IN1140

Høst 2022 eksamen

Regulære Uttrykk

1.1 Norske datoformater

Korrekte norske datoangivelser som ikke angir måned med bokstaver, er i følge Språkrådet de følgende: (Alle i rekkefølgen dag-måned-år)

- 5.6.2014
- 05.06.2014
- 5.6.14
- 05.06.14

Hvilket eller hvilke regulære uttrykk fra listen under gjenkjenner disse datoformater gitt følgende krav:

- Dag skal angis med et tall fra og med 1 til og med 31. Ensifrede tall kan ha 0 foran.
- Måned skal være et tall fra og med 1 til og med 12. Ensifrede tall kan ha 0 foran.
- År skal representeres med et hvilket som helst to-eller firesifret tall.
- Som eksemplene viser, brukes punktum-tegnet for å skille dag, måned og år.

Hvilket eller hvilke regulære uttrykk fra listen under gjenkjenner disse datoformater gitt følgende krav:

- Dag skal angis med et tall fra og med 1 til og med 31. Ensifrede tall kan ha 0 foran.
- Måned skal være et tall fra og med 1 til og med 12. Ensifrede tall kan ha 0 foran.
- År skal representeres med et hvilket som helst to-eller firesifret tall.
- Som eksemplene viser, brukes punktum-tegnet for å skille dag, måned og år.

❏ /([012]?[1-9]|[123]0|31)\.(0?[1-9]|1[012])\.[0-9][0-9][0-9]?[0-9]?/
❏ /([012]?[1-9]|[123]0|31).(0?[1-9]|1[012]).[0-9][0-9][0-9]?[0-9]?/
❏ /([012]?[1-9]|[123]0|31).(0?[1-9]|1[012]).\d\d[\d\d]?/
❏ /([012]?[1-9]|[123]0|31)\.(0?[1-9]|1[012])\.\d\d[\d\d]?/
❏ /([012]?[1-9]|[123]0|31).(0?[1-9]|1[012]).\d\d(\d\d)?/
❏ /([012]?[1-9]|[123]0|31)\.(0?[1-9]|1[012])\.\d\d(\d\d)?/

1.1 Norske datoformater

Hvilket eller hvilke regulære uttrykk fra listen under gjenkjenner disse datoformater gitt følgende krav:

- ~~/([012]?[1-9]|[123]0|31)\.(0?[1-9]|1[012])\.[0-9][0-9][0-9]?[0-9]?/~~
- ~~/([012]?[1-9]|[123]0|31).(0?[1-9]|1[012]).[0-9][0-9][0-9]?[0-9]?/~~
- ~~/([012]?[1-9]|[123]0|31).(0?[1-9]|1[012]).\d\d[\d\d]?/~~
- ~~/([012]?[1-9]|[123]0|31)\.(0?[1-9]|1[012])\.\d\d[\d\d]?/~~
- ~~/([012]?[1-9]|[123]0|31).(0?[1-9]|1[012]).\d\d(\d\d)?/~~
- ➔ /([012]?[1-9]|[123]0|31)\.(0?[1-9]|1[012])\.\d\d(\d\d)?/

(1) Dag skal angis med **et tall fra og med 1 til og med 31**. Ensifrede tall *kan ha 0 foran.*

(2) Måned skal være **et tall fra og med 1 til og med 12**. Ensifrede *tall kan ha 0 foran.*

(3) År skal representeres med **et hvilket som helst to-eller firesifret tall**.

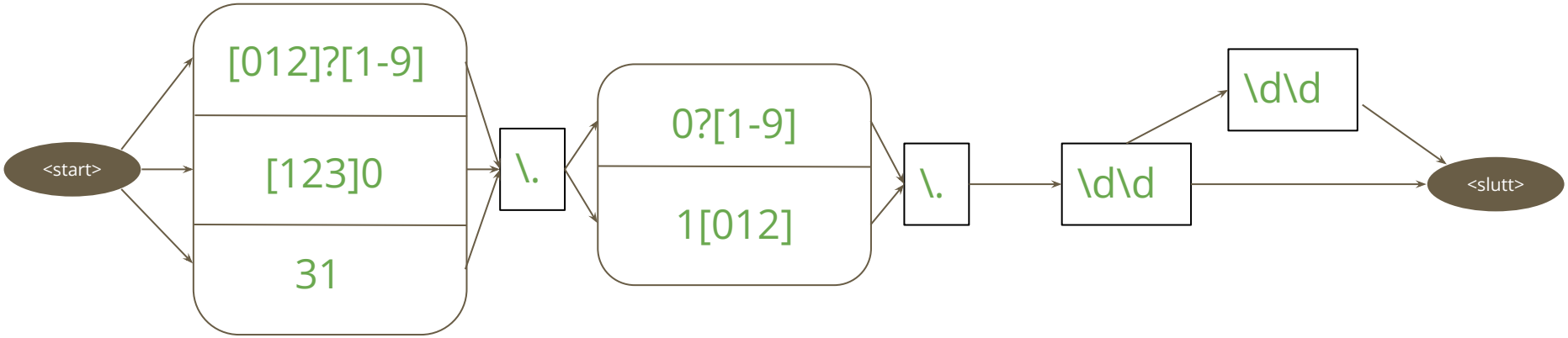


Som eksemplene viser, brukes **punktum**-tegnet for å skille dag, måned og år.

(1) Dag skal angis med **et tall fra og med 1 til og med 31**. Ensifrede tall *kan ha 0 foran.*

(2) Måned skal være **et tall fra og med 1 til og med 12**. Ensifrede *tall kan ha 0 foran.*

(3) År skal representeres med **et hvilket som helst to-eller firesifret tall**.



➔ /([012]?[1-9]|[123]0|31)\.(0?[1-9]|1[012])\.\d\d(\d\d)?/

1.2 Grupperinger

Under gis et regulært uttrykk som gjenkjenner strenger for måned-dag med bindestrek mellom: Måned f.o.m. 1 t.o.m. 12 og dato f.o.m. 1 t.o.m. 31. Dermed vil f.eks. tjuefjerde desember gjenkjennes hvis angitt som "12-24", og første januar gjenkjennes hvis angitt som "1-1", "01-1", "1-01", eller "01-01". I uttrykket er det to bueparenteser ( ):

/(0?[1-9]|1[012])-(31|[123]0|[012]?[1-9])/

Kan vi fjerne én eller begge av de to parentesene i det regulære uttrykket uten at vi endrer det som gjenkjennes?

- ❏ Begge parentesene kan fjernes uten at uttrykket endres.
- ❏ Den siste parentesen kan fjernes men ikke den første.
- ❏ Den første parentesen kan fjernes men ikke den andre.
- ❏ Nei, begge parentesene trengs.

1.2 Grupperinger

Under gis et regulært uttrykk som gjenkjenner strenger for måned-dag med bindestrek mellom: Måned f.o.m. 1 t.o.m. 12 og dato f.o.m. 1 t.o.m. 31. Dermed vil f.eks. tjuefjerde desember gjenkjennes hvis angitt som "12-24", og første januar gjenkjennes hvis angitt som "1-1", "01-1", "1-01", eller "01-01". I uttrykket er det to bueparenteser ( ):

/(0?[1-9]|1[012])-(31|[123]0|[012]?[1-9])/

Kan vi fjerne én eller begge av de to parentesene i det regulære uttrykket uten at vi endrer det som gjenkjennes?

- ~~Begge parentesene kan fjernes uten at uttrykket endres.~~
- ~~Den siste parentesen kan fjernes men ikke den første.~~
- ~~Den første parentesen kan fjernes men ikke den andre.~~
- ➔ Nei, begge parentesene trengs.

1.3 Disjunksjon

Gitt følgende regulære uttrykk:

/ab|cd|efg/

Hvilket eller hvilke tekststrenger blir gjenkjent av uttrykket?

- ❏ abcdefg
- ❏ ab
- ❏ abdfg
- ❏ ace
- ❏ cd
- ❏ acd

1.3 Disjunksjon

Gitt følgende regulære uttrykk:

/ab|cd|efg/

Hvilket eller hvilke tekststrenger blir gjenkjent av uttrykket?

- ~~abcdefg~~
- ➔ ab
- ~~abdfg~~
- ~~ace~~
- ➔ cd
- ~~acd~~

Ordklasser

2.1 Ordklassetagging

Denne oppgaven omhandler ordklassetagging. Tildel korrekt ordklasse i tabellen til ordene markert med **fet** skrift *slik de blir benyttet i teksten nedenfor*.

Forkortelsene for ordklasser i tabellen tilsvarer:

- Pron - Pronomen
- Subs - Subjunksjon
- Prep - Preposisjon
- Det - Determinativ (Artikkel)
- Adj - Adjektiv
- Adv - Adverb
- Konj - Konjunksjon
- Vrb - Verb

Teksten

Pron Subj

"**Hva** er klokken? **Om** jeg kommer for sent til timen, er det ikke sikkert jeg rekker presentasjonen min", sa Mats og **slukte** frokosten sin. Konj

Vrb

"Vil du at jeg skal kjøre deg, **eller** tar du sykkelen?", svarte faren.

Adv

"Den rustne skraphaugen der? Nei, du er nødt til å kjøre meg, og du må kjøre **fort** om vi skal rekke det!", sa Mats.

Adv

"Greit, om vi drar nå rekker vi det **kanskje**" svarte faren oppgitt. Klokken var fem **på** ni, de hadde dårlig tid, men skolen lå heldigvis ikke langt unna, og det var ikke en bil å se på **veien**!

Prep

Subs

- Pron - Pronomen
- Subj - Subjunksjon
- Prep - Preposisjon
- Det - Determinativ
- Adj - Adjektiv
- Adv - Adverb
- Konj - Konjunksjon
- Vrb - Verb

2.2 Orddannelse

I denne oppgaven skal vi jobbe med morfologi.

	ingen affikser	avlednings-prefiks	bøynings-prefiks	avlednings-suffiks	bøynings-suffiks	begge affikstyper
tjene	✓				✓	
tjeneste				✓		
betjener						✓
fortjene		✓				✓
tjenestene						✓
tjente					✓	

Hvis vi tolker infinitivs-e som bøyning

Språkmodeller

3.1 Trigram

Hvor mange trigram forekommer i teksten under?

<s> Det var en mørk og våt dag i november </s>
<s> Studentene ønsket å bestå eksamen i språkteknologi før ferien </s>

- ❏ 14
- ❏ 18
- ❏ 12
- ❏ 16

3.1 Trigram

NB: </s> og <s> skal ikke forekomme i det samme n-grammet

Hvor mange trigram forekommer i teksten under?

<s> Det var en mørk og våt dag i november </s>
<s> Studentene ønsket å bestå eksamen i språkteknologi før ferien </s>

- ~~14~~
- ➔ 18
- ~~12~~
- ~~16~~

NB: inkluderer setningsstart / -slutt markører

1. ('<s>', 'Det', 'var')
2. ('Det', 'var', 'en')
3. ('var', 'en', 'mørk')
4. ('en', 'mørk', 'og')
5. ('mørk', 'og', 'våt')
6. ('og', 'våt', 'dag')
7. ('våt', 'dag', 'i')
8. ('dag', 'i', 'november')
9. ('i', 'november', '</s>')
10. ('<s>', 'Studentene', 'ønsket')
11. ('Studentene', 'ønsket', 'å')
12. ('ønsket', 'å', 'bestå')
13. ('å', 'bestå', 'eksamen')
14. ('bestå', 'eksamen', 'i')
15. ('eksamen', 'i', 'språkteknologi')
16. ('i', 'språkteknologi', 'før')
17. ('språkteknologi', 'før', 'ferien')
18. ('før', 'ferien', '</s>')

3.2 Markovantagelsen

Når vi jobber med n-grammodeller legger vi til grunn den såkalte Markovantagelsen. **Hva er hovedgrunnen til at vi gjør dette?**

- For å utjevne sannsynlighetsmassen slik at høyfrekvente ord får mindre betydning.
- For å kunne gjøre en statistisk tilnærming av betingende sannsynligheter uten å måtte se for langt bakover i historikken.
- For å håndtere ord som ikke forekommer i treningsdataen (OOV-words)

3.2 Markovantagelsen

Eksempel for $n = 2$:

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

Når vi jobber med n-grammodeller legger vi til grunn den såkalte Markovantagelsen. **Hva er hovedgrunnen til at vi gjør dette?**

- ~~For å utjevne sannsynlighetsmassen slik at høyfrekvente ord får mindre betydning.~~
- ➔ For å kunne gjøre en statistisk tilnærming av betingende sannsynligheter uten å måtte se for langt bakover i historikken.
- ~~For å håndtere ord som ikke forekommer i treningsdataen (OOV-words)~~

3.3 Bigram

Du er ansatt i et språkteknologifirma som har fått i oppgave å prosessere bestillinger til et større restaurantkonsern. Du har funnet ut at du trenger en språkmodell som kan hjelpe deg med å regne ut hvor sannsynlige visse bestillinger er, slik at du i saker hvor det er vanskelig å oppfatte kundens ønsker kan velge den mest sannsynlige bestillingen. Det finnes et korpus i bedriften som inneholder bigram-sannsynlighetene for en del kombinasjoner av vanlige ord, og du bestemmer deg for å bruke det til å konstruere en bigram-modell.

Og noen bigram-sannsynligheter som en kollega har regnet ut for hånd:

P(Jeg | <s>) = 0.422
P(pizza | to) = 0.001
P(</s> | pizza) = 0.232
P(</s> | pasta) = 0.001

***NB:** Tabellen kan leses som P(kolonnetittel | radtittel)*
*f.eks. = **P(bestille | pizza) = 0.009 = sannsynligheten for "pizza bestille".***

	jeg	ønsker	å	bestille	pasta	betale	en	pizza
jeg	0.002	0.333	0	0	0	0.001	0	0
ønsker	0.111	0	0.4	0.218	0.4	0.271	0.2	0.211
å	0.002	0	0	0.432	0	0.400	0	0
bestille	0	0.1	*0.022**	0	0.422	0	0.288	0.3
pasta	0	0.002	*0.032**	0	0	0.174	0	0
betale	0	0.003	0.032	0	0.299	0	0.367	0.375
en	0.003	0	0.23	0.2	0.211	0	0.001	0.401
pizza	0	0.101	0.005	0.009	0	0	*0.01**	0

** Skrivefeil i oppgaveteksten som ikke påvirker det vi trenger til oppgaven.*

Gitt dataene over, hva er bigram-sannsynligheten for den følgende setningen?

<s> Jeg ønsker å bestille en pizza </s>

Oppgi de betingede sannsynlighetene du trenger og **sett opp regnestykket** (med faktiske verdier). *Du trenger ikke å regne ut den endelige sannsynligheten*, kun sette opp stykket som ville gitt svaret om man hadde brukt kalkulator.

3.3 Bigram

Hva er bigram-sannsynligheten for den følgende setningen?

<s> Jeg ønsker å bestille en pizza </s>

Oppgi de betingede sannsynlighetene du trenger og **sett opp regnestykket** (med faktiske verdier).

***NB:** Tabellen kan leses som P(kolonnetittel | radtittel)*
*f.eks. = **P(bestille | pizza) = 0.009 = sannsynligheten for "pizza bestille".***

	jeg	ønsker	å	bestille	pasta	betale	en	pizza
jeg	0.002	0.333	0	0	0	0.001	0	0
ønsker	0.111	0	0.4	0.218	0.4	0.271	0.2	0.211
å	0.002	0	0	0.432	0	0.400	0	0
bestille	0	0.1	*0.022**	0	0.422	0	0.288	0.3
pasta	0	0.002	*0.032**	0	0	0.174	0	0
betale	0	0.003	0.032	0	0.299	0	0.367	0.375
en	0.003	0	0.23	0.2	0.211	0	0.001	0.401
pizza	0	0.101	0.005	0.009	0	0	*0.01**	0

** Skrivefeil i oppgaveteksten som ikke påvirker det vi trenger til oppgaven.*

La s = <s> jeg ønsker å bestille en pizza </s>

$P(s) = P(\text{jeg} \mid \text{<s>}) \cdot P(\text{ønsker} \mid \text{jeg}) \cdot P(\text{å} \mid \text{ønsker}) \cdot P(\text{bestille} \mid \text{å}) \cdot P(\text{en} \mid \text{bestille}) \cdot P(\text{pizza} \mid \text{en}) \cdot P(\text{</s>} \mid \text{pizza})$

= 0.422 * 0.333 * 0.4 * 0.432 * 0.288 * 0.401 *0.232

Bigram-sannsynligheter:

P(jeg | <s>) = 0.422
~~P(pizza | to) = 0.001~~

P(</s> | pizza) = 0.232
~~P(</s> | pasta) = 0.001~~

Syntaks

Deloppgavene 4.1–4.3 som følger vil alle ta utgangspunkt i en grammatikk vi antar har følgende regler (der tegnet | angir valg mellom ulike alternativer):

- S → NP VP
- VP → V NP
- VP → VP PP
- NP → NP PP
- PP → P NP
- P → med
- NP → jeg | sushi | pinner | tofu
- V → spiser

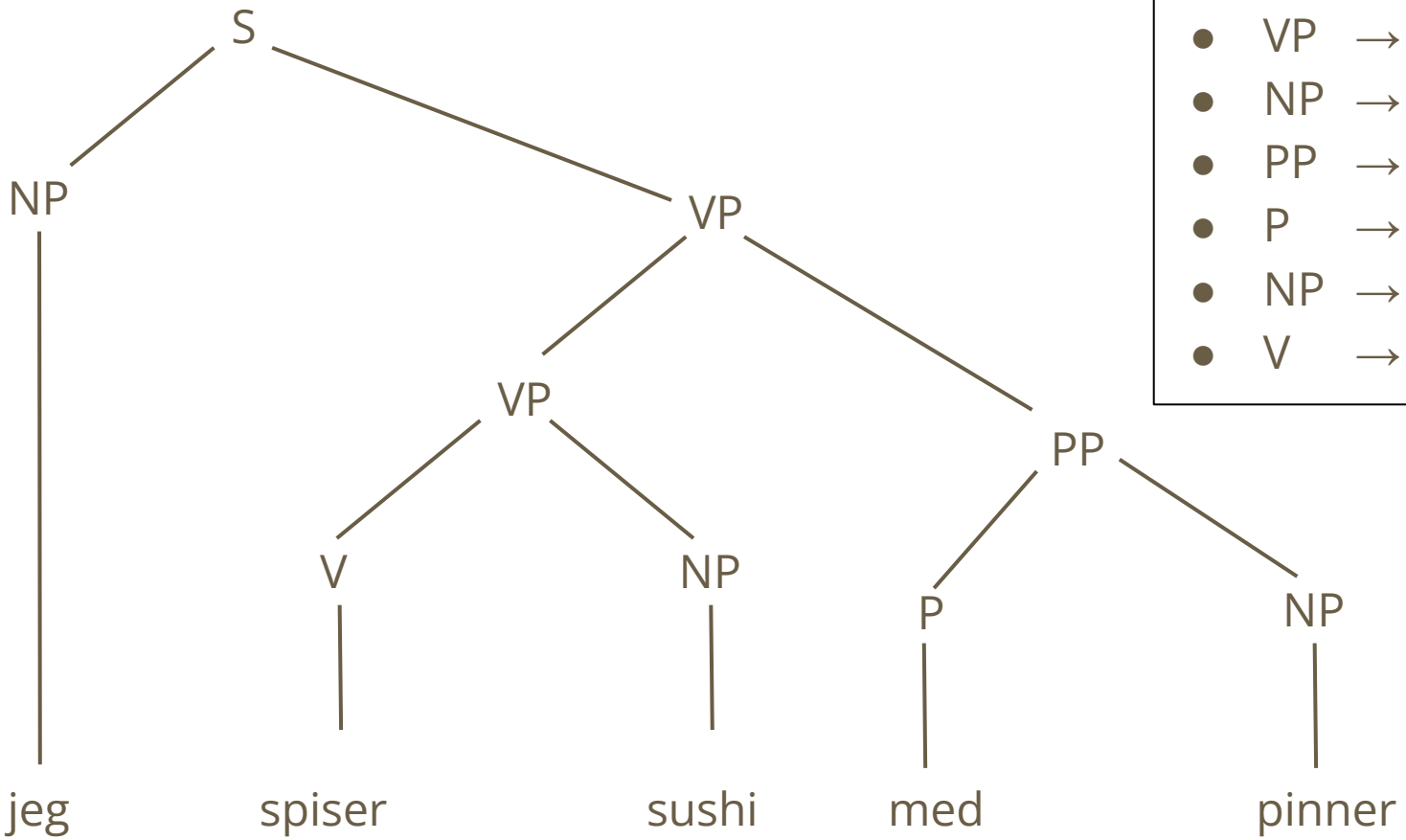
4.1 Frasestrukturtrær

I denne oppgaven kan du svare med digital håndtegning. Bruk eget skisseark (utdelt). Se instruksjon for utfylling av skisseark i lenken under oppgavelinjen.

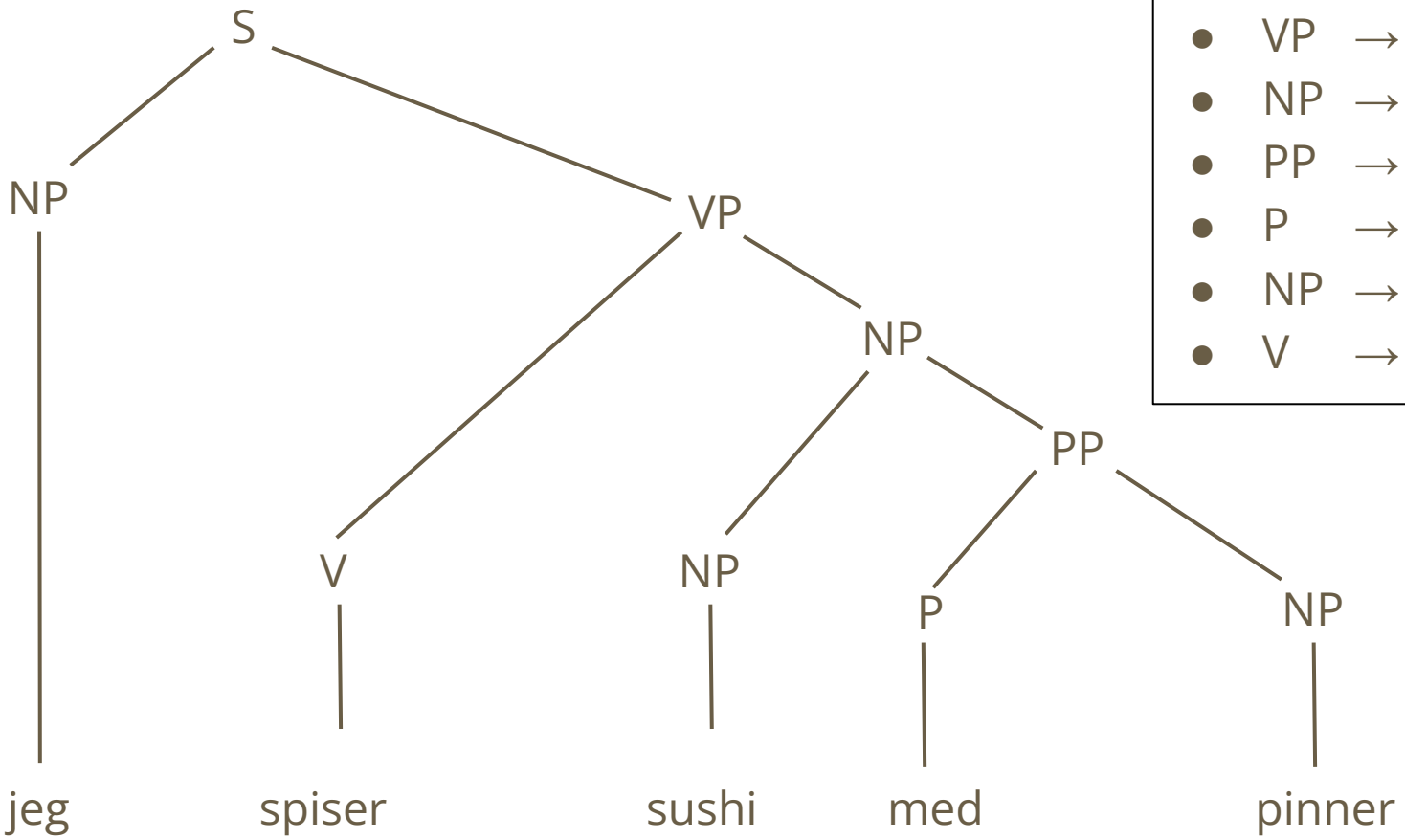
Tegn alle frasestrukturtrær du mener grammatikken tillater for følgende setning:

jeg spiser sushi med pinner

- S → NP VP
- VP → V NP
- VP → VP PP
- NP → NP PP
- PP → P NP
- P → med
- NP → jeg | sushi | ...
- V → spiser



- S → NP VP
- VP → V NP
- VP → VP PP
- NP → NP PP
- PP → P NP
- P → med
- NP → jeg | sushi | …
- V → spiser



4.2 Grammatikalitet

Er de følgende setningene grammatiske i følge grammatikken vår?

	Ja	Nei
jeg spiser		✓
jeg spiser med pinner		✓
jeg spiser pinner med sushi	✓	
tofu spiser pinner med sushi	✓	
sushi med pinner spiser jeg	✓	
Jeg spiser sushi med tofu med pinner	✓	

- S → NP VP
- VP → V NP
- VP → VP PP
- NP → NP PP
- PP → P NP
- P → med
- NP → jeg | sushi | pinner | tofu
- V → spiser

4.1 Rekursjon

Forklar kort hva vi mener med at en grammatikk kan være rekursiv. Er grammatikken vi har brukt i oppgavene over rekursiv? Forklar svaret ditt.

Løsningsforslaget

En grammatikk er rekursiv dersom den inneholder regler der ikke-terminale kategorier på høyresiden kan ekspanderes til produksjoner som inneholder ikke-terminalen på venstre-siden. Forklart med utgangspunkt i frasestrukturtrær kan vi si at grammatikken er rekursiv dersom det er mulig for en node å dominere en annen node av samme type. Rekursjon kan være direkte eller indirekte. Ved *direkte* rekursjon vil kategorien på venstresiden av regelen også forekomme på høyresiden. I det tilsvarende treet vil altså en node direkte dominere en annen node av samme type. Eksempler på direkte rekursive regler i den gitte grammatiken er NP → NP PP og VP → VP PP. Ved *indirekte* rekursjon vil samme node fra venstresiden av en en regel opptre senere i videre ekspansjon av høyresiden. I det tilsvarende treet vil altså en node dominere en annen node av samme type i et subtre. Eksempler på indirekte rekursive regler i den gitte grammatiken er kombinasjonen av NP → NP PP og PP → P NP. Grammatikken er altså rekursiv.

4.3 Rekursjon

Forklar kort hva vi mener med at en grammatikk kan være rekursiv. Er grammatikken vi har brukt i oppgavene over rekursiv? Forklar svaret ditt.

- En node i grammatikken leder til seg selv
  - Enten direkte eller indirekte
    - Et eksempel på hver av disse
  - Hva fører dette til i praksis? (Vilkårlig antall konstituenter)
- Ja/nei, grammatikken er (ikke) rekursiv.
  - Fordi ...

4.4 Konstituenter

Ta utgangspunkt i følgende to setninger:

1. *Hun løp opp bakken.*
2. *Hun slo opp nummeret.*

Er ordsekvensene "opp bakken" og "opp nummeret" konstituenter i setningene over? Forklar hvordan vi kan avgjøre dette ved bruk av minst tre ulike konstituenttester.

“opp bakken”

Tester:

- Kan stå alene:
  - Q: Hvor løp hun?
  - A: Opp bakken.
- Erstattes med pronomen:
  - Hun løp der.
- Flyttes som en enhet:
  - Opp bakken løp hun.



“opp nummeret”

Tester:

- Kan stå alene:
  - Q: Hvor løp hun?
  - A: *Opp nummeret.
- Erstattes med pronomen:
  - *Hun slo det.
- Flyttes som en enhet:
  - *Opp nummeret slo hun.



Semantikk

5.1 Semantiske Roller 1

I denne oppgaven skal du velge hvilken *semantisk rolle* som *best* beskriver ordet markert med fet skrift i setningen under:

> ***Ballen**, som ble kastet hardt, knuste ruta.*

- ❏ Instrument
- ❏ Agent
- ❏ Theme
- ❏ Patient

5.1 Semantiske Roller 1

I denne oppgaven skal du velge hvilken *semantisk rolle* som *best* beskriver ordet markert med fet skrift i setningen under:

***Ballen**, som ble kastet hardt, knuste ruta.*

➔ Instrument

~~❑ Agent~~

~~❑ Theme~~

~~❑ Patient~~



5.2 Semantiske Roller 2

I denne oppgaven skal du velge hvilken *semantisk rolle* som *best* beskriver ordet markert med fet skrift i setningen under:

> *Mina klippet* ***gresset.***

- ❏ Theme
- ❏ Patient
- ❏ Goal
- ❏ Experiencer

5.2 Semantiske Roller 2

I denne oppgaven skal du velge hvilken *semantisk rolle* som *best* beskriver ordet markert med fet skrift i setningen under:

*Mina klippet **gresset.***

- ~~❑ Theme~~
- ➔ Patient
- ~~❑ Goal~~
- ~~❑ Experiencer~~



5.3 Semantiske Roller 3

I denne oppgaven skal du velge hvilken *semantisk rolle* som *best* beskriver ordet markert med fet skrift i setningen under:

> ***Louise*** *hørte et smell bak seg.*

- ❏ Beneficiary
- ❏ Experiencer
- ❏ Agent
- ❏ Patient

5.3 Semantiske Roller 3

I denne oppgaven skal du velge hvilken *semantisk rolle* som *best* beskriver ordet markert med fet skrift i setningen under:

Louise *hørte et smell bak seg.*

- ~~Beneficiary~~
- ➔ Experiencer
- ~~Agent~~
- ~~Patient~~



5.4 Semantiske Roller 4

I denne oppgaven skal du velge hvilken *semantisk rolle* som *best* beskriver ordet markert med fet skrift i setningen under:

> *Den **boka** du bestilte er på vei nå.*

- ❏ Patient
- ❏ Source
- ❏ Theme
- ❏ Instrument

5.4 Semantiske Roller 4

I denne oppgaven skal du velge hvilken *semantisk rolle* som *best* beskriver ordet markert med fet skrift i setningen under:

*Den **boka** du bestilte er på vei nå.*

- ~~Patient~~
- ~~Source~~
- ➔ Theme
- ~~Instrument~~



5.5 Semantiske Roller 5

I denne oppgaven skal du velge hvilken *semantisk rolle* som *best* beskriver ordet markert med fet skrift i setningen under:

> *Jeg ga **ham** en genser til jul.*

- ❏ Beneficiary
- ❏ Patient
- ❏ Theme
- ❏ Experiencer

5.5 Semantiske Roller 5

I denne oppgaven skal du velge hvilken *semantisk rolle* som *best* beskriver ordet markert med fet skrift i setningen under:

*Jeg ga **ham** en genser til jul.*

- ➔ Beneficiary
- ~~Patient~~
- ~~Theme~~
- ~~Experiencer~~



5.6 Semantiske Roller 6

I denne oppgaven skal du velge hvilken *semantisk rolle* som *best* beskriver ordet markert med fet skrift i setningen under:

> *Kaken ble spist opp av **hunden**.*

- ❏ Agent
- ❏ Instrument
- ❏ Theme
- ❏ Patient

5.6 Semantiske Roller 6

I denne oppgaven skal du velge hvilken *semantisk rolle* som *best* beskriver ordet markert med fet skrift i setningen under:

*Kaken ble spist opp av **hunden**.*

➔ Agent

~~❑ Instrument~~

~~❑ Theme~~

~~❑ Patient~~



5.7 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. Gi dem **fingeren**, og de vil ta hele **hånden**.
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. **Liljekonvall** er en **plante** du absolutt ikke bør få i deg.
4. Årstidene **kommer** og **går**, år etter år.
5. Jeg **går** på veien som **går** fra øst til vest.
6. Jeg **går** til en **gård.**

Homonymi	Meronymi	Hyponymi	Antonymi	Polysemi

5.7 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. Gi dem **fingeren**, og de vil ta hele **hånden**.
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. **Liljekonvall** er en **plante** du absolutt ikke bør få i deg.
4. Årstidene **kommer** og **går**, år etter år.
5. Jeg **går** på veien som **går** fra øst til vest.
6. Jeg **går** til en **gård.**

Homonymi	Meronymi	Hyponymi	Antonymi	Polysemi
	✓			
✓				
		✓		
			✓	
				✓
✓				

5.8 Komposisjonalitet

Forklar hva som menes med prinsippet om komposisjonalitet i semantikk. Begrens svaret ditt til 1–3 setninger.

Mange måter å formulere dette på.

Essensen er at **betydningen av et komplekst uttrykk,** altså et uttrykk med flere deler, bestemmes av **betydningen til konstituentene og reglene som brukes for å kombinere dem.** En enkel måte å beskrive dette på kan være å si at *betydningen til en setning er betydningen til hvert enkelt ord + reglene som setter dem sammen.*

Maskinlæring / Naive Bayes

6.1 Veiledet læring

Forklar med 1–2 setninger hva som menes med begrepet "veiledet læring" (*supervised learning*) innen maskinlæring.

Løsningsforslaget: *Veiledet læring referer til maskinlæringsalgoritmer der det brukes annoterte (labeled) treningsdata, dvs. eksempler på input-data samstilt med riktig output-verdi som vi ønsker at modellen skal lære å predikere.*

Det viktigste: Maskinlæring med annotert treningsdata.

6.2 Naive Bayes

Som del av klassifikasjonsmetoden Naive Bayes gjøres en forenklende (eller "naiv") antakelse.

Svar på følgende:

1. Forklar hva denne antakelsen innebærer.

2. Denne naive eller forenklende antakelsen reflekteres også i formelen som brukes for å regne ut sannsynlighetene vi trenger i Naive Bayes. Vis hvordan antakelsen uttrykkes i formelen.

1. Forklar hva denne antakelsen innebærer.

Løsningsforslaget: *Vi antar at sannsynligheten for de ulike trekkene er uavhengige av hverandre, gitt klassen. Dette gjør det mer håndterbart å estimere sannsynlighetene.*

2. Naive Bayes-naivitet

Løsningsforslaget: *Dersom vi antar at et gitt objekt (f.eks et dokument) representeres ved n trekk, f1, f2, . . . , fn, så kan den såkalte likelihood-termen tilnærmes som* $P(f_1, f_2, \ldots, f_n \mid c) \approx \prod^{n}_{i=1} (f_i \mid c)$.

Å gjengi hele NB-formelen gir også uttelling, men det bør gå frem hvilken del av formelen som er relevant og man bør vise hvordan formelen ser ut både før/etter uavhengighetsantakelsen (altså begge sider av ≈).

Hovedantakelse

$$\hat{k} = \underset{k \in K}{\operatorname{argmax}} P(k|\vec{v}) \xrightarrow[\text{ved Bayes teorem}]{P(A|B) = \frac{P(B|A)P(A)}{P(B)}} \hat{k} = \underset{k \in K}{\operatorname{argmax}} \frac{P(\vec{v}|k)P(k)}{P(\vec{v})} \xrightarrow[\text{antakelsen}]{P(\vec{v}|k) \approx \prod_{j=1}^{n} P(v_j|k)} \hat{k} = \underset{k \in K}{\operatorname{argmax}} P(k) \prod_{j=1}^{n} P(v_j|k)$$

Omformulering ved Bayes teorem

Uavhengighets-antakelsen

6.3 Glatting

Forklar med noen få setninger hva som menes med begrepet "glatting" (*smoothing*), inkludert når/hvorfor det brukes. Nevn eksempler fra pensum der glatting har vært relevant å ta i bruk.

Løsningsforslaget

Glatting er en type teknikk som brukes for å *omfordele sannsynlighetsmasse* fra hendelser med mange observasjoner til mer sjeldne hendelser der vi har få eller ingen observasjoner i treningsdataene. Vi har brukt glatting i forbindelse med å estimere de betingede sannsynlighetene i *n*-gram-*språkmodeller* og *klassifikasjon med Naive Bayes*. Der så vi at å bruke såkalte "*maximum likelihood estimates*" (MLE) alene, gitt ved relative frekvenser i treningsdataene, kunne medføre det såkalte "*zero frequency problem*" når modellen skal anvendes på nye usette data: Manglende observasjoner i treningsdataene, for henholdsvis gitte *n*-grammer eller trekk/klasse-par, vil føre til at sannsynligheten for hele sekvensen eller sannsynligheten for en klasse gitt alle trekkene blir null (fordi alle de individuelle sannsynlighetene multipliseres sammen). Det finnes mange ulike måter å gjøre glatting på, men som eksempel har vi sett på metoden *add-1* eller *Laplace smoothing*, som fungerer ved å simpelthen legge til én til frekvensen av alle hendelser i sannsynlighetsestimatet.

Nøkkelord: Få observasjoner, n-gram, add-1

Språkteknologiske anvendelser

7.1 Named Entity Regonition

Den vanligste måten å løse Named Entity Recognition (`navnegjenkjenning') på er ved ord-for-ord klassifisering med bruk av såkalte BIO-labels.

Anta følgende setning:

> *Spilte Toralv Maurstad Henrik Ibsen med Riksteateret, på Stokmarknes i fjor?*

Angi riktig BIO-klassifisering med riktig kategori for hvert ord i setningen. Her opererer vi med følgende kategorier: PER (person), ORG (organisasjon), LOC (location), GPE (geopolitical entity)

7.1 Named Entity Recognition

Spilte Toralv Maurstad Henrik Ibsen med Riksteateret, på Stokmarknes i fjor?

	O	B_PER	I_PER	B_ORG	I_ORG	B_LOC	I_LOC	B_GPE	I_GPE
Spilte	✓								
Toralv		✓							
Maurstad			✓						
Henrik		✓							
Ibsen			✓						
med	✓								
Riksteateret				✓					
,	✓								
på	✓								
Stokmarknes						✓			
i	✓								
fjor	✓								
?	✓								

7.2 Ordsemantikk

Gitt følgende tre setninger:

a) *Kan du kutte gulrøttene?*

b) *Kan du kutte røyken?*

c) *Kan du kutte gulrøttene og røyken?*

Vi ser at verbet "kutte" er et eksempel på et **flertydig ord.**

Betydning:
a) "*kappe*" eller "*skjære*"
b) "*redusere*" eller "*stoppe*"

⇒ "*kutte*" er et **polysem**, altså et flertydig ord med ulike men *relaterte betydninger*

c) er et **zeugma**: en konjunksjon av ledd som hører til ulike måter å bruke verbet på for å vurdere om det er snakk om ulike betydninger. *At setningen ikke fremstår som semantisk velformet, viser oss flertydigheten til ordet*

Bruk noen få setninger på å beskrive og sammenlikne forekomstene av ordet «kutte» i eksempelsetningene a) og b) over. Hvordan kan setning c) hjelpe oss å trekke slutninger om ordet «kutte»? Bruk gjerne relevant fagterminologi der du kan.