Tekstanalyse

Obligatorisk oppgave 6 i INF1000

Frist 20.10.2015 kl 14.00

Sammendrag

Målet med oppgaven er å utvikle et verktøy for enkel analyse av tekster som vi leser fra fil. Vi ønsker for eksempel å finne ut hvilke ord som forekommer ofte i en gitt tekst. Vi vil benytte oss av objekter av en klasse `Ord`, som i tillegg til en tekststreng som representerer ordet selv, vil kunne holde på data om hvor mange forekomster det er av dette ordet i teksten. Vi skal også bruke en klasse `Ordliste`. Ordlisteklassen skal holde oversikt over alle ordene (objekter av klassen `Ord`) som forekommer i teksten.

Deloppgave 6.1

Vi skal jobbe med ord, så vi trenger et redskap for det: Objekter av `class Ord` skal representere hvert sitt unike ord fra teksten. Når vi leser inn teksten og kommer til et ord vi ikke har lest før lager vi et nytt objekt som husker ordet, og at det er funnet én gang. Når vi kommer til et ord vi har lest før, øker vi bare antall forekomster i det objektet vi allerede har laget for dette ordet.

Når teksten er ferdig lest inn i programmet vårt kan vi begynne å bruke de verktøyene vi har laget. Ved at like ord i teksten kun lagres én gang i datastrukturen vår sparer vi plass i datamaskinens minne under kjøring, og slipper å telle forekomster på nytt hver gang vi vil vite antall ganger et ord forekommer.

Klassen skal ha følgende grensesnitt:

```
Konstruktør:
  Ord(String tekst) { ... }
    oppretter et Ord-objekt av den gitte teksten.

    Eksempel:
      new Ord("utmark")

Metoder:
  public String toString() { ... }
    returnerer ordet.

    Eksempel:
      new Ord("skog").toString() returnerer  "skog"

  public int hentAntall() { ... }
    henter data om hvor mange ganger ordet forekommer.

    Eksempel:
      Ord forsteOrd = new Ord("grantre");
      forsteOrd.hentAntall()  gir  1.

  public void oekAntall() { ... }
```

```
    registrerer at ordet har forekommet en gang til.

    Eksempel:
      Ord andreOrd = new Ord("bjerk");
      andreOrd.oekAntall();
      andreOrd.hentAntall()  gir  2.
```

Skriv ferdig klassen Ord. Lag også et lite testprogram som viser at klassen fungerer. Testen skal minimum opprette to pekere og objekter av typen Ord, øke antallet på det ene objektet, og skrive ut ordet og antallet for begge objektene.

Synes du denne oppgaven er vanskelig? Se øvingsoppgaver 7.01 og 7.02.

Deloppgave 6.2

Vi trenger også en liste over alle ordene som forekommer i en bok: `class Ordliste`. Klassen skal følgende grensenitt: (Pass på at du har samme returverdier og public/private deklarasjon i ditt program, som det er definert i grensesnittet her.)

```
Metoder:
  public void lesBok(String filnavn) throws Exception { ... }
    leser alle ordene i en fil og legger dem inn i ordlisten.
    Filen må være organisert slik at det ligger akkurat ett ord på hver
    linje i filen.

    Eksempel:
      Ordliste liste = new Ordliste();
      liste.lesBok("scarlet.text");

  private void leggTilOrd(String ord) { ... }
    legger inn et nytt ord i ordlisten hvis det ikke finnes fra
    før. Hvis ordet allerede er der, skal antallet forekomster økes
    med 1.

    NB! Her regnes store og små bokstaver som like, så "asp", "Asp" og
    "ASP" er alle samme ordet.

    Eksempel:
      liste.leggTilOrd("London");

  public Ord finnOrd(String tekst) { ... }
    finner og returnerer et gitt ord s i ordlisten. Hvis ordet ikke finnes, får vi
    null som svar.

    Eksempel:
      if (liste.finnOrd("Edinburgh") == null) { ... }

  public int antallOrd() { ... }
    finner ut og returnerer hvor mange ulike ord det finnes i ordlisten.

    Eksempel:
      if (liste.antallOrd() > 0) { ... }

  public int antallForekomster(String tekst) { ... }
    finner ut og returnerer hvor mange ganger ordet "tekst" forekommer i ordlisten.

    Eksempel:
      if (liste.antallForekomster("Watson") >= 100) { ... }
```

```
public Ord vanligste() { ... }
  finner og returnerer det ordet som forekommer oftest i boken. Hvis det er flere
  ord som forekommer flest ganger, holder det å finne ett av dem.

  Eksempel:
    System.out.println("Vanligste ord er " + liste.vanligste().toString());

[Frivillig:]
public Ord[] alleVanligste() { ... }
  finner og returnerer det eller de ordene som forekommer flest ganger i boken.
```

Du skal benytte deg av en ArrayList<Ord> for å oppbevare ord i ordlisten.

Skriv ferdig klassen `Ordliste` og lag også et lite testprogram.

Synes du denne oppgaven er vanskelig? Se øvingsoppgaver 7.03 og 7.04.

Deloppgave 6.3

Vi skal teste programmet vårt på den første Sherlock Holmes-boken *A study in scarlet* av Sir Arthur Conan Doyle. Last ned filen `scarlet.text` som inneholder boken formatert riktig (dvs ett ord på hver linje).

Lag et program `Oblig6` som benytter klassene `Ord` og `Ordliste` til å beregne og skrive ut følgende informasjon om denne boken:

a) Hvor mange ulike ord forekommer i boken?

b) Hvor mange ganger forekommer ordet *Holmes*?

c) Hvor mange ganger forekommer *elementary*?

d) Hvilket ord forekommer flest ganger?

Tips Det tar noen sekunder å sjekke hele boken, så du kan spare mye tid på å lage en kort testfil du bruker til programmet er ferdig.

Synes du denne oppgaven er vanskelig? Se øvingsoppgaver 6.3.1, 6.3.2 og 6.3.3.

Fremgangsmåte for innleveringer

1. Lag en fil som heter README.txt. Følgende spørsmål skal være besvart i filen:
   - Hvordan synes du innleveringen var? Hva var enkelt og hva var vanskelig?
   - Hvor lang tid (ca) brukte du på innleveringen?
2. Logg inn på Devilry.
3. Lever alle .java-filene og README.txt.
4. Husk å trykke lever og sjekk deretter at innleveringen din er komplett.

Ifis regler om obligatoriske oppgaver og andre innleveringer[1]

Ved alle pålagte innleveringer av oppgaver ved Ifi, enten det dreier seg om obligatoriske oppgaver, hjemmeeksamen eller annet, kreves det at arbeidet er et resultat av studentens egen innsats.

Krav til innleverte oppgaver

Å utgi andres arbeid for sitt eget er både ulovlig og uetisk og vil medføre sterke reaksjoner fra IFIs og Universitetets side, for eksempel utvisning i ett eller flere semestre; se *Rutiner for behandling av fuskesaker* på http://www.uio.no/om/regelverk/studier/studier-eksamener/fuskesaker/.

Vær derfor oppmerksom på:

- Hvis du tar med tekst, programkode, illustrasjoner og annet som andre har laget, må du tydelig merke det og angi hvor det kommer fra.
- Det er greit å få hint om hvordan en oppgave kan løses, men dette skal eventuelt brukes som grunnlag for egen løsning og ikke kopieres uendret inn.
- Du kan bli innkalt til samtale om dine innleveringer. Du må da kunne forklare innholdet i detalj og redegjøre for hvorledes det innleverte arbeidet er blitt til. Dersom samtalen avdekker at oppgaven ikke er et selvstendig arbeid og/eller du ikke kan gjøre rede for det innleverte arbeidet ditt, kan faglærer velge å ikke godkjenne oppgaven, også selv om gruppelærer har godkjent oppgaven i forkant av samtalen.

Gruppearbeid

I noen emner skal det leveres gruppearbeid. Ifi krever da at alle medlemmer av gruppen kan gjøre rede for hovedtrekkene i det innleverte arbeidet. Dessuten må alle ha utført en rimelig del av det hele, og kunne identifisere og gjøre rede for i detalj sin del.

Samarbeid ved individuelle innleveringer

Disse kravene betyr ikke at Ifi fraråder samarbeid —– tvert imot. Ifi oppfordrer studentene til å utveksle faglige erfaringer. I tilfeller hvor det skal leveres individuelle oppgaver er følgende viktig:

- Du kan diskutere en løsning sammen med andre, men dere skal ikke dele noen deler av løsningen (f eks ved å levere inn lik kode hvor kun variabelnavn er byttet ut).
- Dersom du tar med tekst, programkode, illustrasjoner og annet som andre har laget, må du tydelig merke det og angi hvor det kommer fra — i en selvstendig oppgave er dette noe som sjelden skal forekomme. Hvis du er i tvil om hva som er lovlig samarbeid, må du kontakte gruppelærer eller faglærer.

[1] Disse reglene er hentet fra http://www.mn.uio.no/ifi/studier/admin/obliger/.

Retningslinjer for obligatoriske oppgaver

En obligatorisk oppgave er en oppgave som må besvares og godkjennes for å kunne gå opp til avsluttende eksamen i et emne. Besvarelsen på en obligatorisk oppgave vurderes til godkjent eller ikke godkjent og kan ikke inngå i karaktervurderingen i emnet.

Antall obligatoriske oppgaver som gis i et emne samt tidspunktet for offentliggjøring og innlevering av oppgavene skal være klart på semestersiden for emnet ved studiestart.

Av oppgaveformuleringen skal det gå tydelig frem hva som forventes av innleveringen. Den faglige bakgrunnen som er nødvendig for å kunne løse oppgaven må være forelest eller gjort tilgjengelig på annen måte i god tid før innleveringsfristen.

Retningslinjene i sin helhet finnes på https://www.mn.uio.no/ifi/studier/admin/obliger/oblig-retningslinjer.html.