

i Om eksamen

Eksamen i IN1140/INF1820, høst 2017

Tid:

Fredag 15. desember kl 14.30-18.30

Hjelpemidler:

Ingen.

Før du begynner:

Vi anbefaler å lese gjennom hele oppgaveteksten før du begynner. Hvis du føler du mangler informasjon for å løse en oppgave, gjør dine egne antakelser og redegjør for dem. Ca 30–90 minutter etter påbegynt eksamen vil det bli anledning til å spørre en faglærer om eventuelle oppklaringer til oppgaveteksten. Du kan svare på enten norsk eller engelsk.

1 Regulære uttrykk (2 poeng)

Hvilken av de følgende epost-adressene kan **ikke** gjenkjennes med det følgende regulære uttrykket:

`[aA-zZ0-9\._%+~]+@[aA-zZ0-9\.-]+\.[aA-zZ]{1,2}`

Velg ett alternativ

- olanorman@meg.epost.no
- Ola.Norman@epost.com
- Ola.Norman@epost.no
- OlaNorman@meg.co

Maks poeng: 2

2 Regulære uttrykk for verb (6 poeng)

Skriv ett regulært uttrykk som gjenkjenner formene imperativ, infinitiv, presens, og preteritum av følgende norske verb: spise, kjøpe, tenke, løpe.

Skriv ditt svar her

Maks poeng: 6

3 Ordklasser (5 poeng)

Her skal vi jobbe med følgende setning:

Jeg ser mannen med kikkerten.

Gitt ordklassene i Tabell 1 under, tildel ordklasser til alle ordene i setningen. Du må velge ett alternativ for hvert tilfelle.

Tabell 1

CC	konjunksjon
DET	determinativ
JJ	adjektiv
NN	substantiv
PR	preposisjon
PO	pronomen
RB	adverb
SB	subjunksjon
VB	verb

Finn riktig ordklasse

	CC	VB	RB	PR	SB	NN	DET	JJ	PO
Jeg	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
med	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
ser	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
mannen	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
kikkerten	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Maks poeng: 5

4 Grammatikk og strukturell flertydighet (10 poeng)

Setningen i vårt forrige eksempel er strukturelt flertydig. Definer en kontekstfri grammatikk med regler som kan vise ulike analyser av denne setningen. Altså:

Jeg ser mannen med kikkerten.

Ta også stilling til om grammatikken din er rekursiv. Begrunn svaret ditt.

Skriv ditt svar her

Maks poeng: 10

5 Leksikalsk semantikk/relasjoner (7 poeng)

1. Forklar forskjellen mellom homonymi og polysemi. Gi eksempler på begge.
2. Forklar og gi eksempler på relasjonen hyponymy.
3. Forklar og gi eksempler på relasjonen meronymy.

Skriv ditt svar her

Maks poeng: 7

6 Komposisjonalitet (5 poeng)

Forklar med noen få setninger hva vi mener med at setningssemantikk er komposisjonell (*compositional*)?

Skriv ditt svar her

Maks poeng: 5

7 Semantiske roller (5 poeng)

Gitt følgende setninger:

"**Markus** ryddet **lekene**. Plutselig, kastet Nora **brannbilen** på ham og skadet ham. **Edna** så hva som skjedde. Hun ropte på mor som måtte komme og rense såret med **et antibakterielt middel**."

Angi de semantiske rollene for ordene i tabellen under. Du må velge ett alternativ for hvert tilfelle.

Finn de som passer sammen

	EXPERIENCER	BENEFICIARY	INSTRUMENT	AGENT	PATIENT
lekene	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Edna	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
et antibakterielt middel	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Markus	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
brannbilen	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Maks poeng: 5

8 Maximum likelihood estimering (MLE) (5 poeng)

For å estimere sannsynligheter i n-gram språkmodeller og HMM-modeller har vi benyttet oss av såkalt *maximum likelihood estimering* (MLE). Vis formlene for hvordan vi fra et korpus estimerer sannsynligheten for et ord w_i gitt det foregående ordet w_{i-1} i en sekvens, altså $P(w_i|w_{i-1})$.

Skriv ditt svar her

Maks poeng: 5

9 Glatting (smoothing) (5 poeng)

Å bruke MLE (*maximum likelihood estimation*) alene kan by på problemer. For språkmodeller utvider vi gjerne MLE med såkalt glatting (smoothing).

Forklar kort hvorfor MLE alene byr på problemer. Forklar også generelt hva glatting er, og hvorfor vi bruker det.

Skriv ditt svar her

Maks poeng: 5

10 Transisjonssannsynligheter og emmisjonsannsynligheter (5 poeng)

Forklar kort hva vi i en HMM (Hidden Markov Model) mener med *transisjonssannsynligheter* og *emmisjonsannsynligheter* (observasjonssannsynligheter). Vis også formelene som viser hvordan de kan estimeres fra et treningskorpus.

Skriv ditt svar her

Maks poeng: 5

11 HMM-tagging (5 poeng)

Vis formelen som vi forsøker å maksimere i HMM-tagging, altså for å finne den mest sannsynlige taggsekvensen gitt en ordsekvens. Kommenter kort hvilke forenklingene antagelser som gjøres i formelen.

Skriv ditt svar her

Maks poeng: 5

12 Metodologiske paradigmer (5 poeng)

I språkteknologi kan vi grovt skille mellom regelbaserte metoder og empiriske metoder basert på maskinlæring. Forklar med noen få setninger hva som karakteriserer forskjellene.

Skriv ditt svar her

Maks poeng: 5

13 Lingvistiske nivåer (5 poeng)

Hvilke nivåer av analyse opererer vi med i lingvistikk? Forklar også helt kort (stikkord er nok) hva som er fokuset for analyse på hvert nivå.

Skriv ditt svar her

Maks poeng: 5

14 Named Entity Recognition (5 poeng)

Named Entity Recognition (NER) er automatisk identifisering og kategorisering av egennavn. Forklar kort hvorfor oppslag i en navneliste alene er ikke den beste måten å gjøre NER på. Du kan bruke følgende setning for inspirasjon:

Flyet lander på John F. Kennedy i kveld.

Skriv ditt svar her

Maks poeng: 5

15 Språkmodeller (5 poeng)

Hva er en språkmodell? Forklar med noen få setninger. Nevn også noen eksempler på anvendelser av slike modeller.

Skriv ditt svar her

Maks poeng: 5

16 Informasjonsgjenfinning (5 poeng)

Hvordan kan et dokument representeres som en vektor? Forklar med noen få setninger. Nevn noen eksempler på hvordan en slik representasjon kan være nyttig.

Skriv ditt svar her

Maks poeng: 5

i Introduksjon

Her skal vi tenke oss at vi ønsker å utvikle et system som automatisk skal skille mellom to forskjellige betydninger av substantivet "bønner". Vi antar at dette substantivet har to hovedbetydninger: FRØ og RELIGION. Vedlagt finner du et lite korpus som er annotert med betydningsinformasjon.

Vi ønsker å beregne den mest sannsynlige betydningen \mathbf{b} (fra en mengde mulige betydninger \mathbf{B}) for en forekomst av ordet "bønner", gitt en n -dimensjonal trekkvektor \mathbf{v} som beskriver forekomsten. Vi bestemmer oss for å løse denne oppgaven ved Naive Bayes-klassifisering, angitt ved vedlagte formel.

$$\hat{\mathbf{b}} = \operatorname{argmax}_{\mathbf{b} \in \mathbf{B}} P(\mathbf{b}) \prod_{j=1}^n P(v_j | \mathbf{b})$$

Korpus:

Hun liker bønner. (FRØ)

Jeg sender mine bønner til gud. (RELIGION)

Jeg liker bønner i tomatsaus. (FRØ)

I noen religioner finnes obligatoriske bønner. (RELIGION)

17 Sannsynlighet (5 poeng)

Hva er (prior) sannsynligheten for betydningen **RELIGION** $\in \mathbf{B}$ i korpuset, altså **P(RELIGION)**?

$$\hat{b} = \operatorname{argmax}_{b \in B} P(b) \prod_{j=1}^n P(v_j | b)$$

Korpus:

Hun liker bønner. (*FRØ*)

Jeg sender mine bønner til gud. (*RELIGION*)

Jeg liker bønner i tomatsaus. (*FRØ*)

I noen religioner finnes obligatoriske bønner. (*RELIGION*)

Skriv ditt svar her

Maks poeng: 5

18 Sannsynlighet for individuelle trekk (5 poeng)

Hvordan beregner vi sannsynligheten for et individuelt trekk, gitt en betydning **P(v_j|b)**? Anta en bag-of-words representasjon av forekomstene i korpuset, slik at ordet "liker" er et av trekkene. Hva blir da **P(liker|FRØ)**?

$$\hat{b} = \operatorname{argmax}_{b \in B} P(b) \prod_{j=1}^n P(v_j | b)$$

Hun liker bønner. (*FRØ*)

Jeg sender mine bønner til gud. (*RELIGION*)

Jeg liker bønner i tomatsaus. (*FRØ*)

I noen religioner finnes obligatoriske bønner. (*RELIGION*)

Skriv ditt svar her

Maks poeng: 5

19 Naive Bayes-formelen (5 poeng)

Forklar hvordan Naive Bayes-formelen er utledet fra den opprinnelige sannsynligheten $P(\mathbf{b}|\mathbf{v})$ og beskriv hvilke antagelser som er gjort.

$$\hat{\mathbf{b}} = \operatorname{argmax}_{\mathbf{b} \in B} P(\mathbf{b}) \prod_{j=1}^n P(v_j | \mathbf{b})$$

Korpus:

Hun liker bønner. (*FRØ*)

Jeg sender mine bønner til gud. (*RELIGION*)

Jeg liker bønner i tomatsaus. (*FRØ*)

I noen religioner finnes obligatoriske bønner. (*RELIGION*)

Skriv ditt svar her

Maks poeng: 5