Løsningsforslag til IN1140-eksamen H25

Løsninger til oppgaver med fritekstsvar vises først, så følger eksport fra Inspera under som viser fasit for flervalgsspørsmålene.

Oppgave 3.2 Sannsynligheter for sekvenser

Det er mulig å få full uttelling selv om man har et noe mindre utfyllende svar enn forslaget som gis her.

- Markov-antakelsen.
- Vi antar at den felles sannsynligheten for hele sekvensen kan tilnærmes som produktet av sannsynligheten for hvert ord betinget på kun de n-1 foregående ordene (i stedet for hele historikken). Vi tenker altså at hvert ord er betinget uavhengig av alle de tidligere ordene bortsett fra de n-1 siste.
- Dette gir oss færre parametere, gjør det mye enklere å beregne sannsynlighetene, og gjør det mer overkommelig å samle nok data til å få pålitelig statistikk.

Oppgave 4.2 Konstituenter

Her bør svaret komme innom to av følgende:

- **stå-alene** testen: sekvensen kan stå alene som svar på et spørsmål
    - eksempel på anvendelse: Hva skrev en foreleser? Et opprørt leserinnlegg
- **erstatte med pronomen**: sekvensen kan erstattes med et pronomen
    - eksempel: en foreleser skrev det
- **flytte som enhet**: sekvensen kan flyttes som en enhet og fortsatt gi en grammatisk setning
    - eksempel: et opprørt leserinnlegg skrev en fore

Oppgave 4.2 Utvid grammatikken

Følgende utvidelser / regler må legges til grammatikken:

Det -> ei
N -> dame|penn
VP -> V NP

NB! Her godkjennes **ikke** VP -> VP NP PP da denne ikke er rekursiv for VP.
VP -> VP PP godkjennes også, men må da komme **i tillegg** til VP-regelen over.

Oppgave 6.4 Sentimentanalyse

Det er mulig å få full uttelling selv om man har et noe mindre utfyllende svar enn forslaget som gis her og ikke er innom alle punktene, men det er viktig å ha nevnt noe om flertydighet og komposisjonalitet (selv om man kanskje bruker andre termer).

Ved å kun basere seg på oppslag i ordlister klarer man ikke ta høyde for alle de kontekstuelle faktorene som kan påvirke polariteten til uttrykk. Siden ords polaritet gjerne er avhengig av konteksten de står i, er det problematisk å basere seg på at de har en forhåndsdefinert polaritet og at de kan vurderes i isolasjon. Noen eksempler:

- Ignorerer syntaks og komposisjonalitet, som kan være avgjørende for et uttrykks polaritet, f.eks. ved negasjon ("ikke bra" vs "bra").
- Flertydighet ("sykt" kan være både positivt/negativt).
- Flertydighet er ofte også knyttet til domene, f.eks. "syk" i en medisinsk kontekst ("kvinnen er syk") vs. et diskusjonsforum ("filmen er syk").
- Leksikonoppslag ignorerer også intensitet og grad, slik at «grei» og «fantastisk» gjerne får samme score. Det samme gjelder forskjellen på "passe bra" og "ufattelig bra" – her ser vi hvordan dette også samspiller med komposisjonalitet.
- Andre kontektsavhengigheter (som igjen kan sees som en form for flertydighet) kan være knyttet til ironi og sarkasme, der positive ord i realiteten uttrykker noe negativt eller vice versa.
- Det finnes flere andre faktorer som kan nevnes og gi uttelling utover eksemplene over, som f.eks manglende dekning / begrenset vokabular, osv.

Forside

UNIVERSITETET I OSLO
Det matematisk-naturvitenskapelige fakultet
Skriftlig eksamen i IN1140
2025 HØST

Varighet: 17.12 kl 15:00 til 17.12 kl 19:00
Tillatte hjelpemidler: Ingen.

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. For enkelte av flervalgsoppgavene er det mulig å få minuspoeng ved feil svar.

Om du ønsker å zoome i oppgavesettet, hold ctrl nede og trykk + eller - på numerisk tastatur.

1.1 Regex 1

Vi har følgende regulære uttrykk:

```
/4{3}|56|3{2,4}/
```

Hvilke alternativer under inneholder tekststrenger der hele strengen vil matches av det regulære uttrykket?

- [ ] 4563
- [ ] 444 ✔
- [ ] 34
- [ ] 43
- [ ] 33 ✔
- [ ] 56 ✔

Maks poeng: 3

1.2 Regex 2

Vi har følgende regulære uttrykk:
`/J?[a-z]+/`

Hvilke alternativer under inneholder tekststrenger der hele strengen vil matches i én match av det regulære uttrykket?

- [x] julegaver
- [ ] Juleønsker
- [x] Julegaver
- [x] nisse
- [ ] Jaa!
- [ ] J

Maks poeng: 3

1.3 Regex 3

Vi har følgende regulære uttrykk:

`/f(a|e).*r/`

Hvilke alternativer under inneholder tekststrenger der hele strengen vil matches i én match av det regulære uttrykket?

- [x] feler ✔
- [x] far ✔
- [ ] fallera
- [x] faener ✔
- [ ] fanteri
- [ ] fiene

Maks poeng: 3

1.4 Regex 4

Vi skal lage et regulært uttrykk som skal gjenkjenne norske telefonnumre og har flere utkast å velge mellom.

Hvilket av disse regulære uttrykkene matcher hele strengen for alle disse norske telefonnumrene:

- +47 22 67 57 89
- 0047 22 67 57 89
- 22 67 57 88

og ingen av disse?

- +47 47222222
- 47222222
- 22 67 57

Velg ett alternativ:

- ((\+47|0047)\s)?([0-9][0-9]\s){2}([0-9][0-9]){1}
- /((\+47|0047)\s)?([0-9][0-9]\s){3}([0-9][0-9]){1}/ ✔
- /((\+47|0047)\s)+([0-9][0-9]\s){3}([0-9][0-9]){1}/

Maks poeng: 5

1.5 Regex 5

Hvilket spesialtegn matcher 0 eller flere tegn?

Velg ett alternativ:

- * ✔
- +
- ?
- $

Maks poeng: 1

1.6 Regex 6

Hvordan kan du matche strengen “Jul” på begynnelsen av en linje?

Velg ett alternativ:

- *Jul
- ^Jul ✔
- #Jul
- $Jul

Maks poeng: 2

1.7 Regex 7

Hvilken av følgende matcher et mellomrom?

Velg ett alternativ:

- $
- \d
- \s ✔
- \w

Maks poeng: 1

2.1 Tokenisering

Under ser du en frase fra et av diktene til poeten Gertrude Stein. Hvor mange tokens og typer inneholder den?

en rose er en rose er en rose er en rose

Velg ett alternativ:

- 3 tokens / 11 typer
- 11 tokens / 11 typer
- 11 tokens / 3 typer ✔
- 3 tokens / 3 typer

Maks poeng: 2

2.2 Ordklasser

Tildel ordklasse til hver av ordene i setningen "På låven sitter nissen med sin julegrøt så god og søt***".***

Kolonneforklaring:

Subst=Substantiv
Subj=Subjunksjon
Konj=Konjunksjon
Prep=Preposisjon
Det=Determinativ
Adv=Adverb
Adj=Adjektiv
Pron=Pronomen
Verb=Verb

Finn de som passer sammen:

	Subst	Subj	Konj	Prep	Det	Adv	Adj	Pron	Verb
På	○	○	○	✔	○	○	○	○	○
låven	✔	○	○	○	○	○	○	○	○
sitter	○	○	○	○	○	○	○	○	✔
nissen	✔	○	○	○	○	○	○	○	○
med	○	○	○	✔	○	○	○	○	○
sin	○	○	○	○	✔	○	○	○	○
julegrøt	✔	○	○	○	○	○	○	○	○
så	○	○	○	○	○	✔	○	○	○
god	○	○	○	○	○	○	✔	○	○
og	○	○	✔	○	○	○	○	○	○
søt	○	○	○	○	○	○	✔	○	○

Maks poeng: 6

2.3 Orddannelse

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

Finn de som passer sammen:

	Bøyning	Begge	Avledning	Ingen
venn	○	○	○	○ ✔
gulvet	○ ✔	○	○	○
uvennen	○	○ ✔	○	○
vennlig	○	○	○ ✔	○
vennen	○ ✔	○	○	○
tøffing	○	○	○ ✔	○

Maks poeng: 5

3.1 Trigrams

Hvor mange 3-gram er det i følgende sekvens?

en rose er en rose er en rose er en rose

Velg ett alternativ:

- 3
- 11
- 10
- 9 ✔
- 1
- 2

Maks poeng: 2

3.2 Sannsynligheter for sekvenser

Gitt en sekvens av ord $w_1^k = \langle w_1, w_2, \ldots, w_k \rangle$, hvilken navngitt teoretisk antagelse fra pensum er det som uttrykkes i følgende formel? Forklar kort hva det betyr i praksis og hvorfor gjør vi denne antakelsen? (Det bør holde med et par setninger totalt.)

$$P(w_1^k) \approx \prod_{i=1}^{k} P(w_i | w_{i-n+1}^{i-1})$$

(Du kan anta at sekvensen paddes med start- og stopp-tokens slik at formelen alltid er veldefinert for alle posisjoner i sekvensen.)

Skriv svaret ditt her. Endringer blir lagret automatisk.

Maks poeng: 6

3.3 MLE + glatting

Vi skal estimere sannsynlighetene for en 2-gram-modell $P(w_n|w_{n-1})$ over ord i vokabularet V ved hjelp av MLE (*maximum likelihood estimation*), men vi ønsker også å inkludere såkalt *add-one smoothing* eller Laplace-glatting. Anta at C er en funksjon som teller antall forekomster av en gitt sekvens i treningsdataene. Hva er riktig formel?

$$P(w_n|w_{n-1}) =$$

Velg ett alternativ:

- $\frac{C(w_{i-1},w_i)}{|V|+1}$
- $\frac{C(w_{i-1},w_i)+1}{C(w_{i-1})+|V|}$ ✔
- $\frac{C(w_{i-1},w_i)}{C(w_{i-1})}$
- $\frac{C(w_i)+1}{C(w_{i-1})+|V|}$

Maks poeng: 2.5

3.4 MLE

Sannsynlighetene i *n*-gram-modeller estimeres typisk ved hjelp av såkalt _maximum likelihood estimation_ (MLE). Hva er den vanligste praktiske utfordringen med dette?

Velg ett alternativ:

- Ukjente (usette) n-gram får for høy sannsynlighet.
- Ukjente (usette) n-gram får sannsynlighet 0. ✔
- Modellen kan ikke håndtere start/slutt på setninger.
- Modellens sannsynligheter summerer ikke til 1.
- Alle n-gram får lik sannsynlighet.

Maks poeng: 2.5

4.1 Grammatisk?

Ta utgangspunkt i følgende grammatikk:

```
S -> NP VP
NP -> Det N
NP -> NP PP
VP -> V
PP -> P NP
V -> leser|skriver|spør
N -> student|foreleser|lue|brev
Det -> en|et
P -> med
```

Er følgende setninger grammatiske eller ugrammatiske gitt grammatikken?

	Ugrammatisk	Grammatisk
en foreleser med en lue skriver et brev	✔	
en student spør en foreleser	✔	
en student med en lue leser		✔
en foreleser skriver med en lue	✔	
et foreleser skriver		✔

Maks poeng: 5

4.2 Konstituenter

Ta utgangspunkt i følgende setning:

en foreleser skrev et opprørt leserinnlegg.

Vis hvordan to ulike konstituenttester kan avgjøre konstituentstatus for "et opprørt leserinnlegg". Her skal du for hver konstituenttest:

1. kort beskrive konstituenttesten
2. vise minst ett eksempel på hvordan testen kan anvendes på setningen for sekvensen "et opprørt leserinnlegg" i setningen over.

Skriv ditt svar her

Maks poeng: 4

4.3 Utvid grammatikken

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

ei dame skriver et brev med en penn

NB! Her er det en forutsetning at den utvidede grammatikken også skal dekke følgende setninger uten tillegg av flere regler:

ei dame skriver et brev med en penn med en student
ei dame skriver et brev med en penn med en student med en lue

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

```
S -> NP VP
NP -> Det N
NP -> NP PP
VP -> V
PP -> P NP
V -> leser|skriver|spør
N -> student|foreleser|lue|brev
Det -> en|et
P -> med
```

Hvilke regler skal legges til grammatikken? Skriv ditt svar her

---

Maks poeng: 6

5.1 Leksikale 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 **stolen** opp, holdt den i det ene **stolbeinet** og svingte den over hodet.
2. Dommeren hadde hatt en lang dag i **retten** og ville kose seg med den beste **retten** hun visste om, nemlig pizza!
3. **leggbenet** henger fast i **lårbenet**, og lårbenet henger fast i hoftebenet
4. Han reiste **opp** og **ned** med heisen flere ganger
5. **Stjernen** som var kjent fra TV-skjermen kikket opp på den lysende **stjernen** på himmelen.
6. Jeg **går** til en **gård**.
7. **Tesla** er en populær **el-bil**.

Finn de som passer sammen:

	Homonymi	Meronymi	Hyponymi	Antonymi	Polysemi	Taksonomiske søstre
1.	○	✔	○	○	○	○
2.	✔	○	○	○	○	○
3.	○	○	○	○	○	✔
4.	○	○	○	✔	○	○
5.	○	○	○	○	✔	○
6.	✔	○	○	○	○	○
7.	○	○	✔	○	○	○

Maks poeng: 7

5.2 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 i hver setning. Setningene er nummererte, og indeksene (tallet foran utsagnet) tilsvarer tallet i første kolonne i tabellen under.

1. **Mannen** løp til butikken
2. Mannen skrelte og skar opp **eplene**
3. Han kastet **eplene** i søpla
4. Jeg må kutte de buskene med **hagesaks**
5. **Espen** så på mens moren bakte kake
6. Espen så på mens **moren** bakte kake til farmor
7. Espen så på mens moren bakte kake til **farmor**
8. **Farmor** syntes kaken smakte godt

Finn de som passer sammen:

	Agent	Patient	Theme	Experiencer	Beneficiary	Instrument	Source
1.	✓						
2.		✓					
3.			✓				
4.						✓	
5.				✓			
6.	✓						
7.					✓		
8.				✓			

Maks poeng: 8

6.1 Ekstrinsisk / intrinsisk evaluering

Nedenfor ser du noen ulike påstander om såkalt *ekstrinsisk* og *intrinsisk* evaluering, med ordklassetagging som eksempel. Kryss av for hvilke påstander som er korrekte (det kan være flere).

Velg ett eller flere alternativer

- [ ] Intrinsisk evaluering måler taggerens nytte internt i et større system, som f.eks. maskinoversettelse.
- [ ] Intrinsisk evaluering måler kvaliteten til selve treninsgsdataene mens ekstrinsisk evaluering måler ytelsen på en trent modell.
- [ ] Ekstrinsisk evaluering måler hvor nøyaktig modellen ekstraherer informasjon, i dette tilfellet ordklasser, sammenliknet med en gullstandard.
- [ ] Mens intrinsisk evaluering måler ytelse på utviklingssdata, bruker ekstrinsisk evaluering testdata (såkalt 'held-out testing').
- [ ] Intrinsisk evaluering måler kvaliteten på selve ordklassetaggingen direkte, sammenliknet med en gullstandard. ✔
- [ ] Ekstrinsisk evaluering måler hvor nyttig ordklassetaggingen er for en større oppgave, f.eks. et nedstrøms system som f.eks. maskinoversettelse. ✔

Maks poeng: 5

6.2 Naive Bayes 1

Vi antar her at vi jobbber med tekstklassifikasjon med algoritmen Naive Bayes, med en trekkrepresentasjon basert på såkalt BoW (bag-of-words). For et gitt dokument d, representert ved et sett av k trekk $w_1, w_2, \ldots, w_k$, og et sett av mulige klasser C, så er målet å identifisere den mest sannsynlige klassen $\hat{c} = \underset{c \in C}{argmax}\ P(c|d)$. Men snarere enn å beregne dette direkte, foretas det en del omskrivninger for komme frem til en formel som gir oss en forenklet måte å gjøre klassifikasjonen på. Under viser vi ulike trinn i omskrivningen, med hvert trinn nummerert. Kryss av i tabellen under for hvilken regel eller antakelse som ligger til grunn for hvert trinn.

1. $\hat{c} = \underset{c \in C}{argmax}\ P(c|d) =$
2. $= \underset{c \in C}{argmax}\ \frac{P(d|c)\ P(c)}{P(d)}$
3. $= \underset{c \in C}{argmax}\ P(c)\ P(d|c)$
4. $= \underset{c \in C}{argmax}\ P(c)\ P(w_1, w_2, \ldots, w_k|c)$
5. $= \underset{c \in C}{argmax}\ P(c) \prod_{i=1}^{k} P(w_i|w_1, \ldots, w_{i-1}, c)$
6. $\approx \underset{c \in C}{argmax}\ P(c) \prod_{i=1}^{k} P(w_i|c)$

Finn de som passer sammen:

	6	4	3	2	5
Produktregelen	○	○	○	○	○ ✔
Bayes regel	○	○	○	○ ✔	○
Ignorerer en konstant	○	○	○ ✔	○	○
Antar at d representeres ved BoW-trekk	○	○ ✔	○	○	○
Antar at trekkene er betinget uavhengige	○ ✔	○	○	○	○

Maks poeng: 8

6.3 Naiv Bayes 2

Nedenfor ser du noen påstander om tekstklassifikasjon med algoritmen Naive Bayes. Vi antar at vi bruker såkalte BoW-trekk (bag-of-words). Kryss av for påstandene som er korrekte (det kan være flere).

Velg ett eller flere alternativer

- [ ] Naive Bayes ignorerer sannsynligheten for den enkelte klasse, den såkalte prior sannsynligheten P(c).
- [ ] Naive Bayes kalles naiv fordi den antar at trekkene er gjensidig avhengige av hverandre.
- [ ] Såkalt glatting (f.eks. add-one smoothing) kan være nødvendig for å unngå proble[illegible]r med nullsannsynligheter. ✔
- [ ] Naiv Bayes ignorerer sannsynligheten for det gitte dokumentet, den såkalte margi[illegible]'e sannsynligheten P(d). ✔
- [ ] Naive Bayes antar at trekkene er betinget uavhengige av hverandre gitt klassen. ✔
- [ ] Algoritmen kalles "naiv" fordi den antar at alle klasser forekommer like ofte i treningsdataene.
- [ ] For tekstklassifisering estimeres gjerne de betingete sannsynlighetene for hvert tre[illegible] gitt klassen ved hjelp av de relative frekvensene for trekkene i hver klasse. ✔

Maks poeng: 7

6.4 Sentimentanalyse

En utbredt anvendelse av språkteknologi er såkalt *sentimentanalyse*, som enkelt sagt handler om å klassisfisere tekster som positive eller negative. Som et alternativ til maskinlæring brukes enkelte ganger såkalte *sentimentleksikon* for dette. Disse består av lister av ord som på forhånd er kategorisert som positive eller negative. Klassifikasjon av tekster som positive/negative gjøres så på basis av å telle antall treff for ord fra disse listene.

Diskuter kort (noen få setninger) hvilke svakheter som er forbundet med en naiv bruk av en slik leksikontilnærming til sentimentanalyse. Gi eksempler på relevante språklige fenomener.

Skriv svaret ditt her. Endringer blir lagret automatisk.

Maks poeng: 6

7 Kommentarer og uklarheter

Hvis du ønsker å redegjøre for forutsetninger du la til grunn ved svar på flervalgsoppgaver, for eksempel fordi du hadde problemer med å forstå noen av oppgavene, kan du beskrive dette i feltet under. NB! Dette er ikke nødvendig hvis du ikke hadde noen slike utfordringer. Denne oppgaven gir ikke poeng.

Tips: Skriv et punktum hvis du ikke har noen kommentarer, så registreres oppgaven som besvart.

Skriv eventuelle kommentarer her.

Maks poeng: 0