Obligatorisk oppgave nr 2

INF2270 – Datamaskinarkitektur

Våren 2010

Oversikt

Oppgaven dreier seg om å skrive fire funksjoner i x86-kode for å håndtere tekster med **Unicode**-tegn kodet i **UTF-8**.[1]

Vi definerer en **UTF-8-tekst** til å være en byte-sekvens av UTF-8-kodete tegn avsluttet av en 0-byte.

Eksempel

Teksten «Ås» består av de to tegnene «Å» (Unicode-kodepunkt **U+00C5** som UTF-8-kodes som to byte $C3_{hex}$ og 85_{hex}) og «s» (Unicode-kodepunkt **U+0073** som UTF-8-kodes som én byte 73_{hex}) og lagres i en UTF-8-tekst som

$C3_{hex}$	85_{hex}	73_{hex}	00_{hex}

Funksjonene

Filen ~inf2270/programmer/Oblig-2/inf2270-utf8.h (også tilgjengelig fra en nettleser som http://www.ifi.uio.no/~inf2270/programmer/Oblig-2/inf2270-utf8.h) inneholder signaturen til de fire funksjonene:

inf2270-utf8.h

```
void latin1_to_utf8 (unsigned char *utf, unsigned char *lat);
void utf8_cat (unsigned char *utf, unsigned int c);
int utf8_get (unsigned char *utf, int pos);
void utf8_to_latin1 (unsigned char *lat, unsigned char *utf);
```

Funksjonen latin1_to_utf8

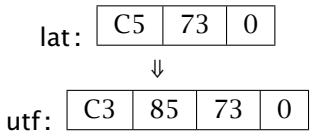
Denne funksjonen omformer en vanlig C-tekst (der tegnene er lagret som Latin-1) til en UTF-8-tekst; parametrene er

unsigned char *utf peker der den nye UTF-8-teksten skal ligge.

unsigned char *lat angir Latin-1-teksten.

[1] En kort introduksjon til Unicode, UTF-8 og Latin-1 finnes i kurskompendiet; dessuten er det meget gode artikler om dette i Wikipedia.

Eksempel



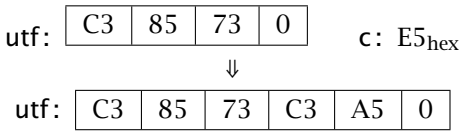
Funksjonen utf8_cat

Denne funksjonen utvider en UTF-8-tekst med ett Unicode-tegn. Parametrene er

unsigned char *utf peker på teksten som skal utvides.

unsigned int c er tegnet som skal settes inn; det er en Unicode-verdi i intervallet $[0, 2^{21}\rangle$.

Eksempel



Funksjonen utf8_get

Denne funksjonen henter frem et gitt tegn i teksten. Parametre er

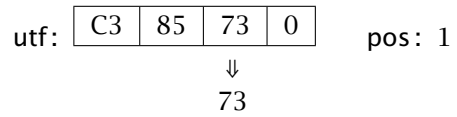
unsigned char *utf er UTF-8-teksten vi skal hente ut tegnet fra.

int pos er posisjonen; det første tegnet i teksten er nr 0.

Hvis oppgitte posisjon ikke finnes i teksten, skal funksjonen returnere -1.

NB! Posisjonen er angitt som en nummerering av UTF-8-tegnene i teksten og *ikke* som et byte-nummer.

Eksempel



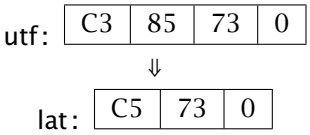
Funksjonen utf8_to_latin1

Denne funksjonen omformer en UTF-8-tekst til en vanlig C-tekst (der tegnene lagres som Latin-1). Hvis UTF-8-teksten inneholder tegn som ikke kan representeres i Latin-1 (dvs ikke ligger i intervallet $[0, 256\rangle$), skal det settes inn et spørsmålstegn i stedet. Funksjonsparametrene er

unsigned char *lat peker dit den nye teksten skal legges.

unsigned char *utf peker på UTF-8-teksten.

Eksempel



Testprogram

Programmet ~inf2270/programmer/Oblig-2/test-utf8.c (også tilgjengelig fra en nettleser som http://www.ifi.uio.no/~inf2270/programmer/Oblig-2/test-utf8.c) kan brukes til å teste funksjonene ganske grundig. (Gruppelærerne vil sikkert bruke dette til å teste innsendte løsninger.)

Annet

- Oppgaven skal løses enkeltvis så vi forventer at innleverte besvarelsene er klart forskjellig fra hverandre. Det kan bli aktuelt å innkalle studenter til en samtale der de kan forklare den koden de har skrevet. Forøvrig gjelder Ifis reglement for obligatoriske oppgaver og krav til innleverte oppgaver.
- Innlevert kode skal skrives i x86-assemblerkode og skal virke på Ifis Linux-maskiner med gcc-kommandoen.
- Dere kan anta at all kode dere ikke selv skriver, oppfører seg perfekt. Dette innebærer blant annet at det aldri vil være ulovlige UTF-8-koder i tekstene (med mindre dere selv gjør noe dumt).

Gode råd

- Tenk en del på hva funksjonene skal gjøre, hvordan de bør skrives og hvilke situasjoner som kan oppstå. I mange tilfeller vil dette gi mye kortere og enklere kode enn en rett-fram-løsning.
- Skriv funksjonene først i et passende høynivåspråk, for eksempel **C**.
- Legg vekt på oversiktlig kode og gode kommentarer. Gruppelæreren kan nekte å rette uforståelig kode.