IN1140

Eksamen H2021

1 Regulære uttrykk (4 poeng)

Anta følgende regulære uttrykk:

```
regex = "((\+47|0047)\s)?([0-9][0-9]\s){3}([0-9][0-9]){1}"
```

Hvilken av følgende påstander om uttrykket er usann? Begrunn svaret ditt.

1. Det regulære uttrykket gjenkjenner kun norske telefonnumre som begynner med +47 eller 0047
2. Det regulære uttrykket gjenkjenner norske telefonnumre som kan begynne med +47 eller 0047

Løsningsforslag

Den første er feil, telefonnumre kan starte uten +47 eller 0047, da det er opsjonelt pga bruk av "?"

---

2 Regulære uttrykk (15 poeng)

Skriv et regulært uttrykk som sjekker om en postadresse er gyldig i Norge. Merk at for at posten skal kunne levere brevene/pakkene, må adressen skrives på ett av følgende formater:

```
Fornavn Etternavn
Gatenavn Husnummer
Postkode By

Fornavn Etternavn
P.O. BOX Nummer
Postkode By
```

```
Fornavn Etternavn
Postboks Nummer
Postkode By
```

Merk følgende formateringer:

- Husnummer kan være tall mellom 01 og 99.
- Nummer kan være tall mellom 0001 og 9999.
- Postkode kan være tall mellom 0001 og 5999.
- Det skal være kun ett fornavn, og kun ett etternavn.
- Det er mellomrom mellom hver element, utenom etter Etternavn, Husnummer, og Nummer, der skal det være ny linje.
- Uttrykket ditt må kunne håndtere de norske bokstavene æ, ø, å, Æ, Ø, og Å.

Eksempelvis skal følgende strenger gjenkjennes av ditt regulære utrykk:

```
Per Post
Storgata 15
0155 Oslo

Bedriften AS
P.O. BOX 9999
3705 Skien

Julius Box
Postboks 6890
2712 Brandbu
```

Løsningsforslag

```
regex = "[a-zA-ZæøåÆØÅ]+\s[a-zA-ZæøåÆØÅ]+\n([a-zA-ZæøåÆØÅ]+
\s[0-9][1-9]|(P\.\O.\sBOX|Postbox)\s[0-9]{3}[1-9])\n
[0-5][0-9]{2}[1-9]\s[a-zA-ZæøåÆØÅ]+"
```

---

3 Tokenisering (15 poeng)

- Beskriv kort hva tokenisering er og hvordan det typisk inngår i et språkteknologisk system. Hvilke andre språkteknologiske oppgaver er avhengig av tokenisering? Gi minst ett eksempel.
- Hvordan kan output fra en tokeniserer se ut for denne teksten? Skriv en tokenisert versjon av teksten under.
- Tenk deg at du skal lage en regelbasert tokeniserer for teksten under. Hvilke utfordringer ser du som du må ta høyde for i tokenisereren din? Gi minst tre eksempler fra teksten.

```
Vi søker motiverte medarbeidere og "Team-players" til vår nye super-satsing!
Du må kunne salg og være deg selv 110%. For mer informasjon om lønn,
arbeidvilkår, osv., ta kontakt via vår hjemmeside http://www.supersjansen.no.
Vi ser fram til å høre fra deg!
```

Løsningsforslag

- En tokeniserer er et system som deler opp en tekst i løpende ord. Dette gjøres typisk ved å skille ut tegnsetting. Tokenisering er første ledd i mange språkteknologiske oppgaver. Eksempelvis vil både ordklassetagging og Word Sense Disambiguation avhenge av tokenisering som første skritt for å få tak i ordene som skal være input til ordklassetaggeren eller WSD-systemet.
- En foreslått tokenisert tekst:

  ```
  Vi søker motiverte medarbeidere og " Team-players " til vår nye
  super-satsing ! Du må kunne salg og være deg selv 110 % .
  For mer informasjon om lønn , arbeidvilkår , osv. , ta kontakt
  via vår hjemmeside http://www.supersjansen.no . Vi ser fram til å
  høre fra deg !
  ```

  (Merknad: Ved gjennomgang ser jeg at en gjennomgående "misforståelse" i tokeniseringsoppgaven er at de først angir en forenklet tokenisering der man kun splitter på mellomrom (og muligens tegnsetting) i oppgave 2 og deretter beskriver problematiske tilfeller i oppgave 3 (urler, forkortelser osv). Denne tolkningen av oppgaven er helt fin og bør gi full uttelling da den viser forståelse og speiler arbeidsflyten vi hadde i kurset, der de først implementerte en forenkelt tokeniserer og deretter forbedret den. Med en slik fordeling er det viktig at besvarelsen tydelig viser (enten i deloppgave 2 eller 3) at de har forstått at tokenisering skiller ut tegnsetting fra ordene (slik at det ikke feks har "arbeidsvilkår," eller "deg!").

- I teksten er det en rekke forekomster av ulike typer tegnsetting som må behandles særskilt i en regelbasert tokeniserer, særlig gjelder det tegnsetting som ikke skal skilles ut som egne tokens, eksempelvis når de forekommer som del av en forkortelse (f.eks.), et sammensatt ord (Team-players) eller en URL (http://www.supersjansen.no)

---

4 Språkmodell (15 poeng)

Anta at vi vil trene en bigram-modell på følgende korpus:

```
<s> jeg er rar </s>
<s> rar er jeg </s>
<s> rar er rar </s>
```

1. Hvor mange unike bigram forekommer i korpuset?
2. Hva er deres sannsynlighet? Vis hvordan du beregner sannsynligheten for hvert bigram fra korpuset.
3. Gitt bigram-sannsynlighetene du beregnet over, hva er sannsynligheten for følgende setning: ¡s¿ jeg er jeg ¡/s¿
   Vis hvordan du kom fram til svaret ditt.

Løsningsforslag

1. 8 bigram
2. Vi beregner sannsynlighetene slik:

   ```
   P(jeg|<s>) = C(<s> jeg/<s>) = 1/3 = 0.33
   P(er|jeg) = C(jeg er/jeg) = 1/2 = 0.5
   P(rar|er) = C(er rar/er) = 2/3 = 0.67
   P(</s>|rar) = C(rar </s>/rar) = 2/4 = 0.5
   P(rar|<s>) = C(<s> rar/<s>) = 2/3 = 0.67
   P(er|rar) = C(rar er/rar) = 2/4 = 0.5
   P(jeg|er) = C(er jeg/er) = 1/3 = 0.33
   P(</s>|jeg) = C(jeg </s>/jeg) = 1/2 = 0.5
   ```

3. Basert på sannsynlighetene beregnet i forrige oppgave for følgende bigram:

   ```
   <s> jeg
   jeg er
   er jeg
   jeg </s>
   ```

kan vi regne ut sannsynligheten for setningen slik:
0.33 x 0.5 x 0.33 x 0.5 = 0.027225

---

5 Ordklasser (7 poeng)

Gitt ordklassene i Tabell 1 under, tildel ordklasser til alle ordene i setningen. Du må velge ett alternativ for hvert ord.
Når vinteren senker seg sakte over byen er det tid for et varmt bad

---

Løsningsforslag

Når SUBJ
vinteren SUBST
senker VERB
seg PRON
sakte ADV
over PREP
byen SUBST
er VERB
det PRON
tid SUBST
for PREP
et DET
varmt ADJ
bad SUBST

---

6 Ordklassetagging (8 poeng)

Vi ønsker å lage et system for å automatiske merke opp en tekst med ordklasser. Som et utgangspunkt utvikler vi et system der vi slår opp i et leksikon der hvert ord er assosiert med én ordklasse. Systemet vårt viser seg dessverre å ikke fungere særlig bra.

- Hva er grunnen til det tror du? Illustrer svaret ditt med eksempler.
- Hva kan vi gjøre for å forbedre systemet vårt?

Løsningsforslag

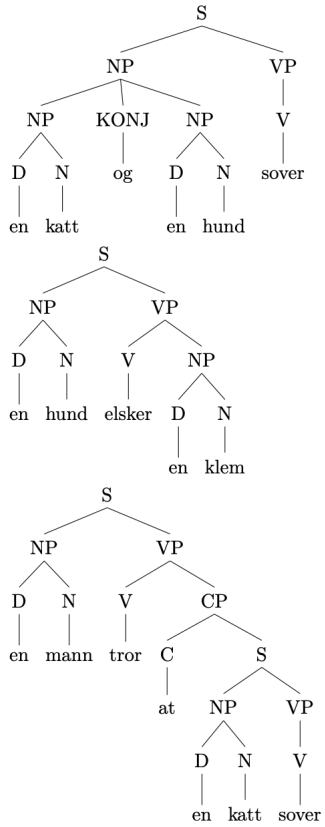
Eksempelsvar:
1.Hovedproblemet her er knyttet til flertydighet. Mange ord er flertydige, dvs de kan tildeles flere ulike ordklassetagger avhengig av hvilken kontekst de forekommer i. Eksempelvis kan ordet "rett" analyseres som både substantiv, adjektiv og adverb. Særlig er det slik i språk at mange av de mest frekvente ordene i språk også er blant de mest flertydige.

2.Vi trenger derfor et system som tar konteksten med i betraktningen og ikke blindt tildeler en tagg for hvert ord. Dette kan vi gjøre både med regelbaserte og statistiske metoder, men her har statistiske modeller en fordel. De kan trenes på manuelt annoterte datasett og generalisere til nye og usette data og er utformet for å betinge på konteksten.

7 Grammatikk (6 poeng)

Trærne på bildet utgjør en liten trebank, et syntaktisk annotert korpus som vi skal benytte oss av i denne oppgaven.

- 1. Utled en frasestrukturgrammatikk fra de syntaktiske analysene i korpuset og angi grammatikken under.
- 2. Er grammatikken din rekursiv? Begrunn svaret ditt.



Løsningsforslag

Eksempelsvar:

1.

```
S -> NP VP
NP -> D N
NP -> NP KONJ NP
VP -> V
VP -> V NP
VP -> V CP
CP -> C S

D -> en
N -> katt, hund, trapp, klem, mann
V -> sover, elsker, tror
```

```
KONJ -> og
C -> at
```

2.
Ja, grammatikken er rekursiv, samme syntaktiske kategori forekommer på venstre og høyresiden i en regel. Den inneholder følgende (direkte) rekursive regler:

```
NP -> NP KONJ NP
```

Grammatikken inneholder også indirekte rekursjon ved følgende regelkombinasjon, der en setning kan forekomme inne i en annen setning:

```
S -> NP VP
VP -> V CP
CP -> C S
```

---

8 Utvidelse av grammatikken (5 poeng)

Vi vil at grammatikken vår videre skal kunne gi en analyse for setninger som: Hvilke regler må du utvide grammatikken din med for at den skal kunne gi en analyse for følgende setninger?

En katt sitter eller sover
En hund elsker en mann og hater en katt

Løsningsforslag

```
VP -> VP KONJ VP
V -> sitter
V -> hater
KONJ -> eller
```

---

9 Ugrammatiske analyser (4 poeng)

Grammatikken kan gi opphav til noen ugrammatiske setninger. Diskutér påstanden kort og illustrér med et eksempel.

Løsningsforslag

Eksempelsvar:

en mann sover at en katt sitter
en katt elsker

---

10 Konstituentkriterier (5 poeng)

Vi ønsker å avgjøre hvorvidt "en hund og en katt" er en konstituent i følgende setning: *en hund og en katt fanger en fugl*
Vis hvordan du kan bruke minst to konstituentkriterier for å avgjøre dette.

Løsningsforslag

Eksempelsvar:

- stå alene: Hvem fanger en fugl? En hund og en katt
- erstatte med pronomen: De fanger en fugl.
- flytte som enhet: Det er en hund og en katt som fanger en fugl. Alternativt: En fugl blir fanget av en hund og en katt.

Ja, dette er en konstituent.

---

11 Leksikale Relasjoner (3 poeng)

Hvilken semantisk relasjon holder mellom følgende ord-par? NB! Her får du poeng for riktig svar, men ikke negative poeng for feil svar.

	antonymi	hyponymi	meronymi	synonymi
fot -- tå	○	○	○	○
singel -- gift	○	○	○	○
inne -- ute	○	○	○	○
sommerfugl -- insekt	○	○	○	○
demokrati -- folkestyre	○	○	○	○
pen -- vakker	○	○	○	○

Løsningsforslag

Eksempelsvar:

- fot – tå **meronymi**
- singel – gift **antonymi**
- inne – ute **antonymi**
- sommerfugl – insekt **hyponymi**
- demokrati – folkestyre **synonymi**
- pen – vakker **synonymi**

---

12 Semantiske roller (3 poeng)

Angi semantisk rolle for de *uthevede* ordene (her får du poeng for riktig svar, men ikke negative poeng for feil svar):

	GOAL	SOURCE	INSTRUMENT	AGENT	THEME	EXPERIENCE
Jon rører i suppa	○	○	○	○	○	○
Jon kjører mot *Bergen*	○	○	○	○	○	○
Han kommer fra *Oslo*	○	○	○	○	○	○
Jon hører en lyd	○	○	○	○	○	○
Nøkkelen åpner døren	○	○	○	○	○	○
Jon skjærer brød med *en kniv*	○	○	○	○	○	○

Løsningsforslag

Eksempelsvar:

- Jon **agent**
- Bergen **goal**
- Oslo **source**
- Jon **experiencer**
- Nøkkelen **instrument**
- En kniv **instrument**

---

13 Evaluering (10 poeng)

En måte å evaluere dialogsystemer på er ved å be mennesker manuelt analysere og og se gjennom to dialoger produsert av systemet, og velge den dialogen de synes var best. Da vi evaluerte vår Naive Bayes sentimentklassifiserer, brukte vi imidlertid en annen måte

å evaluere modellen vår på. Da hadde vi en såkalt "gullstandard", og vi sammenlignet klassene til klassifisereren vår med de virkelige sanne klassene i gullstandarden.

- 1. Beskriv kort hvordan vi evaluerte maskinlæringsmodellen vår ved å bruke disse gullstandardklassene.
- 2. Å ha gullstandardannotasjoner for et dialogsystem er ikke alltid ønskelig. Om vi ønsker åha et system som kan kommunisere med mennesker, og som kan håndtere flere typersamtaler trenger vi et system som kan "kopiere" menneskelige samtaler uten å følge en mal. Kan du tenke deg noen utfordringer innen menneske-menneske kommunikasjon som er viktige å kunne håndtere og som gjør det vanskelig å forholde seg til en gullstandard?

Løsningsforslag

1. En maskinlæringsmodell evalueres ved bruk av accuracy, precision, recall, og F1. For å kunne regne ut disse målene, teller vi antall sanne positive (true positive), falske negative (false negative), falske positive (false positive), og sanne negative (true negative) disse kombineres i forskjellige måter og gir oss en oversikt over hva klassifisereren klarte å klassifisere rikitig, og hvilke feil den har gjort.

2. Menneske-menneske kommunikasjon er utfordrende fordi et system må forstå at under en samtale, har hver sin tur til å snakke. Av og til stiller vi spørsmål og forventer svar, andre ganger stiller vi spørsmål for å be den andre om å utføre noe. Vi gjentar det den andre sier for å vise at vi følger med eller at vi har forstått. Rekkefølgen i en dialog er ikke alltid konsekvent. Vi hopper av og til fra ett tema til et annet, og ombestemmer oss ofte. Mye av samtalene innebærer inferens og implikasjon.
