Informasjon

Eksamen i IN1140

Introduksjon til språkteknologi

25 november kl. 09:00 - 13:30 (4,5 timer)

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.
httlps://www.uio.no/studier/eksamen/innlevering/alternativer-for-handtegninger.htm

Man kan trekkes ut til samtale for å kontrollere eierskap til sin besvarelse
https://www.mn.uio.no/om/hms/koronavirus/kontrollsamtale/
Samtalen har ikke innvirkning på sensuren/karakteren, men kan lede til at instituttet oppretter fuskesak. Les mer om hva som regnes som fusk på UiOs nettsider:
https://www.uio.no/om/regelverk/studier/studier-eksamener/fuskesaker/

For øvrig gjelder informasjonen på nettsiden om eksamensavvikling ved MN høsten 2020:
https://www.mn.uio.no/om/hms/koronavirus/eksamen-2020.html

Kontaktinfo:
https://www.mn.uio.no/om/hms/koronavirus/brukerstotte/brukerstotte-eksamen-h20.html

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\nylinjeprisklasse 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 2020299kr kl. 09:30 Standard`**
2. **`fredag 21 januar 20211199kr kl. 21:30 Premium`**
3. **`torsdag 31 januar 20211999nok kl. 22:00 Standard`**

Men følgende strenger skal ***ikke*** gjenkjennes:

1. **`lørdag 01 februar 2021299kr kl. 20:30 Standard`**
2. **`søndag 01 desember 20202299kr kl. 09:30 Standard`**
3. **`tirsdag 31 januar 20211999nok kl. 22:01 Standard`**
4. **`fredag 21 januar 20211199kr kl.21:30 Flex`**

Skriv ditt svar her

Maks poeng: 15

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.

Skriv ditt svar her

Maks poeng: 8

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.

Maks poeng: 10

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.

Skriv ditt svar her

Maks poeng: 4

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

Maks poeng: 4

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).
Skriv ditt svar her

Maks poeng: 6

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.
Skriv ditt svar her

Maks poeng: 6

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.

Skriv ditt svar her

Maks poeng: 10

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.

Skriv ditt svar her

Maks poeng: 6

10 Naive Bayes klassifisering (20 poeng)

I denne oppgaven har vi et lite utvalg setninger som hører til klassene nynorsk (nn) og bokmål (nb).

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

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.

Skriv ditt svar her

Maks poeng: 20

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?

Skriv ditt svar her

Maks poeng: 7

12 Coreference Resolution (4 poeng)

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

Skriv ditt svar her

Maks poeng: 4