

Málvinnsla - Skiladæmi 1

Háskóinn í Reykjavík - Tölvunarfræðideild

Ágúst 2007

1 Dæmi I - 20%

Í þessu dæmi eigið þið að nota *egrep* til að sækja upplýsingar úr hluta úr málheildinni “Íslensk orðtíðnibók” (sjá skrá 01PM.txt og lýsingu á íslenska markamenginu sem er aðgengileg undir “Annað efni” á vefsíðu námskeiðsins). Í öllum tilvikum eigið þið að sýna skipunina sem þið notið. Athugið: Kafli 2.3.4 í kennslubók gæti hjálpað við lausnina.

Notið *egrep* til að:

1. Skila öllum þeim línnum (þ.e. orðum og mörkum), úr málheildinni sem innhalda sögn, í framsöguhátti, germynd og fyrstu persónu.
2. Skila öllum þeim línnum sem innihalda forsetninguna “til”.
3. Skila öllum karlkyns nafnorðum í eintölu og nefnifalli sem hafa beygingarendinguna “ur” (t.d. “maður”). Sérnöfn eiga ekki að koma með.
4. Skila öllum nafnorðum og lýsingarorðum í eignarfalli (sérnöfn mega vera með).

2 Dæmi II - 30%

Skrifið reglulegar segðir sem bera kennsl á strengi ...

1. af lengd 4 sem byrja á a eða b , innihalda hvaða tvo stafi sem er í miðjunni og enda með a eða b í lokin.
2. af lengd 2 sem byrja á a og innihalda síðan hvaða staf sem er nema x , y eða z .
3. af lengd 3 sem innihalda eingöngu tölustafi.

4. úr inntaksstafrófinu $\Sigma = \{a, b\}$ sem inniheldur hlutstrenginn *abba*
5. úr inntaksstafrófinu $\Sigma = \{a, b\}$ þar sem engir tveir samliggjandi stafir eru eins (t.d., *abab* er í málinu en *abba* ekki). ϵ er í málinu.
6. úr inntaksstafrófinu $\Sigma = \{a, b\}$ sem byrja og enda á sama staf. ϵ er ekki í málinu.

3 Dæmi III - 15%

1. Teiknið endanlega stöðuvél, **DFA** vél, sem samsvarar reglulegu segðinni $(ab)^*c$
2. Teiknið endanlega stöðuvél, **DFA** vél, sem ber kennsl á alla strengi úr inntaksstafrófinu $\Sigma = \{0, 1\}$ (einnig ϵ), nema þá sem innihalda hlutstrenginn *001*.
3. Teiknið endanlega stöðuvél, **NFA** vél, með sem fæstum stöðum, sem samsvarar reglulegu segðinni $a|b^*|(a|b)^*a$

4 Dæmi IV - 35%

1. (15%) Skrifðu Perl-forrit, *findLongest.pl*, sem tekur inn eitt skráarnafn, finnur lengsta orðið í skránni og prentar það út ásamt lengd þess. Athugið að forritið skal virka fyrir textaskrá á hvaða sniði sem er. Prófið forritið á málheildinni 01PM.txt, skilið forritskóðanum og úttakinu miðað við þessa málheild. Prófið forritið ykkar einnig á annarri textaskrá sem inniheldur fleiri en tvo tóka í hverri línu, t.d. á einhverri frétt á www.mbl.is.
2. (20%) Skrifðu Perl-forrit, *tagFrequency.pl*, sem tekur inn markaða málheild (eitt orð/mark í hverri línu) og tíðnitölu t , og prentar út þau t mörk sem koma oftast fyrir í málheildinni. Skilið forritskóðanum og úttakinu miðað við málheildina 01PM.txt og tíðnitöluna 10.

5 Skilafrestur

Verkefninu skal skila í síðasta lagi þriðjudaginn 4. september, kl. 23:59.