Tentamen 2017-01-11

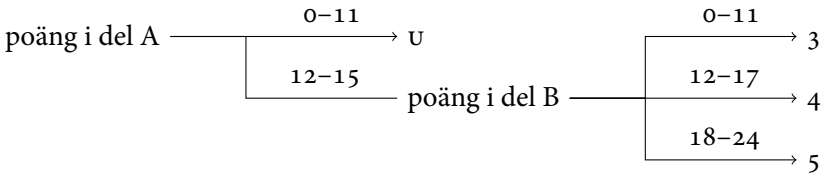
Examinator: Marco Kuhlmann

Denna tentamen består av två delar:

1. **Del A** består av 5 uppgifter som prövar din förståelse av de grundläggande begrepp och procedurer som behandlas på kursen. Dessa uppgifter kräver endast kompakta redogörelser, t.ex. en kort text, en uträkning eller ett diagram. Varje uppgift är värd 3 poäng, och för att bli godkänd på tentan behöver du ha minst 12 poäng totalt i denna del. Uppnår du detta mål avgörs ditt betyg av dina poäng i del B.

2. **Del B** består av 4 uppgifter: De tre första uppgifterna prövar din förståelse av de mera avancerade begrepp och procedurer som behandlas på kursen, samt hur olika metoder från kursen hänger ihop. Den sista uppgiften är en uppgift av essäkaraktär som prövar din förståelse av den artikel som behandlas på seminariet. Dessa uppgifter kräver utförliga och sammanhängande redogörelser med korrekt terminologi. Varje uppgift är värd 6 poäng.

Ditt betyg sätts enligt följande:



Lycka till!

För att korta rättningstiden rättar vi del B endast om du fått minst 12 poäng i del A. Vill du ha personlig återkoppling på uppgifter som vi inte rättat är du välkommen att kontakta examinator. En utförlig rättningsmall kommer att finnas för alla uppgifter.

Del A

01 **Textklassificering** **(3 poäng)**

Ett system för textklassificering baserat på metoden Naive Bayes ska avgöra om dokumentet ”Stockholm Oslo” är en nyhet om Sverige (klass `S`) eller en nyhet om Danmark (klass `D`).

a) För att predicera dokumentets klass använder systemet bl.a. sannolikheterna $P(\texttt{D})$ och $P(\text{Oslo} \mid \texttt{D})$. Lista de övriga sannolikheter som är relevanta.

b) Skatta de relevanta sannolikheterna med Maximum Likelihood-metoden utifrån följande dokumentsamling. Ställ upp bråk.

	dokument	klass
1	Stockholm Oslo	`S`
2	Köpenhamn Stockholm	`D`
3	Stockholm Köpenhamn	`S`
4	Köpenhamn Oslo	`D`

c) Utifrån de skattade sannolikheterna, vilken klass predicerar systemet? Redovisa hur du räknat. Visa tydligt att du förstått Naive Bayes-klassificeringsregeln.

02 **Ordpredicering** **(3 poäng)**

I en text bestående av 1 215 396 ord hittas ordet *det* 13 694 gånger, ordet *är* 13 700 gånger, ordet *våras* 2 gånger, bigrammet *det är* 927 gånger och bigrammet *det våras* 0 gånger.

a) Skatta unigramsannolikheten $P(det)$ och bigramsannolikheten $P(är \mid det)$ med Maximum Likelihood-metoden. Ställ upp bråk.

b) Vad händer när man skattar bigramsannolikheten $P(våras \mid det)$ med Maximum Likelihood-metoden? Varför kan detta vara ett problem för praktiska system? Ange en exempelmening som illustrerar problemet.

c) Skatta bigramsannolikheten $P(våras \mid det)$ med Maximum Likelihood-metoden och addera k-utjämning med $k = 0{,}01$. Antag att vokabulären definieras av Svenska Akademiens ordlista, som innehåller 126 000 ord. Ställ upp bråk.

03 Ordklasstaggning (3 poäng)

Vid utvärderingen av en ordklasstaggare fick man ut nedanstående förväxlingsmatris. Den markerade cellen anger antalet gånger systemet taggade ett ord som verb (`VB`) medan det enligt guldstandarden borde ha taggats som substantiv (`NN`).

	NN	JJ	VB
NN	58	6	1
JJ	5	11	2
VB	0	7	43

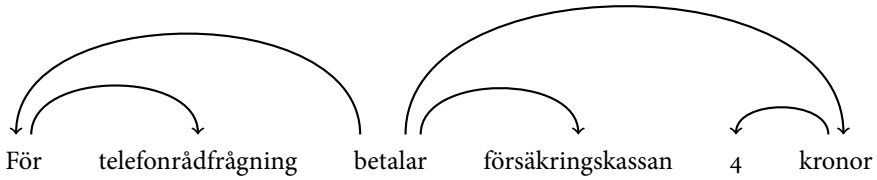
a) Ställ upp ett bråk för taggarens täckning (recall) på verb.

b) Ställ upp ett bråk för taggarens precision på substantiv.

c) Ange en annan förväxlingsmatris där taggarens korrekthet är samma som i matrisen ovan men där taggarens precision på adjektiv är 100%.

04 Syntaktisk analys (3 poäng)

En transitionsbaserad dependensparser ska parsa meningen *För telefonrådfrågning betalar försäkringskassan 4 kronor*. Här är guldstandardträdet för denna mening:



a) Parserns initiala konfiguration för meningen ser ut så här:

stack: [] **buffert:** [För, telefonrådfrågning, betalar, försäkringskassan, 4, kronor]

Ange den terminala konfigurationen för guldstandardträdet.

b) Efter två transitioner har parsern kommit till följande konfiguration:

stack: [För, telefonrådfrågning] **buffert:** [betalar, försäkringskassan, 4, kronor]

Ange de transitioner som parsern har att välja mellan i denna konfiguration. Markera den transition som parsern behöver ta för att återskapa alla bågar i guldstandardträdet.

c) Ange en transitionssekvens som tar parsern hela vägen från den initiala konfigurationen för meningen till den terminala konfigurationen, och som återskapar hela guldstandardträdet.

05 **Semantisk analys** **(3 poäng)**

Betrakta följande dokumentsamling:

(1) automobile wheel motor vehicle transport passenger

(2) car form transport wheel capacity carry five passenger

(3) transport London game spectator advise avoid use car

(4) London soccer tournament begin goal match

(5) Giggs score goal football tournament Wembley London

(6) Bellamy passenger football match play part goal

a) Fyll i följande matris med samförekomster. Varje cell ska innehålla antalet gånger målordet (rad) förekommer i samma dokument som kontextordet (kolumn).

	passenger	transport	goal	match
automobile	☐	☐	☐	☐
car	☐	☐	☐	☐
soccer	☐	☐	☐	☐
football	☐	☐	☐	☐

b) Rita målorden som vektorer i ett koordinatsystem där x-axeln svarar mot det totala antalet förekomster i kontexterna *passenger*, *transport* och y-axeln svarar mot det totala antalet förekomster i kontexterna *goal*, *match*.

c) Hur kan man med hjälp av sådana vektorrepresentationer mäta likhet mellan målorden? Vilka resultat skulle detta ge för de angivna målorden?

Del B

06 **Rättstavningskorrektur** **(6 poäng)**

När man googlar på ett felstavat ord som t.ex. *rätstavning* så tar det endast några bråkdelar av en sekund och Google frågar:

> Menade du: *rättstavning*

I den här uppgiften ska du fundera på hur denna teknik kan implementeras.

a) Förklara hur en enkel algoritm skulle kunna fungera som tar ett ord w och beräknar alla ord vars Levenshtein-avstånd till w är högst ett.

b) Hur skulle man kunna utöka denna algoritm så att den liksom Googles föreslår det mest sannolika rättstavade ordet i ett givet språk? Koppla till metoder du har lärt känna under kursen.

c) Hur skulle man kunna generalisera ansatsen till ord vars Levenshtein-avstånd till det felstavade ordet är större än ett? Vilket beräkningsmässigt problem uppstår?

07 **Informationsextraktion** **(6 poäng)**

Informationsextraktion är uppgiften att extrahera strukturerad information från textdokument. Begreppet *strukturerad information* syftar dels på namngivna entiteter och deras attribut, dels på semantiska relationer mellan dessa entiteter.

a) Förklara begreppen *namngivna entiteter* och *semantiska relationer*. Vilka typer av entiteter och relationer kan vara intressanta att extrahera? Ge exempel.

b) Förklara hur du skulle kunna angripa uppgiften att hitta namngivna entiteter i en text med hjälp av de metoder som vi lärt känna inom området ordklasstaggning. Vilka problem eller begränsningar ser du med detta?

c) Förklara hur du skulle kunna angripa uppgiften att hitta semantiska relationer mellan namngivna entiteter med hjälp av de metoder som vi lärt känna inom området syntaktisk analys. Vilka problem eller begränsningar ser du med detta?

08 **Viterbi-algoritmen** **(6 poäng)**

Följande matriser specificerar en Hidden Markov-modell. Istället för sannolikheter anges kostnader (negativa log-sannolikheter).

	PL	PN	PP	VB	EOS
BOS	11	2	3	4	19
PL	17	3	2	5	7
PN	5	4	3	1	8
PP	12	4	6	7	9
VB	3	2	3	3	7

	hen	vilar	ut
PL	17	17	4
PN	3	19	19
PP	19	19	3
VB	19	8	19

När man använder Viterbi-algoritmen för att beräkna den mest sannolika (minst kostsamma) taggsekvensen för meningen ”hen vilar ut” enligt denna modell får man ut följande matris. Notera att matrisen saknar tre värden (markerade celler).

		hen	vilar	ut
BOS	0			
PL		28	27	21
PN		A	28	35
PP		22	27	20
VB		23	B	36
EOS				C

a) Beräkna de saknade värdena. Redovisa hur du räknat.

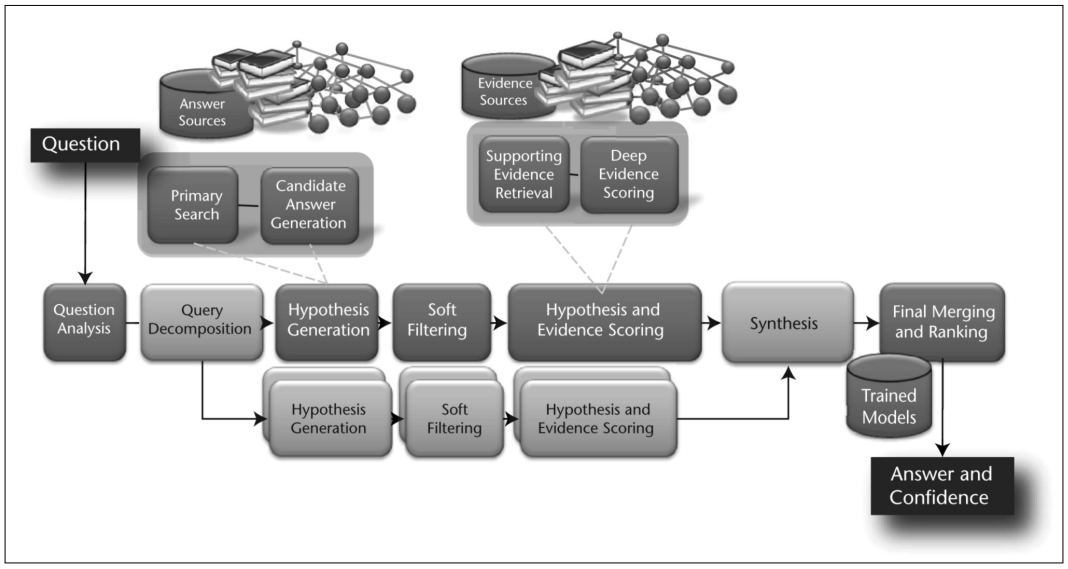
b) Ange den mest sannolika taggsekvensen för meningen. Förklara hur den kan fås från (den fullständiga) matrisen.

c) Förutom Viterbi-algoritmen har du även lärt känna en annan metod för ordklasstaggning. Förklara denna metod kortfattat. Resonera kring möjligheterna och begränsningarna som finns med de två metoderna.

09 **Frågebesvarande system** **(6 poäng)**

Denna uppgift syftar på följande artikel:

David Ferrucci, Eric Brown, Jennifer Chu-Carroll, James Fan, David Gondek, Aditya A. Kalyanpur, Adam Lally, J. William Murdock, Eric Nyberg, John Prager, Nico Schlaefer, Chris Welty. *Building Watson: An Overview of the DeepQA Project.* AI Magazine 31(3):59–79, 2010.



a) Med stöd i bilden ovan, beskriv DeepQA-arkitekturen översiktligt och redogör för ett av stegen mera ingående. Ge exempel från artikeln.

b) Beskriv de mått som använts för att bedöma Watsons prestation. Hur har utvecklarna gått tillväga för att optimera dessa mått? Resonera kring vilket det viktigaste måttet är.

c) Författarna skriver: ”Our results strongly suggest that DeepQA is an effective an extensible architecture that may be used as a foundation for combining, deploying, evaluating, and advancing a wide range of algorithmic techniques to rapidly advance the field of QA.” Diskutera vilka specifika resultat i artikeln som stödjer denna slutsats. Håller du med författarnas bedömning?