Málvinnsla: Forritunarverkefni I - Tilreiðing texta

Háskóinn í Reykjavík - Tölvunarfræðideild

September 2007

1 Lýsing

Í þessu verkefni eigið þið að búa til tilreiðara fyrir (íslenskan) texta. Með tilreiðara er átt við forrit sem framkvæmir bæði orðaskiptingu og setningaskiptingu (sjá að neðan).

Í raun væri best að forritið myndi einnig geta ráðið við texta í öðrum germönskum málum, eins og sænsku, dönsku, norsku, ensku, o.s.frv. Með það í huga þá er best að forritið nýti sér lista yfir skammstafanir í íslensku í mjög takmörkuðum mæli (helst ekkert!), heldur frekar beri kennsl á þær með því að nota reglulegar segðir.

Þið ráðið því hvaða forritunarmál/tól þið notið en mælt er með JFlex og/eða Perl. Athugið að vel kemur til greina að nota bæði málin, t.d. að gera "einfalda" tilreiðingu með JFlex og lesa svo úttakið inn í Perl-forrit sem sér um "flóknari" tilreiðingu sem ekki er leyst á auðveldan hátt með reglulegum segðum (þetta er í samræmi við umræður um tilreiðingu í fyrirlestri). Það er því ekkert því til fyrirstöðu að þið leysið verkefnið með því að búa til röð forritseininga sem keyra hver á eftir annarri. Í því tilviki tekur eining inn skrá sem undanfarandi eining hefur skilað af sér.

Hvernig þið skiptið forritinu í einingar á þó að vera "hulið" fyrir notandanum, þ.e. hann á eingöngu að þurfa að slá inn skipun eins og:

```
tokenise inntak.txt uttak.txt
```

þar sem *inntak.txt* er inntaksskráin og *uttak.txt* er samsvarandi tilreiddur texti (tokenise forritið er þá einhvers konar .bat skrá eða skel sem keyrir nokkrar forritseiningar).

Gróflega séð er verkefnið í tveimur hlutum sem er lýst hér á eftir í köflum 1.1 og 1.2.

1.1 Orðaskiptir (e. word tokeniser)

- Inntak í forritið er textaskrá á "frjálsu sniði", þ.e. hún getur innihaldið eina setningu í hverri línu, margar setningar í hverri línu, einn tóka í hverri línu, nokkra tóka í hverri línu, o.s.frv.
- Úttakið skal vera einn tóki í hverri línu.

1.2 Setningaskiptir (e. sentence segmentiser)

- Inntak í forritið er textaskrá með einum tóka í hverri línu (þ.e. skrá sem Orðaskiptirinn skilar af sér).
- Úttakið skal vera einn tóki í hverri línu ásamt auðri línu til að tákna bil á milli setninga.

2 Markmið og prófun

Markmiðið hjá ykkur ætti að vera að búa til tilreiðara sem er "nokkuð" nákvæmur en hann þarf ekki að vera fullkominn (enda er það illgerlegt!). Fyrir prófanir getið þið t.d. notað textaskrána *visindavefur2.txt* sem er aðgengileg undir "Annað efni" á vef námskeiðsins. Passið ykkur þó að prófa forritið ykkar á öðrum textum, t.d. af www.mbl.is eða erlendum vefsíðum.

3 Skilafrestur

Forritskóða, ásamt dæmum um inntak og tilsvarandi úttak, skal skila í síðasta lagi þriðjudaginn 18. september, kl. 23:59.