Løsningsforslag eksamen IN1140 2024

**Merk at for tekstoppgavene er løsningsforslaget nettopp et *forslag* til løsning. Det vil typisk være mulig å løse oppgaven på mange ulike måter og ved bruk av ulike formuleringer.*

UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

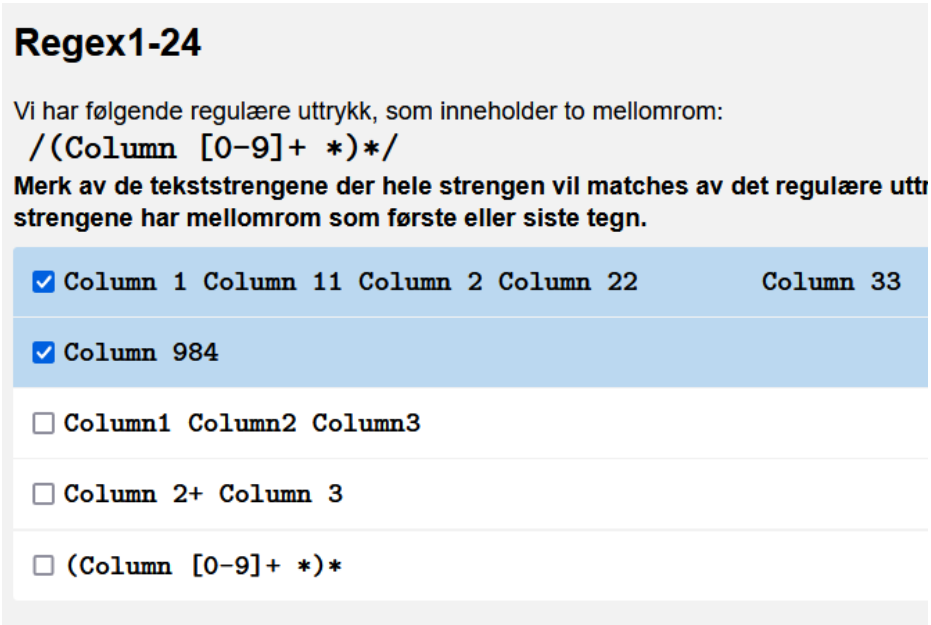
Skriftlig eksamen i IN1140
2024 HØST
Varighet: 28.11 kl 15:00 til 28.11 kl 19:00
Tillatte hjelpemidler: Ingen. En kalkulator er tilgjengelig i Inspera nederst på siden.

Det er viktig at du leser hele forsiden før du begynner å besvare eksamensoppgavene.

Til slutt i oppgaven finnes det et blankt dokument der eventuelle kommentarer til flervalgsoppgavene kan legges til.

Totalt antall mulige poeng for denne oppgaven er 100 poeng.

Regulære uttrykk



RegEx2-24

Kva for nokre e-postadresser kan kjennast igjen med det følgjande regulære uttrykket:
`/[a-zA-Z0-9\._%+-]+@[a-zA-Z0-9\.-]+\.[a-zA-Z]{2,}/`

Vel eitt eller fleire alternativ

- [x] OlaNorman@m.co ✓
- [x] olanorman@meg.epost.no ✓
- [x] Ola.Norman@epost.com ✓
- [ ] 2+2=4@math.styx

Regex 3:

Kva for eitt av alternativa utgjer eit slikt regulært uttrykk?
Vel eitt alternativ

- (•) `/([A-Z]{3})([0-9]+)\+\1\2/` ✓
- ( ) `/([A-Z]*3[0-9]*3)\+(\1\2)/`
- ( ) `/([A-Z]*3)([0-9]+)\+\1*3\2/`
- ( ) `/[A-Z]{3}([0-9]+)\+[A-Z]{3}\1/`
- ( ) `/([A-Z]{3})\1\+([0-9]+)\2/`

Regex 4

```
/=\d{1,4}([+\-*]\d{1,4}){1,3}/
```

Ord og ordklasser

5 Lingvistiske nivåer (6 poeng)

I løpet av året har vi tatt for oss flere lingvistiske nivåer. Navngi 3 lingvistiske nivåer og gi en kort beskrivelse av hver av dem.

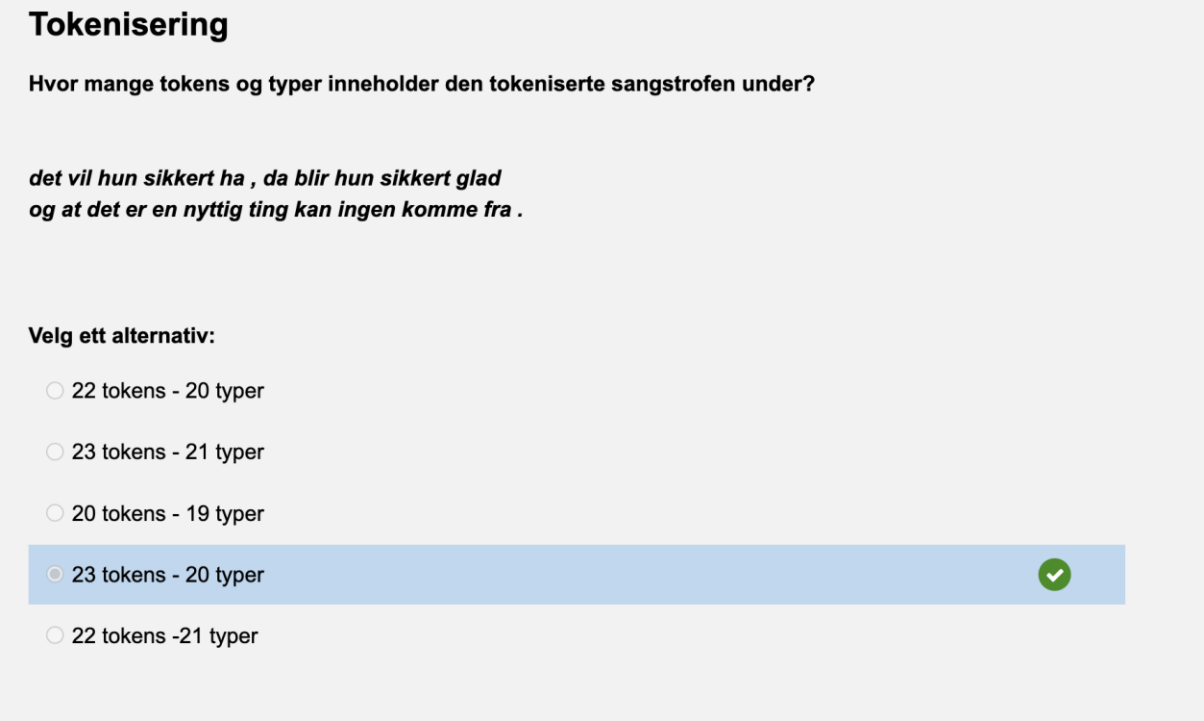
Svar:
Morfologi: beskriver hvordan ord er bygget opp av *morfemer*, de minste meningsbærende enhetene i språk samt hvordan nye ord dannes.

Syntaks: beskriver regler for hvordan setninger er bygget opp ord og grupperinger av ord, såkalte fraser.

Semantikk: omhandler ord og setningers *betydning.* Underområder er leksikalsk semantikk som beskriver ords betydning samt setningssemantikk.

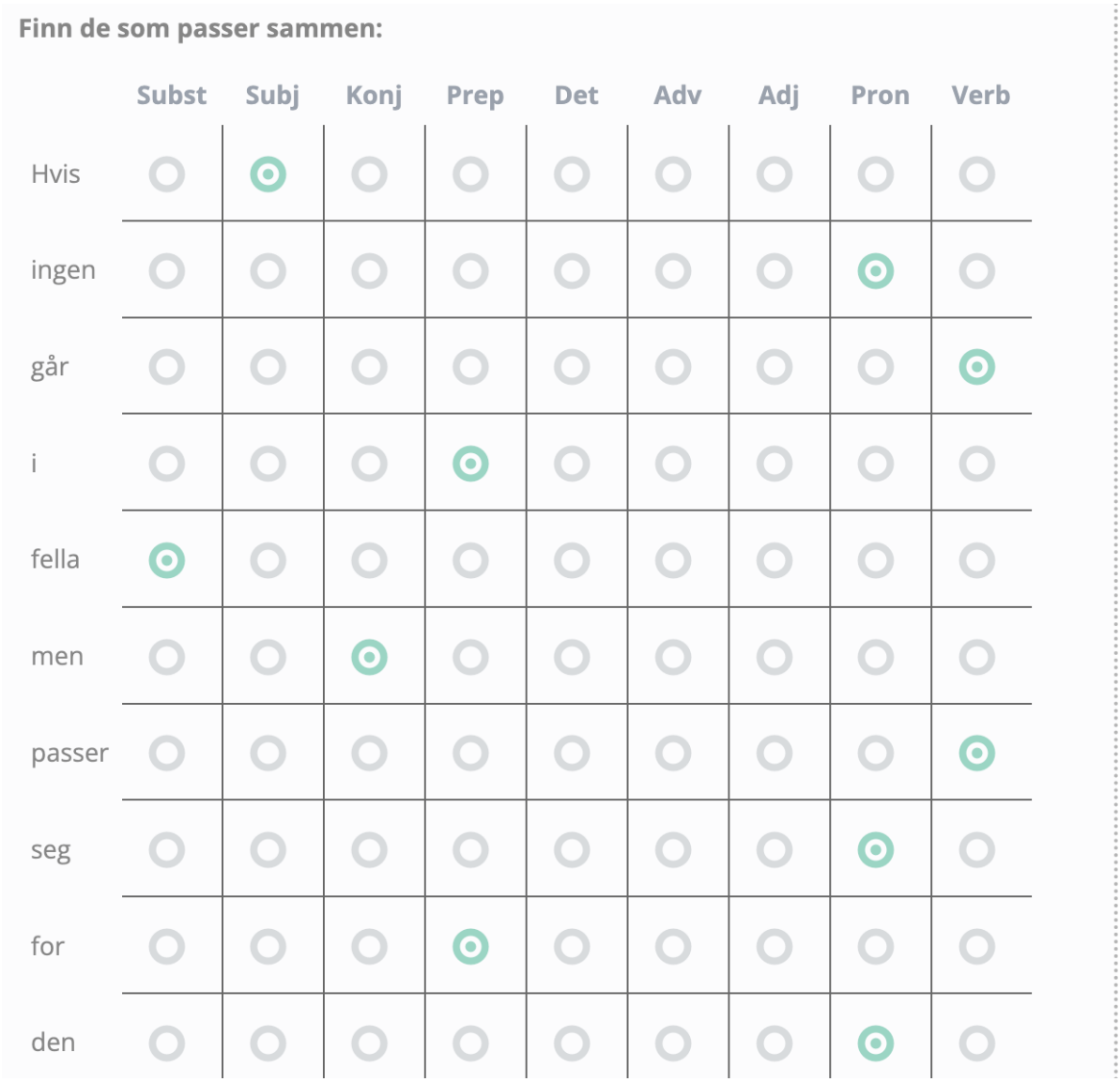
Dessuten Pragmatikk, fonologi, fonetikk.

6 Tokenisering (4 poeng)



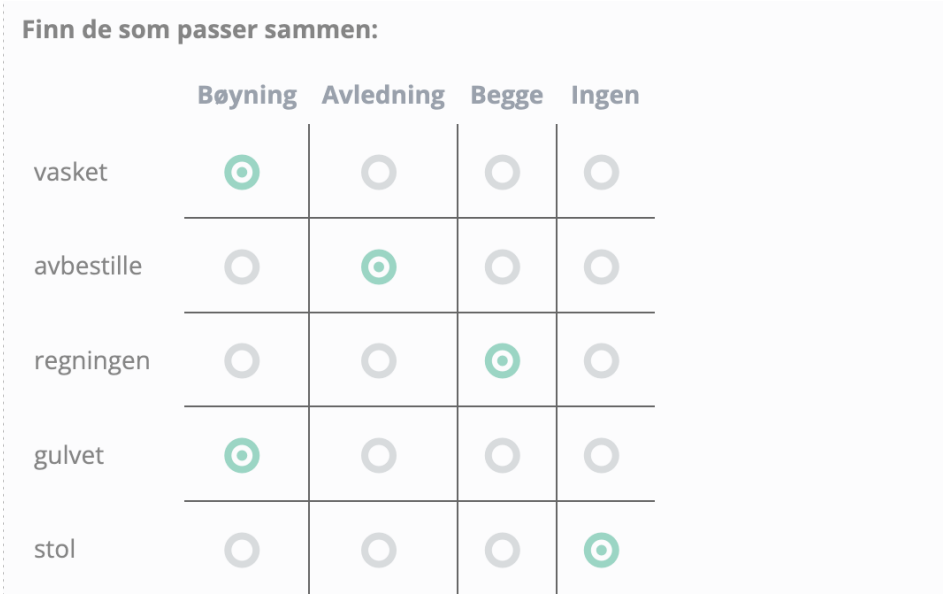
7 Ordklasser

Tildel ordklasse til hver av ordene i setningen "*Hvis ingen går i fella men passer seg for den*".



8 Orddannelse

For hver av ordene under: er det dannet som resultat av bøyning, avledning, begge eller ingen av delene?



9 Et problem som må løses

Du er ansatt i bedriften NorGram AS og har fått i oppgave å lage en bigram-modell for det følgende korpuset:

<s> Mina liker iskrem </s>
<s> Sara liker ikke kake </s>
<s> Tor liker iskrem og kake>

Sjefen er kun interessert i å vite hva sannsynligheten er for følgende setning:

<s> Tor liker ikke paprika </s>

Hvilket problem vil du møte på her, og hvilken teknikk fra faget kan du bruke for å løse det?

Svar:

Paprika har ingen frekvens i korpuset. Vi vil derfor få en sannsynlighet på 0 for et bigram som inneholder paprika, noe som igjen gir en total sannsynlighet på 0 for hele setningen. Dette kan løses ved hjelp av glatting, f.eks laplace-glatting.

10

N-gram

I dette faget har vi for det meste jobbet med *n*-gram modeller der *n* har vært 2, 3 eller 4. Hva er hovedproblemet med å øke *n* for mye?
Velg ett alternativ:

- Vi vil ende opp med å glatte/smoothe for mye.
- Produktet av sannsynlighetene vil nærme seg null, og da vil datamaskinen få problemer med å representere det (numerisk underflow).
- Jo større n er, jo vanskligere blir det å få et godt estimat for n-grammet. ✓
- Jo lavere n er, jo bedre er modellen, fordi den da fanger opp små, lokale kontekstforskjeller mellom de ulike n-grammene.

11 Bigram

Gitt følgende korpus:

<s> syre i vann det går an </s>
<s> vann i syre det blir uhyre </s>
<s> blir syre i vann uhyre </s>

Får vi sannsynlighetene:

P(syre | <s>) = 1/3
P(i | syre) = 2/3
P(vann | i) = 2/3
P(det | vann) = 1/3
P(går | det) = 1/2
P(an | går) = 1/1
P(</s> | an) = 1/1

P(vann | <s>) = 1/3
P(i | vann) = 1/3
P(syre | i) = 1/3
P(det | syre) = 1/3
P(blir | det) = 1/2
P(uhyre | blir) = 1/2
P(</s> | uhyre) = 2/2
P(blir | <s>) = 1/3
P(syre | blir) = 1/2

P(uhyre | vann) = 1/3

Bigramsannsynligheten for setningen:

<s> vann i syre det går an </s>

Blir:

P(vann | <s>) * P(i | vann) * P(syre | i) * P(det | syre) * P(går | det) * P(an | går) * P(</s> | an)

= 1/3 * 1/3 * 1/3 * 1/3 * 1/2 * 1 * 1 = **0.00617**

12

Term i nevneren

I dette faget har vi brukt formelen nedenfor når vi har jobbet med språkmodeller. Hva kan skje med sannsynlighetsfunksjonen P om vi ikke legger til termen $|V|$ i nevneren? Og hvorfor er dette et problem? Svaret ditt trenger ikke være lengre enn 2-3 setninger.

$$P(w_n|w_{n-1}) = \frac{C(w_{n-1},w_n)+1}{C(w_{n-1})+|V|}$$

C er en opptellingsfunksjon som teller forekomsten av argumentet sitt i et korpus.

Formelen viser add-one smoothing. Om vi ikke legger til V risikerer vi å få verdier av P som er større enn 1, altså at enkelte bigrammer har mer enn 100% sannsynlighet for å forekomme, noe som konseptuelt ikke gir mening og som vil skape en skjevhet i fordelingen.

13 Grammatisk?

MERK: På grunn av mangler i en oppgitte grammatikken, ble de tre deloppgavene preposisjon ikke tellende.

Ta utgangspunkt i følgende grammatikk:

S -> NP VP

NP -> Det N

NP -> N

NP -> NP PP

VP -> V

PP -> P NP

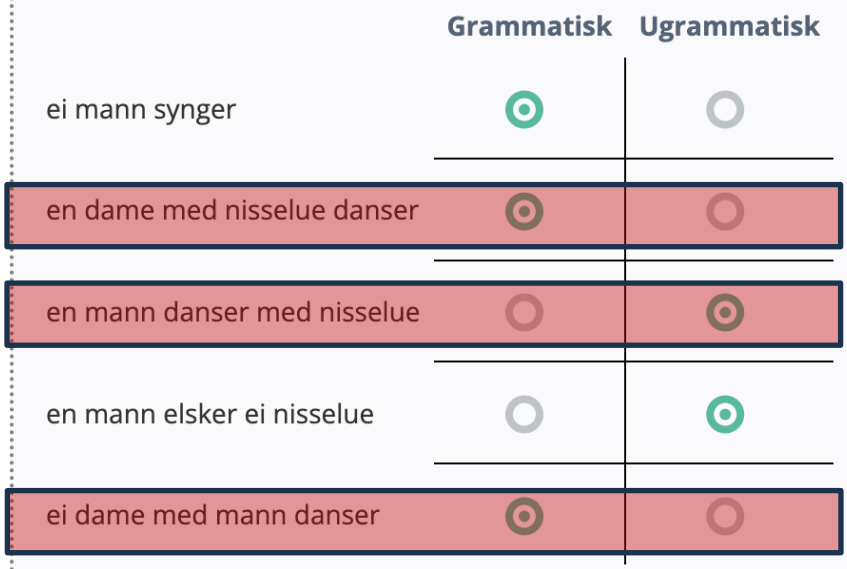
V -> danser|synger|aker|elsker

N -> mann|dame|nisselue

Det -> en|ei

P -> med

Er følgende setninger grammatiske eller ugrammatiske gitt grammatikken?



14 Konstituenter (6 poeng)

Ta utgangspunkt i følgende setning:

et lystig reinsdyr danser over himmelen

Vis hvordan to ulike konstituenttester kan avgjøre konstituentstatus for “et lystig reinsdyr”. Her skal du:

1. kort beskrive konstituenttesten
2. vise hvordan den kan anvendes på setningen over for sekvensen *et lystig reinsdyr*

Svar:
Her bør svaret komme innom to av følgende:

- stå-alene testen
    - sekvensen kan stå alene som svar på et spørsmål
    - eksempel på anvendelse: *Hvem danser over himmelen? Et lystig reinsdyr*
- erstatte med pronomen
    - sekvensen kan erstattes med et pronomen
    - eksempel: *Den danser over himmelen*
- flytte som enhet
    - sekvensen kan flyttes som en enhet og fortsatt gi en grammatisk setning
    - eksempel: *et lystig reinsdyr er det som danser over himmelen* eller *det er et lystig reinsdyr som danser over himmelen*

15 Utvid grammatikken (4 poeng)

Hvordan kan vi utvide grammatikken fra forrige oppgave (gjentatt under) for at den skal gi en korrekt syntaktisk analyse av setningen:

et reinsdyr danser med en nisselue

Husk å angi både leksikalske og ikke-leksikalske regler som må legges til.

S -> NP VP

NP -> Det N

NP -> NP PP

VP -> V

V -> danser|synger|aker|elsker

N -> mann|dame|nisselue

Det -> en|ei

P -> med

Svar:

Følgende regler må legges til:

Det -> et
N -> reinsdyr

VP -> VP PP eller VP -> V PP

PP -> P NP (Manglet i opprinnelig oppgave. Kan legges til her eller tas for gitt.)

Semantikk

16 Leksikalske relasjoner

I denne oppgaven skal du fra et sett med leksikale relasjoner velge den som passer best for noen korte utsagn. For hvert utsagn er det markert to ord med fet skrift, og det er relasjonen som best beskriver forholdet mellom disse du skal klassifisere. Utsagnene er nummererte, og indeksene (tallet foran utsagnet) tilsvarer tallet i første kolonne i tabellen under.

1) hun løftet **gryta** opp, holdt den i det ene **øret** mens hun stakk en sleiv nedi

2) Myrna Vilhamer, som var **gift** og hadde to barn, ble i går kveld funnet drept. I blodet hennes er det funnet spor av store mengder **gift**.

3) **leggbenet** henger fast i **lårbenet**, og lårbenet henger fast i hoftebenet

4) Hun har reist **fram** og **tilbake** mellom butikkene og leiligheten i hele dag

5) Hilde satte en **friske** blomster på bordet, nå som hun var **frisk** av forkjølelsen.

6) Jeg **går** til en **gård**

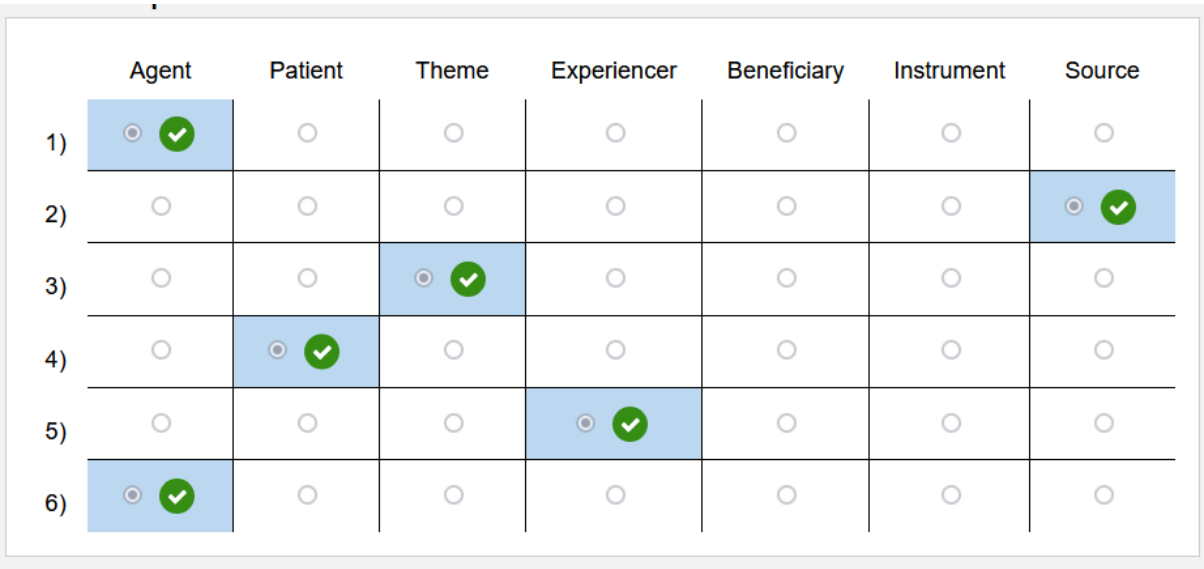
Finn de som passer sammen:

	Homonymi	Meronymi	Hyponymi	Antonymi	Polysemi	Taksonomiske søstre
1)	○	◉ ✓	○	○	○	○
2)	◉ ✓	○	○	○	○	○
3)	○	○	○	○	○	◉ ✓
4)	○	○	○	◉ ✓	○	○
5)	○	○	○	○	◉ ✓	○
6)	◉ ✓	○	○	○	○	○

17 Semantiske roller

I denne oppgaven skal du fra et sett med semantiske roller velge den som passer best for ordet som er markert med fet skrift.

1) **Reven** hoppet ut av grøfta
2) Reven hoppet ut av **grøfta**
3) **Reven** er i buskene
4) Jeg må kutte de **buskene**
5) **Espen** så på mens moren bakte kake
6) Espen så på mens **moren** bakte kake til farmor
Finn de som passer sammen:



18 Trekkvektorer

Vi skal lage trekkvektorer for to ord i en setning. Teksten er ferdig tokenisert, hvert token er adskilt med mellomrom.
Vektorene skal ha følgende trekk: "w-1": Token (ordet) til venstre. "w": Token, "w+1": Token til høyre, "pos-1": Ordklasse til venstre for token, "pos": ordklasse til valgt token, "pos+1": ordklasse for token til høyre.
Hvis valgte token er først i setningen, brukes vi "<s>" for token til venstre. Tilsvarende bruker vi "</s>" for token til høyre for siste token. Disse token får ordklasse "None".
Alle trekk angis med heltall utfra de gitte tabeller.
Setningen vi skal bruke er:

`Nå eller aldri , sa katta .`

Vi har brukt en modell som ga oss følgende ordklasser pr token:

nå	ADV
eller	CCONJ
aldri	ADV
,	PUNCT
sa	VERB
katta	NOUN
.	PUNCT

Fyll ut tabellen med heltall, ved hjelp av utdraget av vokabularet og ordklasse-indeks nederst

Token	w-1	w	w+1	pos-1	pos	pos+1
nå	0 ✓	8 ✓	5 ✓	99 ✓	4 ✓	3 ✓
katta	12 ✓	7 ✓	3 ✓	0 ✓	2 ✓	1 ✓

Token

Indeks	Token
0	<s>
1	</s>
2	,
3	.
4	aldri
5	eller
6	isen
7	katta
8	nå
9	over
10	rasker
11	reven
12	sa

Ordklasser

Indeks	Ordklasse
0	VERB
1	PUNCT
2	NOUN
3	CCONJ
4	ADV
99	None

19 Naive Bayes

Klassifiseringsalgoritmen Naive Bayes gjør en såkalt *naiv* antagelse, derav navnet. Hva innebærer denne antagelsen, og hvordan uttrykkes den i formelen for Naive Bayes?

Den naive antagelsen i NB innebærer at vi behandler alle trekk som uavhengige, gitt en klasse. Konsekvensen av dette er typisk at rekkefølgen på ord ikke har noen betydning innad i en klasse.

Dette uttrykkes i formelen ved at sannsynligheten for en serie med trekk $f_1, \ldots, f_n$ for en klasse c kan tilnærmes som:

$$P(f_1, \ldots, f_n) \approx \Pi_{i=1}^{n} P(f_i|c)$$

20 – 24 IE

Vi skal trekke ut informasjon fra en tekst ved hjelp av forskjellige metoder for informasjonsekstrahering. Teksten består av disse to setningene:

Setning 1:*"John Doe jobbet ved William Carey University fra april 2020 til april 2023."*
Setning 2: *"Han er to meter høy og veldig snill."*

Angi hvilken metode fra språkteknologi vi bruker i hvert tilfelle:

IE1

"Han" i setning 2 peker til "John Doe" i setning 1.

Velg ett alternativ:

- Sentimentanalyse
- Named Entity Resolution
- Relasjonsekstraksjon
- Temporal Expression Extraction
- Coreference Resolution ✓
- Event Extraction

IE2

Setning 2 er en positiv setning.

Velg ett alternativ:

- Named Entity Resolution
- Temporal Expression Extraction
- Relasjonsekstraksjon
- Coreference Resolution
- Event Extraction
- Sentimentanalyse ✓

IE3

Setning 1 inneholder en tidsangivelse med varighet 3 år.

Velg ett alternativ:

- Sentimentanalyse
- Event Extraction
- Coreference Resolution
- Named Entity Resolution
- Temporal Expression Extraction ✓
- Relasjonsekstraksjon

IE4

Ansatt_hos("John Doe", "William Carey University")

Velg ett alternativ:

- Named Entity Resolution
- Sentimentanalyse
- Relasjonsekstraksjon ✓
- Event Extraction
- Coreference Resolution
- Temporal Expression Extraction

IE5

Setning 1 nevner organisasjonen "William Carey University" og personen "John Doe".

Velg ett alternativ:

- Event Extraction
- Coreference Resolution
- Temporal Expression Extraction
- Named Entity Resolution 
- Sentimentanalyse
- Relasjonsekstraksjon