Tentamen 2018-01-13

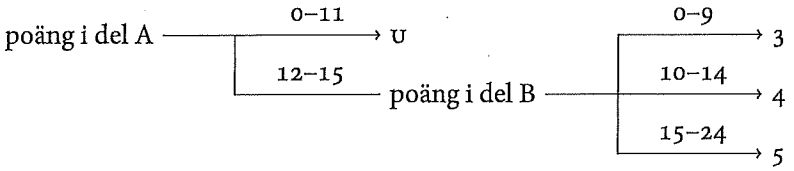
Examinator: Marco Kuhlmann

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 "London Paris" är en nyhet om Storbritannien (klass U) eller en nyhet om Spanien (klass S).

a) Ange klassificeringsregeln för Naive Bayes som en formel och förklara de olika delarna i den.

b) Skatta de sannolikheterna som är relevanta för detta beslut med Maximum Likelihood-metoden (utan utjämning) utifrån följande dokumentsamling. Svara med bråk.

	dokument	klass
1	London Paris	U
2	Madrid London	S
3	London Madrid	U
4	Madrid Paris	S

c) I praktiska implementationer av en Naive Bayes-klassificerare brukar man räkna om sannolikheter till logaritmer. Skissa grafen för funktionen $f(p) = \log p$ där p är ett sannolikhetsvärde. Förklara hur klassificeringsregeln behöver ändras när man använder logaritmiserade sannolikheter.

02 Ordpredicering (3 poäng)

Datamängden *Corpus of Contemporary American English* består av 520 miljoner token och innehåller 1 254 193 unika ord. Vi hittar följande frekvenser av unigram och bigram: *your*, 883 614; *rights*, 80 891; *doorposts*, 21; *your rights*, 378; *your doorposts*, 0.

a) Skatta sannolikheterna $P(\textit{rights})$ och $P(\textit{rights} \mid \textit{your})$ med hjälp av Maximum Likelihood-metoden (utan utjämning). Svara med bråk.

b) Skatta bigramsannolikheten $P(\textit{doorposts} \mid \textit{your})$ med hjälp av Maximum Likelihood-metoden med addera-k utjämning med $k = 0{,}01$. Svara med ett bråk.

c) Vi tränar tre n-gram-modeller på nyhetstexter (38 miljoner ord): en unigram-modell, en bigram-modell och en trigram-modell. Vi utvärderar dessa modeller på 1,5 miljoner ord med samma vokabulär och får ut följande entropivärden: 7,409; 6,768; 9,910. Vilket entropivärde hör till vilken modell? Förklara.

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

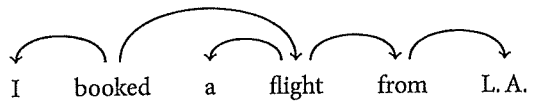
a) Ange taggarens korrekthet. Svara med ett bråk.

b) Ange taggarens täckning (recall) på adjektiv och dess precision på substantiv. Svara med bråk.

c) Ange en annan förväxlingsmatris där taggarens korrekthet är samma som i matrisen ovan men där taggarens täckning (recall) på verb är 100%.

04 Syntaktisk analys (3 poäng)

En transitionsbaserad dependensparser analyserar meningen *I booked a flight from L. A.* Här är guldstandardträdet för denna mening:



a) Antag att parsern börjar i den initiala konfigurationen för meningen och tar transitionerna SH, SH, RA. Ange den nya konfigurationen. Representera det partiella dependensträdet genom att lista de bågar som finns i det.

b) Ange en fullständig transitionssekvens som tar parsern hela vägen från den initiala konfigurationen till en terminal konfiguration, och som skapar alla bågar i guldstandardträdet.

c) För en mening med n ord, hur många transitioner gör parsern för att komma från den initiala konfigurationen till en terminal konfiguration? Förklara ditt resonemang.

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. 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)

Nedan visas den matris som Wagner–Fisher-algoritmen producerar när den beräknar Levenshtein-avståndet mellan orden *intention* och *execution*. Observera att matrisen saknar några värden (markerade celler).

n	A	8	8	8	8	8	8	7	6	5
o	A	7	7	7	7	7	7	6	5	6
i	A	6	6	6	6	6	6	5	6	7
t	A	5	5	5	5	5	5	6	7	8
n	A	4	4	4	4	5	6	7	7	7
e	A	3	4	B	4	5	6	6	7	8
t	A	3	3	3	4	5	6	6	7	8
n	A	2	2	3	4	5	6	7	7	7
i	A	1	2	3	4	5	6	6	7	8
#	A	A	A	A	A	A	A	A	A	A
	#	e	x	e	c	u	t	i	o	n

a) Definiera begreppet Levenshtein-avstånd. Din definition ska vara förståelig även för folk som inte tagit denna kurs.

b) Ange värdena för cellerna markerade med bokstaven A. Förklara.

c) Beräkna värdet för cellen markerad med bokstaven B. Förklara. Visa tydligt att du har förstått Wagner–Fisher-algoritmen.

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). Den markerade cellen ange övergångskostnaden från BOS till PL.

	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	5	B	35
PP	22	27	20
VB	A	14	36
EOS			C

a) Beräkna värdet för cellen A. Förklara din beräkning.

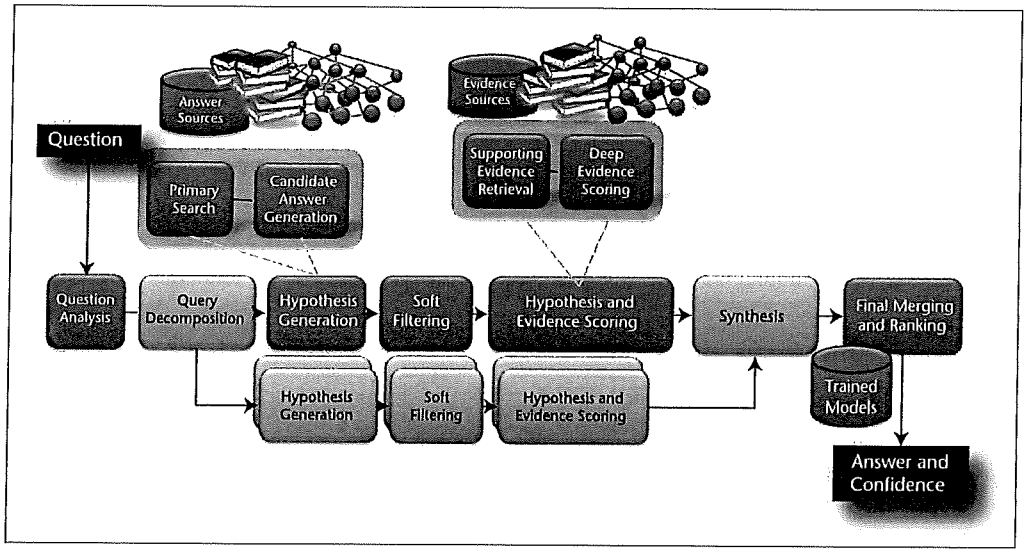
b) Beräkna värdena för cellerna B och C. Förklara dina beräkningar.

c) Rita in de "backpointers" som identifierar den mest sannolika sekvensen för meningen. Ange denna taggsekvens.

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 om författarnas bedömning?