Løsningsforslag til eksamen i IN1140, høst 2022

- Husk på at løsningsforslagene som står her er nettopp det: forslag. Det finnes typisk flere måter å gjøre ting på.

1 Regulære uttrykk (12 poeng)

1.1 Norske datoformater (6 poeng)

Velg ett eller flere alternativer

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

1.2 Grupperinger (3 poeng)

Velg ett alternativ:

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

1.3 Disjunksjon (3 poeng)

Velg ett eller flere alternativer

- [x] ab ✔
- [x] cd ✔
- [ ] acd
- [ ] abcdefg
- [ ] ace
- [ ] abdfg

2 Ordklasser (9.5 poeng)

2.1 Ordklassetagging (5 poeng)

MERKNAD: Her ble det knot i forklaringen av forkortelsene for hver ordklasse som var plassert sammen med oppgaveteksten. Feilen var at **Subs** pekte til Subjunksjon, men skulle ha vært **Subj**, og at ingenting pekte til Substantiv. Om det er noen som har blandet disse to, altså satt *veien* til å være **Subj** og *om* til å være **Subs**, så gis full uttelling.

	Prep	Pron	Vrb	Adj	Adv	Subs	Subj	Det	Konj
på	✓								
fort					✓				
kanskje					✓				
veien						✓			
Hva		✓							
en								✓	
Om							✓		
slukte			✓						
eller									✓
rustne				✓					

2.2 Orddannelse (4.5 poeng)

	avlednings-prefiks	både bøynings- og avlednings-affiks	bøynings-suffiks	bøynings-prefiks	avlednings-suffiks	ingen affikser
tjenestene		✓				
tjeneste					✓	
tjene			✓			✓
fortjene	✓	✓				
tjente			✓			
betjener		✓				

Å konsekvent anse infinitiven for å ha bøyningssuffiks gir også poeng.

3 Språkmodeller (13 poeng)

3.1 Trigram (3 poeng)

Riktig svar er **18**

3.2 Markovantagelsen (4 poeng)

Riktig svar er: *For å kunne gjøre en statistisk tilnærming av betingende sannsynligheter uten å måtte se for langt bakover i historikken.*

3.3 Bigram (6 poeng)

Studenten må benytte både utdrag fra tabell + de oppgitte sannsynlighetene fra en fiktiv kollega for å sette opp stykket. Oppgaven ber om bigram-sannsynligheten for en setning, noe som innebærer at alle bigram-sannsynlighetene må ganges sammen. Bigrammene vi trenger er:

P(jeg | <s>) * P(ønsker | jeg) * P(å | ønsker) * P(bestille | å) * P(en | bestille) P(pizza | en) * P(</s> | pizza)

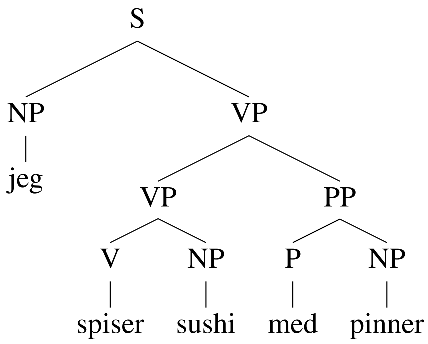
Med faktiske verdier blir dette:

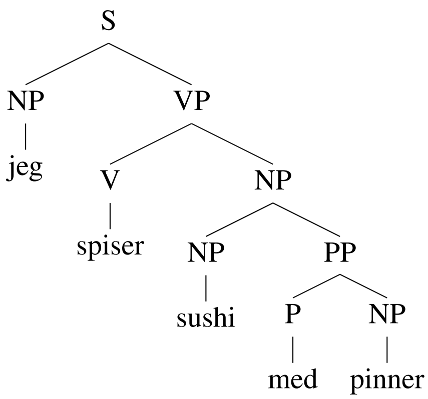
= 0.422 * 0.333 * 0.4 * 0.432 * 0.288 * 0.401 *0.232

Merknad Det manglet en verdi i cellen for P(en | pizza), som altså svarer til teksten *pizza en*, som kan ha skapt forvirring, men om oppgaven er løst riktig har ikke dette noen innflytelse, da bigrammet som behøves her er P(pizza | en), som svarer til strengen *en pizza*. Se også eksempel i Jurafsky & Martin, kapittel 3 side 6.

4 Syntaks (24 poeng)

4.1 Frasestrukturtrær (6 poeng)





4.2 Grammatikalitet (6 poeng)

Er de følgende setningene grammatiske i følge grammatikken vår?

	Ja	Nei
jeg spiser	○	◉
jeg spiser sushi	◉	○
jeg spiser med pinner	○	◉
tofu spiser pinner med sushi	◉	○
sushi med pinner spiser jeg	◉	○
jeg spiser sushi med tofu med pinner	◉	○

4.3 Rekursjon (5 poeng)

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 $\rightarrow$ NP PP og VP $\rightarrow$ 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 $\rightarrow$ NP PP og PP $\rightarrow$ P NP. Grammatikken er altså rekursiv.

4.4 Konstituenter (7 poeng)

- "opp bakken" er en konstituent i den første setningen.
- Tester:
  - Kan stå alene: *Hvor løp hun? Opp bakken.*
  - Erstattes med pronomen: *Hun løp der.*
  - Flyttes som en enhet: *Opp bakken løp hun.*
- "opp nummeret" er IKKE en konstituent i den andre setningen.
- Tester (der * indikerer at setningen ikke høres velformet ut, semantisk eller syntaktisk):
  - Kan stå alene: *Hva slo hun? *Opp nummeret.*
  - Erstattes med pronomen: **Hun slo det.*
  - Flyttes som en enhet: **Opp nummeret slo hun.*

5 Semantikk (15 poeng)

5.1 Semantiske roller 1 (1 poeng)

Instrument: det uttrykkes eksplisitt at noen har kastet ballen.

5.2 Semantiske roller 2 (1 poeng)

Patient: Gresset endrer tilstand.

5.3 Semantiske roller 3 (1 poeng)

Experiencer: Louise opplever noe (en lyd), men hun gjør det ikke aktivt.

5.4 Semantiske roller 4 (1 poeng)

Theme: Boka er i bevegelse

5.5 Semantiske roller 5 (1 poeng)

Beneficiary: Mottager av gave

5.6 Semantiske roller 6 (1 poeng)

Agent: Det er hunden som spiser kaken (en aktiv handling, litt avhengig hvor mye bevissthet man ønsker å tillegge en hund).

5.7 Leksikale relasjoner (6 poeng)

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

5.8 Komposisjonalitet (3 poeng)

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.

6 ML / NB (14 poeng)

6.1 Veiledet læring (3 poeng)

Veildet 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.

6.2 Naive Bayes (6 poeng)

- Vi antar at sannsynligheten for de ulike trekkene er uavhengige av hverandre, gitt klassen. Dette gjør det mer håndterbart å estimere sannsynlighetene.
- Dersom vi antar at et gitt objekt (f.eks et dokument) representeres ved n trekk, $f_1, f_2, \ldots, f_n$, så kan den såkalte *likelihood*-termen tilnærmes som $P(f_1, f_2, \ldots, f_n|c) \approx \prod_{i=1}^{n}(f_i|c)$. 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 $\approx$).

6.3 Glatting (5 poeng)

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 betingete 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 frequencey 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 inviduelle 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.

7 Språkteknologiske anvendelser (12.5 poeng)

7.1 Named Entity Recognition (6 poeng)

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

7.2 Ordsemantikk (6.5 poeng)

- Vi ser at verbet “kutte” er et eksempel på et *flertydig* ord.
- I setning a) brukes det i betydningen “kappe” eller “skjære”, mens det i b) brukes i betydningen “redusere” eller “stoppe”.
- Vi ser at verbet “kutte” et *polysem*, altså et flertydig ord med ulike men relaterte betydninger. (Vi gir også halv uttelling for å heller si at det er et *homonym / homograf*, altså er ord med flere distinkte betydninger.)
- Setning c) ser vi et såkalt *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.