $P, S \in \Sigma^*$, missä Σ on järjestetty aakkosto.
   - Anna algoritmi, joka tekee aakkostomuunnoksen $f : \Sigma \to \Sigma'$ siten, että jos ja vain jos $p_i = s_j$ niin $f(p_i) = f(s_j)$, missä $\Sigma' = \{1, \ldots, \sigma'\}$ ja $\sigma' \leq m + 1$. Algoritmin tulee toimia ajassa $O(n \log m)$, missä $|P| = m \leq n = |S|$.