COMP 61332

Two hours

UNIVERSITY OF MANCHESTER
SCHOOL OF COMPUTER SCIENCE

Text Mining

Date: Tuesday 22nd May 2018

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 under the tree with the bone is happy.

ii) The window is too small.

iii) This is a sentence that has no adjectives.

iv) The fish is ready to eat.

(2 marks)

b)

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

"In March 2019, the UK will leave the European Union", said Juncker last week.

(2 marks)

ii) Why is BILOU notation often preferred to BIO notation? (1 mark) Under what circumstances might you prefer BIO notation to BILOU notation? (1 mark)

(2 marks)

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

(3 marks)

c) 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 (POS) tagger
(6) Machine-learning-based named entity recogniser (NER)
(7) Tokeniser
(8) Sentence splitter
(9) Rule-based named entity recogniser
(10) Dictionary-based named entity recogniser

[Question 1 continues on the following page]

[Question 1 continues from the previous page]

Assume that you have just created components (5)–(10).
Assume a UIMA-based environment.
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, for a named entity recognition task. You do **not** need to refer to any specific UIMA-based type systems.

(4 marks)

d) Explain the role of the following in UIMA:

- Annotations
- Type System
- Common Analysis Structure (CAS)

(3 marks)

e) You are given the job of recommending a text mining architecture for your large multinational company which has interests in many kinds of product in many kinds of domain for many different types of customer, and which is rapidly expanding its activities. On your desk are advertisements from several companies that provide off-the-shelf text mining tools for specific tasks. In your inbox is a mail from the Chief Software Engineer saying "Don't worry about complex architectures, my team can write a text mining program that will do everything you want.". A colleague texts you to ask "Should we use this UIMA stuff?".

What recommendations would you make for a text mining architecture? Justify your decisions.

(4 marks)

[PTO]

2.

a) 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)

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

i) How are the lexicon and the rules induced from a corpus of tagged text?

(4 marks)

ii) Order the following major components of the TBL method in terms of the amount of their contribution to performance:

- Lexical rules
- Initial state
- Non-lexical rules

Explain and justify your ordering.

(1 mark)

c) Consider the following sentence:

The girl watched the man with the binoculars.

The sentence is ambiguous and has two possible interpretations: that the girl is using the binoculars, and that the man has the binoculars.

i) Draw the dependency graph for each of the two possible interpretations according to the notation of your choice but using the following label set: `det` (determiner), `nmod` (nominal modifier), `obj` (object), `pobj` (object of the preposition), `punct` (punctuation), `sbj` (subject), `vmod` (verb modifier).

(4 marks)

ii) Draw the phrase structure tree for each of the two possible interpretations, using the following label set: `DT` (determiner), `IN` (preposition), `JJ` (adjective), `NN` (singular noun), `NNS` (plural noun), `RB` (adverb), `VB` (verb), `AdjP` (adjectival phrase), `AdvP` (adverbial phrase), `NP` (noun phrase), `PP` (prepositional phrase), `VP` (verb phrase) and `S` (sentence).

(4 marks)

iii) For each of the two possible interpretations, list the predicate-argument structures for *watched* and *with*. Referring to the predicate-argument structures that you listed, comment on the underlying reason for the ambiguity of the sentence.

(3 marks)

3.

a) For years now, an American man called John Lewis (person) has been confused with the retail shop John Lewis (organisation). Consider the following lines of text from news headlines and articles written about each of them.

- Cold weather clothing sales leaped by 22% at John Lewis last week.
- John Lewis Partnership names new group procurement director.
- The John Lewis advert was very popular.
- A Twitter user named John Lewis is bombarded with misdirected tweets every year.
- John Lewis said that he does not mind the unexpected attention.
- Mr John Lewis is from Blacksburg, Virginia.

If you are going to develop a machine learning-based named entity recogniser (NER), what features would you include to help the NER distinguish between these two entities? Why do think these will work? Illustrate with examples.

(4 marks)

b) Consider the following sentences S1–S3:

S1: The Faculty of Science and Engineering will run the Research Data Management Workshop in summer, to be held in the Renold Building.

S2: TweetHarvester ran on my Linux machine for a month.

S3: Trump ran against Clinton for US President in 2016.

i) Annotate the events in the sentences using the templates below. Put "N/A" if a role does not have any value. Do **not** reproduce the template explanations or the role definitions in column 2.

(6 marks)

[Question 3 continues on the following page]

[Question 3 continues from the previous page]

Event type: `PROCESS_EXECUTION` (execution of a process or program on a computer)

`Trigger`	the word signifying the event	
`Process`	the program that is executed	
`Duration`	how long the execution is	
`Platform-or-System`	the platform on which the program has been executed	
`Start-time`	when the execution started	
`Completion-time`	when the execution finished	

Event type: `POLITICAL_COMPETITION` (a political race, e.g., an election)

`Trigger`	the word signifying the event	
`Candidate1`	a candidate or contender	
`Candidate2`	another candidate or contender	
`Time-period`	when the competition took place	
`Position`	the desired political position	

Event type: `OCCASION_HOSTING` (hosting or organising an occasion or affair)

`Trigger`	the word signifying the event	
`Occasion`	the occasion or affair	
`Host`	the organisation hosting the occasion	
`Venue`	where the occasion is held	
`Schedule`	when the occasion is held	
`Frequency`	how often the occasion is held	

[Question 3 continues on the following page]

[Question 3 continues from the previous page]

ii) The verb *run* has different meanings across sentences S1–S3. Imagine that you are about to develop a machine learning-based event trigger detector that should be able to assign the correct event type to each instance of *run*. What features will you use? Explain why, by giving examples. Keep in mind that, apart from the sentences, you also have named entities as input.

(4 marks)

iii) Discuss how text mining helps in the development of search engines that use semantic facets.

(2 marks)

c) A collection of N=100000 documents has been indexed. Four of the index terms and their postings list sizes are as follows:

Index term	Posting list size
bread	45000
flour	12000
butter	32000
toast	90000

A user gives the Boolean query:

```
bread AND (NOT flour) AND butter AND (NOT toast)
```

Recommend a query processing order for this query. Justify your recommendation.

(2 marks)

d) I calculate how many times each term appears in each document in a collection, to yield $tf_{t,d}$. I then calculate a tf-idf weight for each term in each document. I note that, for some term, its weight is less for document *A* than for document *B*.

What is this telling me? (1 mark)
How would I use this information to rank documents for relevance in response to a query? (1 mark)

[PTO]

4.

a) Consider the following two sentences:

The young woman talked about planting rose cuttings.
The old man shouted about growing flowers.

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

(3 marks)

b)

i) Explain what would be the result of applying Lesk's algorithm to find the correct sense for *table* in the phrase *The row of water bottles on the table.* In your answer, refer to the following data. Assume that stemming has taken place and stop words have been removed.

(4 marks)

Senses for table:

table_1: piece of furniture having a smooth flat top that is usually supported by one or more vertical legs.

table_2: set of data arranged in rows and columns.

table_3: underground surface below which the ground is wholly saturated with water.

Context definitions:

row_1: arrangement of objects or people side by side in a line.

row_2: linear array of data side by side, as in a table.

water_1: compound that occurs at room temperature as a clear colourless odourless tasteless liquid.

water_2: part of the earth's surface covered with water (as a river, lake or ocean).

bottle: glass or plastic vessel used for storing water or other liquids.

ii) Lesk's algorithm has proven to be very popular over the years. Discuss reasons for its popularity, and state your position on using it, with justifications.

(2 marks)

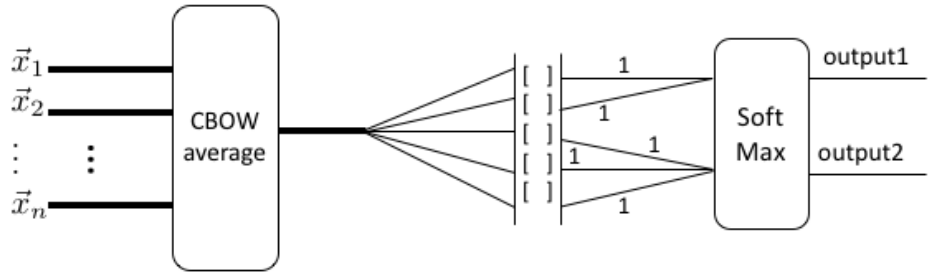
c) We observe that a phrase such as “frying pan” cannot appear more frequently than its constituent words (“frying” or “pan”). How can we use this observation to represent meaning and what are the advantages and disadvantages of the associated approaches compared to other distributional semantics techniques?

(5 marks)

[Question 4 continues on the following page]

[Question 4 continues from the previous page]

d) The following network operates on sentences (sequences of tokens). The weights of the linear layer are all 1 or zero, only the non-zero weights are shown. CBOW stands for Continuous Bag of Words.



There are 5 possible tokens t_1, t_2, t_3, t_4 and t_5 each with a one-hot token embedding vector:

$$\vec{t}_1 = \begin{bmatrix}1\\0\\0\\0\\0\end{bmatrix} \quad \vec{t}_2 = \begin{bmatrix}0\\1\\0\\0\\0\end{bmatrix} \quad \vec{t}_3 = \begin{bmatrix}0\\0\\1\\0\\0\end{bmatrix} \quad \vec{t}_4 = \begin{bmatrix}0\\0\\0\\1\\0\end{bmatrix} \quad \vec{t}_5 = \begin{bmatrix}0\\0\\0\\0\\1\end{bmatrix}$$

For example, with the sentence "t_2 t_3 t_1" *n*=3, the inputs to the CBOW layer are

$$\vec{x}_1 = \begin{bmatrix}0\\1\\0\\0\\0\end{bmatrix} \quad \vec{x}_2 = \begin{bmatrix}0\\0\\1\\0\\0\end{bmatrix} \quad \vec{x}_3 = \begin{bmatrix}1\\0\\0\\0\\0\end{bmatrix}$$

i) Given the network and the token embeddings above, for each of the following sentences state which output of the SoftMax layer will be higher or, if the input values are equal, state that there will be a tie. (Hint: the highest output value will correspond to the highest input value, *there is no need* to compute the formulae for the SoftMax equation.)

Sentence	Higher: output1, output2? Or tie?
t_2 t_3 t_1	
t_2 t_2 t_4	
t_3 t_4 t_1	
t_2 t_1 t_5 t_3	

(2 marks)

[Question 4 continues on the following page]

[Question 4 continues from the previous page]

ii) Could this network be used to detect if an arbitrary-length sentence has either token t_1 or t_2 immediately preceding t_3, t_4 or t_5? If yes, explain how. If not, what is the simplest sort of neural network that could be used to detect this and what other layers, if any, would be required?

(4 marks)

5.

a) Imagine that for a text mining project, you have been allowed to use ready-made tools that can recognise names of persons and locations. You found two different named entity recognisers (NER-A and NER-B) and after running each of them on some text with gold standard named entity annotations, you obtained the following number of true positives (TPs), false positives (FPs) and false negatives (FNs) for each entity type:

For NER-A:

Entity Type	TPs	FPs	FNs
Person	53	9	76
Location	32	15	10

For NER-B:

Entity Type	TPs	FPs	FNs
Person	70	15	59
Location	40	8	2

i) Your supervisor asked you to choose only one NER to use in your project. He said that he wants a tool that will produce minimal noise in terms of person names (i.e., he does not want to see tokens recognised as person names when in fact they are not). Which of the two NERs will you choose and what led you to that decision?

(2 marks)

ii) Calculate the performance of NER-B for each entity type in terms of precision, recall and F1-score.

(3 marks)

iii) Based on the results of NER-B, calculate the value of: (i) micro-averaged and (ii) macro-averaged precision and recall for the Person and Location types.

(2 marks)

iv) Which of the micro-averaged or macro-averaged scores give a better indication of overall performance and why? Make use of your answer in the previous item (iii) to explain and provide examples.

(3 marks)

[Question 5 continues on the following page]

[Question 5 continues from the previous page]

b) Liu notes of aspect-based sentiment analysis systems that "almost all real-life sentiment analysis systems in industry are based on this [fine-grained] level of analysis" (Liu, 2015).

Briefly comment on differences between conventional approaches to sentiment analysis and the aspect-based approach, assess what progress has been made towards achieving usable results with the aspect-based approach, and comment on any remaining challenges for this approach that you identify.

(5 marks)

c) "Many researchers have developed systems that can be adapted by other text mining specialists, but applications that can be tuned by [end users] are mostly lacking." (Gonzalez et al., 2016)

Discuss to what extent you agree with the above claim, referring to both the current state-of-the-art and to your view of the near- and medium-term future of the field. Justify your arguments.

(5 marks)

END OF EXAMINATION