COMP61332

Two hours

Formula Sheet attached for use with Question 4.

UNIVERSITY OF MANCHESTER
SCHOOL OF COMPUTER SCIENCE

Text Mining

Date: Tuesday 31st May 2016

Time: 14:00 - 16:00

Please answer any THREE Questions from the FIVE Questions provided

Each question is worth 20 marks

This is a CLOSED book examination

The use of electronic calculators is permitted provided they
are not programmable and do not store text

[PTO]

1.

a) Classify the following examples according to the type of ambiguity they display:

i) The dog walked down the road barked.

ii) He ordered the pizzas.

iii) Can you open the window?

iv) The boy kicked the ball between the goalposts.

(2 marks)

b) In relation to Brill's Transformation-based Learning (TBL) algorithm:

i) Explain how the lexicon, the transformation rules and the part of speech guesser are used to tag a text. In your answer, illustrate the format of a transformation rule, and indicate the format of the lexicon.

(3 marks)

ii) State why, when having used a corpus for training, you might reasonably expect only a small gain in performance on a test corpus, for *lexically-based* data.

(1 mark)

iii) Are there any benefits to using a TBL-trained tagger in contrast to one trained using other machine-learning approaches?

(2 marks)

c) i) What kind of resource is ideally needed to produce a probabilistic phrase structure grammar (PCFG) (1 mark) and how is it used to do this (1 mark)?

ii) Why is a lexicalised PCFG of more interest than a non-lexicalised PCFG?

(1 mark)

[Question 1 continues on the following page]

[Question 1 continues from the previous page]

d) Consider the following grammar and lexicon (numbers are given for reference only):

```
1) S -> NP VP          7) PRO -> many        13) V -> flooded
2) NP -> ADJ NNS       8) ADJ -> many        14) NNS -> rivers
3) NP -> PRO           9) NNS -> fish
4) NP -> NNS           10) N -> fish
5) VP -> V NP          11) V -> fish
6) VP -> V             12) ADJ -> flooded
```

i) Show, by constructing two parse trees, that the string "many fish flooded rivers" is ambiguous according to the above grammar.

(2 marks)

ii) Show the steps (rule numbers) that a naïve top-down (goal driven) depth-first parser would take in parsing the string "fish flooded rivers" according to the above grammar.

(2 marks)

iii) Jurafsky & Martin (2009) state, in relation to naïve, top-down (goal-driven) parsing and naïve, bottom-up (data-driven) parsing: "Neither of these approaches adequately exploits the constraints presented by the grammar and the input words". What evidence is there to support this view?

(2 marks)

iv) Assume that the Earley algorithm has been applied to parse the string "many rivers flooded" with the grammar and lexicon shown above. The parser has paused at the state shown below as a chart. Reproduce this chart, and add the *next* nine labelled edges (representing scanning, prediction or completion steps) that would be produced by the algorithm.

(3 marks)

[PTO]

2.

a) What is meant by BILOU notation, and why is it particularly relevant to machine learning methods in text mining?

(2 marks)

b) Annotate the following sentence, using BILOU notation, to show the boundaries of its (underlined) noun phrase chunks:

"In June, the United Kingdom will have a referendum on membership of the European Union", said Cameron.

(2 marks)

c) What are the relative merits of in-line and stand-off annotations in representing the input and output of text mining components?

(3 marks)

d) You are asked to write a tokeniser for a text mining system. Discuss tokenisation issues you may expect to arise for this task, with appropriate examples, explaining the impact that certain tokenisation decisions may have on later processing components.

(4 marks)

e) In relation to supervised machine-learning based named entity recognition:

i) State four features that would be useful in ***detecting*** named entities in general news text, i.e., in answering the question "are there named entities present?".

(1 mark)

ii) State four features that would be useful in ***classifying*** named entities in general news text, i.e., in answering the question "what named entities are present?".

(1 mark)

iii) Discuss advantages and disadvantages of dictionary-based, rule-based and machine learning approaches to named entity recognition.

(4 marks)

[Question 2 continues on the following page]

[Question 2 continues from the previous page]

f) Consider the following processing components:

(1) Reference evaluator: Reports annotation effectiveness comparing two inputs of which one is indicated to be a reference (gold data) input. The report is saved to a file and includes common performance metrics.
(2) Annotation remover
(3) Gold standard corpus reader
(4) Syntactic parser
(5) Part-of-speech tagger
(6) Named entity recogniser
(7) Sentence splitter
(8) Tokeniser
(9) Chunker

Assume that you have just created components (5)..(9).
Design a workflow, by drawing a diagram, which would allow you to evaluate the combined effectiveness of your components against a given gold standard corpus. You do **not** need to refer to any specific UIMA-based type systems.

(3 marks)

[PTO]

3.

a) Give a dependency representation for the following sentence, for the interpretation that the eating was done with a fork:

I ate the fish with a fork.

For your analysis, you must select from the following set of dependency relations: `SUBJ` (subject), `DOBJ` (direct object), `IOBJ` (indirect object), `NMOD` (noun modifier), `PMOD` (prepositional modifier), `POBJ` (prepositional object), `DETMOD` (determiner modifier), `PU` (punctuation), `RT` (root).

(2 marks)

b) Consider the following sentences S1..S3:

S1: In 2010, Prime Minister David Cameron appointed May to become Head of the Home Office.

S2: Many English Football Supporter Clubs will flock to see this year's FA Cup Final which will be held in May at London's iconic Wembley Stadium.

S3: Since 1956, the Isle of May has been managed by Scottish Natural Heritage as a National Nature Reserve.

i) List all instances of the following named entities in each sentence S1..S3: Person, Location, TemporalExpression, Organisation, NamedOccasion.

(1 mark)

ii) A named entity recogniser for the above four entity types sometimes labels sequences that contain both false positives and false negatives when there is an exact matching requirement. What sequences in the above sentences S1..S3 might give rise to such errors, and why?

(1 mark)

iii) Annotate the events in the three sentences S1..S3 using the three templates below. Put "N/A" if a role does not have a value. Do **not** reproduce the template explanations or role definitions.

START-POSITION: occurs whenever a person begins working for (or changes offices within) an organisation. This includes government officials starting their terms, whether elected or appointed.

Trigger	the word signifying the event	
Person-Arg	the employee	
Entity-Arg	the employer	
Position-Arg	the job title	
Time-Arg	when the employment begins	
Place-Arg	where the employment begins	

[Question 3 continues on the following page]

[Question 3 continues from the previous page]

TRANSFER-OWNERSHIP: refers to the buying, selling, loaning, borrowing, giving, or receiving of artefacts or organisations.

Trigger	the word signifying the event	
Buyer-Arg	the receiver	
Artefact-Arg	the item that was received	
Price-Arg	the sale price	
Time-Arg	when the receiving takes place	
Place-Arg	where the receiving takes place	

MEET: occurs whenever two or more entities come together at a single location and interact with one another face-to-face. This includes talks, summits, conferences, meetings, visits, and any other gatherings where two or more parties get together at some location.

Trigger	the word signifying the event	
Entity-Arg	who are gathering	
Time-Arg	when the gathering takes place	
Place-Arg	where the gathering takes place	
Occasion-Arg	The named occasion or reason for the gathering	

(6 marks)

c) Assume that the three sentences S1..S3 of **3b above** came from three different documents that are searchable by some search engine.

i) Comment on what the search results will typically look like if the documents are not indexed using any annotations (i.e., only exact string matching is used), and a user inputs as a query "May".

(1 mark)

ii) Assume we now index the documents with part-of-speech (POS) tags for each token and re-run the same query. Would this make a difference to search results? Say why, or why not.

(1 mark)

iii) Assume we have a user who is interested in the FA Cup this year. Explain how a semantic search engine could make use of the semantic annotations you produced in **3b i and 3b iii above**, once appropriately indexed, in order to return information relevant to this user's interest. (4 marks) What would sample queries look like now? (2 marks)

iv) Why might you need to persuade a user to adopt a semantic search engine of the type you described in **3 c iii** and what arguments would you use to convince him? (2 marks)

[PTO]

4.

a) Consider the following two sentences:

> The boy ran happily along the road holding his coat.
> The girl walked sadly down the path carrying her coat button.

Explain how, by accessing lexical relations in the Princeton WordNet, an analyser could determine that these two sentences involved similar events with similar participants and closely related interpretations.

(2 marks)

b)

i) Explain (informally) how Lesk's algorithm for word sense disambiguation (WSD) would help to resolve the correct sense for *bank* in the sentence *The waves hit the steep bank*. In your answer, refer to the following data which are available for this task:

Senses for bank:

bank_1: sloping land (especially the slope beside a body of water).

bank_2: a financial institution that accepts deposits and moves the money into lending activities.

bank_3: a building in which the business of banking is transacted.

Context definitions:

wave: one of a series of ridges that moves across the surface of a liquid (especially across a large body of water).

hit: hit against; come into sudden contact with.

steep: of a slope; set at a high angle.

(3 marks)

ii) Would you choose Lesk's algorithm to do WSD? Say why, or why not.

(1 mark)

c)

i) Discuss advantages and disadvantages of the count-based, context-predicting and compositional distributional semantic models.

(2 marks)

ii) Once we have compiled our context vectors, how can we then use these to help us construct a thesaurus?

(2 marks)

d)

i) What is the advantage of X^2 *(chi-square) test* over *t test*?

(1 mark)

[Question 4 continues on the following page]

[Question 4 continues from the previous page]

ii) What is the advantage of likelihood ratios over X^2 *(chi-square) test*?

(1 mark)

e) ***For this question part, consult the provided formula sheet***.

A newswire corpus consists of 10000 bigrams, where *black* occurs in 200 bigrams and *maria* occurs in 150 bigrams. The bigram *black maria* occurs 5 times.

i) State the *t test* null hypothesis and decide if the co-occurrence of *black* and *maria* is random or not using the *t test*. Show your working. (The critical value for a confidence level $\alpha = 0.005$ is 2.576.)

(3 marks)

ii) Compute the observed values contingency table of the X^2 *(chi-square) test*.

(1 mark)

iii) Compute the expected values contingency table of the X^2 *(chi-square) test*.

(1 mark)

iv) Decide if the co-occurrence of *black* and *maria* is random or not using the X^2 *(chi-square) test*. Show your working. (For 1 degree of freedom and at a probability level of $\alpha = 0.05$ the critical value is 3.841.)

(3 marks)

[PTO]

5.

a) Two annotators annotate a corpus for PERSON entity instances. We calculate how many times they agree/disagree, and obtain the following table:

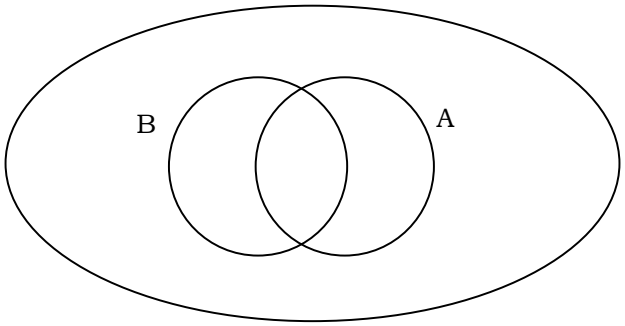
		Annotator 1 Results		
		yes	no	total
Annotator 2 Results	yes	300	20	320
	no	10	70	80
	total	310	90	400

Using the above values and the kappa statistic, calculate the inter-annotator agreement rate. Show your working.

(3 marks)

Hint: The formula for the kappa statistic is: $(P(a) – P(e))/(1 – P(e))$, where $P(a)$ is the observed proportion of times the annotators agreed, and $P(e)$ is the proportion of times they would be expected to agree by chance.

b) Consider the following partially completed diagram:



The outer ellipse represents a text corpus. Circle A represents the gold standard annotations for the entity `DISEASE` in this corpus. Circle B represents the result of a named entity recogniser attempting to annotate the corpus for `DISEASE`.
Reproduce this diagram and clearly label appropriate parts to indicate:

- True positives (TP)
- False positives (FP)
- True negatives (TN)
- False negatives (FN)

(2 marks)

[Question 5 continues on the following page]

[Question 5 continues from the previous page]

c) In text mining, why do we usually prefer to use *F measure* when scoring the results of a system against a gold standard annotated corpus?

(1 mark)

d)

i) Show, with examples, how the lexicon, syntax, rhetoric and pragmatics are all involved in explaining how subjective sentiment is conveyed in natural language.

(4 marks)

ii) Bing Liu considers that an opinion is represented by a quintuple (5-tuple). What does he mean by this, and what technique or techniques would be required to capture information to fill the quintuple representation he proposes?

(5 marks)

e) Briefly set out what you consider to be the major current challenges for text mining, and then discuss what if any progress we may expect towards meeting these challenges in the near to medium future. Justify your views and conclusions, giving appropriate examples to back up your arguments.

(5 marks)

END OF EXAMINATION

COMP61332 Examination: Formula Sheet

Pointwise Mutual Information (PMI)

$$I(x,y) = log_2 \frac{P(x,y)}{P(x)\,P(y)} = log_2 \frac{P(x|y)}{P(x)} = log_2 \frac{P(y|x)}{P(y)} \tag{1}$$

where x and y are events.

T statistic

$$t = \frac{\overline{x} - \mu}{\sqrt{\frac{s^2}{N}}} \tag{2}$$

where $\overline{x}$ is the sample mean, s^2 the sample variance, N the sample size, and μ the mean of the distribution.

$\nu \backslash \alpha$	0.05	0.025	0.01	0.005	0.0025	0.001	0.0005
1	6.314	12.71	31.82	63.66	127.3	318.3	636.6
2	2.920	4.303	6.965	9.925	14.09	22.33	31.60
3	2.353	3.182	4.541	5.841	7.453	10.21	12.92
4	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	1.812	2.228	2.764	3.169	3.581	4.144	4.587
20	1.725	2.086	2.528	2.845	3.153	3.552	3.850
30	1.697	2.042	2.457	2.750	3.030	3.385	3.646
40	1.684	2.021	2.423	2.704	2.971	3.307	3.551
50	1.676	2.009	2.403	2.678	2.937	3.261	3.496
60	1.671	2.000	2.390	2.660	2.915	3.232	3.460
80	1.664	1.990	2.374	2.639	2.887	3.195	3.416
100	1.660	1.984	2.364	2.626	2.871	3.174	3.390
120	1.658	1.980	2.358	2.617	2.860	3.160	3.373
∞	1.645	1.960	2.326	2.576	2.807	3.090	3.291

Table 1: t distribution table (ν: degrees of freedom)

Continued on next page

X^2 statistic

$$X^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{3}$$

where O denotes the contingency table of observed values and E the contingency table of expected values.

$\nu \backslash \alpha$	0.250	0.100	0.050	0.025	0.010	0.005
1	1.32330	2.70554	3.84146	5.02389	6.63490	7.87944
2	2.77259	4.60517	5.99146	7.37776	9.21034	10.59663
3	4.10834	6.25139	7.81473	9.34840	11.34487	12.83816
4	5.38527	7.77944	9.48773	11.14329	13.27670	14.86026
5	6.62568	9.23636	11.07050	12.83250	15.08627	16.74960
6	7.84080	10.64464	12.59159	14.44938	16.81189	18.54758
7	9.03715	12.01704	14.06714	16.01276	18.47531	20.27774
8	10.21885	13.36157	15.50731	17.53455	20.09024	21.95495
9	11.38875	14.68366	16.91898	19.02277	21.66599	23.58935
10	12.54886	15.98718	18.30704	20.48318	23.20925	25.18818
12	14.84540	18.54935	21.02607	23.33666	26.21697	28.29952
14	17.11693	21.06414	23.68479	26.11895	29.14124	31.31935
16	19.36886	23.54183	26.29623	28.84535	31.99993	34.26719
18	21.60489	25.98942	28.86930	31.52638	34.80531	37.15645
20	23.82769	28.41198	31.41043	34.16961	37.56623	39.99685
22	26.03927	30.81328	33.92444	36.78071	40.28936	42.79565
24	28.24115	33.19624	36.41503	39.36408	42.97982	45.55851
26	30.43457	35.56317	38.88514	41.92317	45.64168	48.28988
28	32.62049	37.91592	41.33714	44.46079	48.27824	50.99338
30	34.79974	40.25602	43.77297	46.97924	50.89218	53.67196

Table 2: χ^2 distribution table (ν: degrees of freedom)