IN1140

Eksamen H2020

Informasjon

Eksamenssettet inneholder 12 oppgaver.
Alle hjelpemidler er tillatt (lærebok, nettressurser, notater osv.) Det er ikke tillatt å samarbeide eller kommunisere med andre under eksamen om oppgavene.
Oppgaven har én oppgave som krever filopplasting.

---

1 Regulære uttrykk (15 poeng)

Du vil kjøpe en togreise fra Oslo S til Bodø. Du er fleksibel når det gjelder dato, så lenge det er før (men ikke inkludert) 01.02.2021. Du vil ha en billett som koster maks 1999kr, og vil reise fra Oslo S tidligst kl. 08:00 og seinest kl. 22:00. Du ønsker også å se prisene for både Standard og Premium billetter.

Skriv et regulært utrykk som fanger resultatene av søket ditt slik at resultatene alltid er skrevet i følgende format:

```
ukedag dag måned år–nylinje–
prisklasse tid billettype
```

Merk følgende formateringer:

- dag skal være melllom 01 og 31.
- måned skal være november, desember, og januar.
- år skal være 2020 eller 2021.
- prisene skal være mellom 1 og 1999 kroner. Merk at
- prisen kan bli skrevet som kr eller nok
- klokkeslett skal alltid starte med kl. og være mellom 08 og 22. Minutter skal være mellom 00 og 59. Klokkeslett skal være skrevet som kl. mellomrom time:minutter.
- billettype skal være enten Standard eller Premium.
- det er mellomrom mellom hver element, utenom etter år, der skal det være ny linje.

For eksempel skal følgende strenger gjenkjennes av ditt regulære utrykk:

1. `mandag 01 desember 2020`
   `299kr kl.  09:30 Standard`
2. `fredag 21 januar 2021`
   `1199kr kl.  21:30 Premium`
3. `torsdag 31 januar 2021`
   `1999nok kl.  22:00 Standard`

Men følgende strenger skal ikke gjenkjennes:

1. `lørdag 01 februar 2021`
   `299kr kl.  20:30 Standard`
2. `søndag 01 desember 2020`
   `2299kr kl.  09:30 Standard`
3. `tirsdag 31 januar 2021`
   `1999nok kl.  22:01 Standard`
4. `fredag 21 januar 2021`
   `1199kr kl.21:30 Flex`

Løsningsforslag

Eksempelsvar:

```
regex = r'(man|tir|ons|tors|fre|lør|søn)dag\s((0[1-9]|[12][0-9]|3[01])\s
(desember\s2020|januar\s2021)|(0[1-9]|[12][1-9]|30)\snovember\s2020)\n
([1-9][0-9]{0,2}|1[1-9][0-9]{0,2})(kr|nok)\skl\.\s
(((0[8-9]|1[0-9]|2[0-1]):[0-5][0-9])|22:00)\s(Standard|Premium)'
```

---

2 Bøyning og orddanning (8 poeng)

Ta utgangspunkt i følgende tekst og besvar spørsmålene under:

`Det var en gang en far som het snekker Andersen, og han hadde mange unger slik som farer bruker å ha, og så var det en julekveld at han lista seg ut`

mens ungene og fru snekker Andersen satt og knekte nøtter for å spise filipine. Han skulle nedi vedskjulet sitt for der hang det en julenissedrakt, og på ei kjelke lå det en stor sekk med julegaver. Så tok snekker Andersen på seg julenissedrakten og dro kjelken med julegavesekken ut på gardsplassen.

1. Finn to eksempler på bøyning i teksten. Hvilken kategori er ordene bøyd for?
2. Finner du noen eksempler på orddanning (avledning eller sammensetning)? Illustrér med eksempler.

Løsningsforslag

1. Eksempelsvar: Noen eksempler på bøyning er feks substantivet "unge" som er bøyd for ubestemt flertall "unge+er" og bestemt flertall "unge+ene". Andre eksempler: "lista" preteritumsformen av "liste", "farer" presens/nåtid av "fare", "julegavesekken" bestemt form av "julegavesekk".
2. Det er flere eksempler på sammensetninger i teksten, eksempelvis "julegave+sekken" eller "julenissedrakt".

---

3 Ordklassekriterier (10 poeng)

Denne oppgaven omhandler **ordklassekriterier**. Ta utgangspunkt i teksten fra forrige oppgave, gjentatt under:

Det var en gang en far som het snekker Andersen, og han hadde mange unger slik som farer bruker å ha, og så var det en julekveld at han lista seg ut mens ungene og fru snekker Andersen satt og knekte nøtter for å spise filipine. Han skulle nedi vedskjulet sitt for der hang det en julenissedrakt, og på ei kjelke lå det en stor sekk med julegaver. Så tok snekker Andersen på seg julenissedrakten og dro kjelken med julegavesekken ut på gardsplassen.

Besvar følgende spørsmål:

1. Gi en kort beskrivelse av de tre vanligste kriteriene for tildeling av ordklasse.
2. Ta deretter for deg teksten om snekker Andersen og vis hvordan du kan benytte de tre kriteriene for å tildele ordklasse til to ulike ord hentet fra teksten.

Løsningsforslag

1) Eksempelsvar:

De tre ordklassekriteriene er som følger:

- formelle (morfologiske) kriteriet: hvordan bøyes ordet?
- funksjonelle (syntaktiske) kriteriet: hvordan kombineres ordet med andre ord?
- betydningsmessige (semantiske) kriteriet: hva betyr ordet?

2) Eksempelsvar (besvarelsen ber kun om to, her angis flere for å illustrere):

- "unger"
  - formelt: bøyd med flertallsendelse -er, kan bøyes for bestemthet "ungene"
  - funksjonelt: forekommer etter determinativ "mange unger", eller "mange bråkete unger"
  - betydning: levende vesen

  → alle kriteriene peker i retning SUBSTANTIV
- "knekte"
  - formelt: bøyes i fortid: "knekker - knekte"
  - funksjonelt: fungerer som predikat, feks "de knekte nøtter"
  - betydning: betegner en handling

  → alle kriteriene peker i retning VERB
- "han"
  - forrmelt: kan bøyes for kasus "han - ham"
  - funksjonelt: kan erstatte en NP "Julenissen spiste vs han spiste"
  - betydning: henter betydning fra sammenhengen

  → alle kriteriene peker i retning PRONOMEN
- "ei"
  - formelt: bøyes ifht kjønn ei/en/et
  - funksjonelt: bestemmer til substantiv "ei kjelke"
  - betydning: bestemmer substantivets referanse

  → alle kriteriene peker i retning DETERMINATIV

- ”på”
  - formelt: ubøyelig
  - funkjsonelt: tar substantiv argument ”på ei kjelke”
  - betydning: betegner beliggenhet

  → alle kriteriene peker i retning PREPOSISJON

---

4 Grammatikk (4 poeng)

Under ser du en liten kontekstfri grammatikk for norsk:

```
S -> NP VP
VP -> V
VP -> V NP
NP -> julenissen, snekkeren, ungene, kjelken
V -> danset, snekret, så, trakk
```

Grammatikken er langt fra noen komplett grammatikk for norsk. Gi ett eksempel på en grammatisk, norsk setning som ikke gis noen analyse av denne grammatikken, samt ett eksempel på en ugrammatisk setning som gis en analyse.

Løsningsforslag

Eksempelsvar:
En grammatisk setninge som ikke gis en analyse vil feks være “julenissen tror at ungene danset” eller “en julenisse danset”
En ugrammatisk setninge som gis en analys kan feks være “julenissen danset ungene”, eller “snekkeren så”

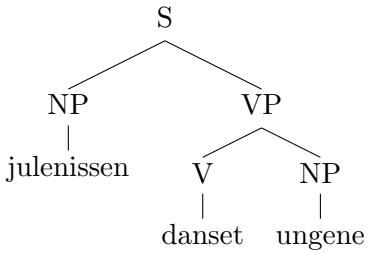
---

5 Syntaktisk tre (4 poeng)

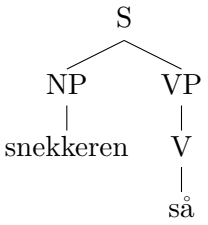
Last opp det syntaktiske treet som grammatikken tildeler den ugrammatiske setningen fra forrige oppgave.

```
S -> NP VP
VP -> V
VP -> V NP
NP -> julenissen, snekkeren, ungene, kjelken
V -> danset, snekret, så, trakk
```

Løsningsforslag



eller



6 Utvidelse av grammatikken (6 poeng)

Gitt den samme grammatikken som tidligere (gjentatt under):

```
S -> NP VP
VP -> V
VP -> V NP
NP -> julenissen, snekkeren, ungene, kjelken
V -> danset, snekret, så, trakk
```

ønsker vi å kunne gi en analyse av følgende setninger:

julenissen og snekkeren snekret
julenissen og snekkeren trakk kjelken og ungene
julenissen og snekkeren og ungene danset
julenissen og snekkeren og ungene og kjelken danset

Hvordan kan vi utvide grammatikken vår slik at disse gis en analyse? Du skal her benytte deg av en rekursiv regel som lar deg analysere NP'er av uviss lengde (i tillegg til eventuelle leksikale regler).

Løsningsforslag

Eksempelsvar: Vi legger til følgende regler:

```
NP -> NP C NP
C -> og
```

---

7 Utvidelse av grammatikken (6 poeng)

Vi vil at grammatikken vår videre skal kunne gi en analyse for setninger som:

snekkeren snekret på julaften
snekkeren trakk kjelken på julaften
snekkeren danset på stuegulvet på julaften

Hvordan vil du utvide grammatikken for å få til dette? Husk å angi både syntaktiske og leksikalske regler.

Løsningsforslag

Eksempelsvar: (Vi legger til preposisjonsfraser i VP)

```
VP -> VP PP
PP -> P NP
P -> på
NP -> stuegulvet, julaften
```

---

8 Flertydighet (10 poeng)

Flertydighet er en egenskap som kjennetegner naturlige språk på alle lingvistiske nivåer.

1. Diskutér denne påstanden og illustrér med eksempler.
2. Hvordan kan vi håndtere flertydighet i språkteknologi? Ta utgangspunkt i én språkteknologisk oppgave og utdyp svaret ditt.

Løsningsforslag

Eksempelsvar (noe fyldigere enn forventet):

1. Vi finner flertydighet på alle lingvistiske nivåer, eksempelvis kan ett ord ha flere mulige ordklasser. Det norske ordet "rett" feks kan ha ordklassene substantiv, adjektiv og verb. Med ulik ordklasse kommer også ulik betydning, og det samme ordet "rett" vil ha en rekke betydninger, eksempelvis (og oversatt til engelsk for å illustrere) "straight", "court", "correct".

   I naturlige språk finner vi også syntaktisk flertydighet, dvs at én og samme setning vil kunne gi opphav til flere syntaktiske analyser. Et eksempel på dette er "Jenta så mannen med teleskopet", der setningen kan gis en tolkning som tilsvarer at mannen blir sett gjennom teleskopet, alternativt at mannen bærer et teleskop og disse to tolkningene vil gis ulik syntaktisk analyse.

2. Alle språkteknologiske oppgaver må håndtere flertydighet. Vi skal her ta utgangspunkt i ordklassetagging. Tidlige regelbaserte systemer for ordklassetagging kunne gi opphav til flere mulige analyser. Statistiske metoder brukes i utstrakt grad for å håndtere flertydighet i dag og her vil man kunne tildele ulik sannsynlighet til ulike ordklasser basert på modeller trent på treningsdata, dvs data som er manuelt annotert med ordklasseinformasjon.

---

9 Leksikalske relasjoner (6 poeng)

Ta utgangspunkt i ordet *nisse* og vis hvordan ordet kan inngå i to ulike leksikalske relasjoner. For hver av relasjonene skal du navngi relasjonen og vise ord-paret som illustrerer relasjonen.

Løsningsforslag

Eksempelsvar (merk at svaret kun skal inneholde to relasjoner):

```
skjegg - nisse: meronymi
nisse - vesen: hyponymi
nisse - gnom: synonymi
```

1	nn	alle døma er gode
2	nn	fiskerne fekk mykje fisk i dag
3	nn	eg elskar sol og varme
4	nb	du har satt et dårlig eksempel
5	nb	jeg liker å spise kransekake
6	nb	det er ikke for mye sol
S1	?	eg åt for mykje kransekake
S2	?	eg liker kransekake

10 Naive Bayes klassifisering (20 poeng)

I denne oppgaven har vi et lite utvalg setninger som hører til klassene nynorsk (nn) og bokmål (nb).
Gitt de to nye test-setningene S1 og S2 bruk Naive Bayes-formelen til å klassifisere test-setningene S1 og S2.

Her skal du:

1. Regne ut sannsynlighetene for de forskjellige ordene. Du trenger bare å regne ut for ordene i test-setningene. **Bruk** glatting.

2. Regne ut hvilken verdi som er størst. Blir setningene klassifisert som nynorsk eller bokmål?

3. Er klassifiseringen av setningene S1 og S2 riktig? For begge setningene? Hvis den er det, begrunn dette. Hvis ikke, forklar årsaken til feil klassifisering og gi forslag til hvordan vi kan unngå slike feil.

Løsningsforslag

P(nn) = $\frac{3}{6}$

P(nb) = $\frac{3}{6}$

1: P(eg|nn)= $\frac{1+1}{15+30}$

P(åt|nn)= $\frac{0+1}{15+30}$

P(for|nn)= $\frac{0+1}{15+30}$

P(mykje|nn)= $\frac{1+1}{15+30}$

P(kransekake|nn)= $\frac{0+1}{15+30}$

—

P(eg|nb)= $\frac{0+1}{17+30}$

P(åt|nb)= $\frac{0+1}{17+30}$

P(for|nb)= $\frac{1+1}{17+30}$

P(mykje|nb)= $\frac{0+1}{17+30}$

P(kransekake|nb)= $\frac{1+1}{17+30}$

—

P(nn)P(S|nn) = $\frac{3}{6} * \frac{4}{45^5}$ = 0,000000011 = $1.1 * 10^{-8}$

P(nb)P(S|nb) = $\frac{3}{6} * \frac{4}{47^5}$ = 0,000000009 = $0.9 * 10^{-8}$

Blir klassifisert som nn

2: P(eg|nn)= $\frac{1+1}{15+30}$

P(liker|nn)= $\frac{0+1}{15+30}$

P(kransekake|nn)= $\frac{0+1}{15+30}$

—

P(eg|nb)= $\frac{0+1}{17+30}$

P(liker|nb)= $\frac{1+1}{17+30}$

P(kransekake|nb)= $\frac{1+1}{17+30}$

—

P(nn)P(S|nn) = $\frac{3}{6} * \frac{2}{45^3}$ = 0,000010974 = $10.97 * 10^{-6}$

P(nb)P(S|nb) = $\frac{3}{6} * \frac{4}{47^3}$ = 0,000019264 = $19.26 * 10^{-6}$

Blir klassifisert som nb

3: S1 er riktig klassifisert, S2 er feil klassifisert. Det er hovedsakelig pga datamangel, og at Naive Bayes er naive. Her gis det poeng for god diskusjon og argumentasjon.

---

11 Named Entity Recognition (7 poeng)

Anta følgende tekst (fra Kaptein Sorte Bill av Thorbjørn Egner):

```
Jeg er kaptein Sorte Bill fra femten hundre og fjorten,
hei fadderi fadderullan dei,
en sjørøverkap'ten av den gamle gode sorten,
hei fadderi fadderullan dei.

Skuta heter Klara og er very well bekrutta
hei fadderi fadderullan dei,
```

```
Og kjent og fryktet var'a ifra Moss og til Calcutta,
hei fadderi fadderullan dei.
```

1. Hva er ord-for-ord klassifisering, så kalt BIO klassifisering, i Named Entity Recognition?
2. Du skal nå hente ut egennavn fra teksten over og klassifisere dem. Du kan her benytte deg av følgende kategorier: PER (person), ORG (organisajon), LOC (location), DT (dato), GPE (geopolitical entity) .
3. Er det noen av entitetene som ikke passer inn i noen av de oppgitte kategoriene? Hva er i såfall grunnen til det?

Løsningsforslag

1. BIO-klassifisering: taggen indikerer om ordet befinner seg i begynnelsen(B), innenfor (I) eller utenfor (O) et egennavn, samt indikerer kategori.
2. Person: Sorte Bill
   Date: femten hundre og fjorten
   GPE: Moss
   GPE: Calcutta
3. Klara identifiseres som vanskelig men skal ikke tagges som person da den her refererer til et skip

---

12 Coreference Resolution (4 poeng)

1. Hva er Coreference Resolution?
2. Hvorfor er det en vanskelig oppgave i språkteknologi?

Løsningsforslag

1. Det er å automatisk finne uttrykkene i en tekst som refererer til samme person eller ting; samme referent, f.eks. Jon sa at han ville komme; selve substantivet Jon og pronomenet han refererer til samme person, nemlig Jon.
2. Det er en vanskelig oppgave f.eks. fordi samme pronomen kan referere til forskjellige personer eller steder.