Two hours

Formula Sheet attached for use with Question 4.

UNIVERSITY OF MANCHESTER
SCHOOL OF COMPUTER SCIENCE

Text Mining

Date:  Friday 17th May 2013
Time:  09:45 - 11:45

Please answer any THREE Questions from the FIVE Questions provided

Each question is worth 20 marks

Clearly cross out anything you do not wish to be marked.

This is a CLOSED book examination

The use of electronic calculators is permitted provided they are not programmable and do not store text.
1. 
   a) Classify the following examples according to the type of ambiguity they display:
      
      i) Sister accidentally shoots dead brother.
      ii) Stolen painting found by tree.
      iii) Police help dog bite victim.
      iv) The boy chased the dog with a bone.  

      (2 marks)

   b) 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 stages.

      (3 marks)

   c) How might publishers’ markup be used in determining what regions of a scientific article should be subject to text mining analysis and how types of text region should be treated?

      (3 marks)

   d) Briefly explain the distinction between in-line and stand-off markup, and discuss which is the better output format for text analysis.

      (3 marks)

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

      (2 marks)

   f) Annotate the following sentence to show the boundaries of its (underlined) noun phrase chunks, using the BILOU notation:
      
      Investigation into allegations of horsemeat mislabelling will be “relentless”, the Food Standards Agency says.

      (2 marks)

   g) State the sequence in which you would expect to apply the following components, explaining your choice, and noting any inherent ordering constraints that may apply:

      i) chunker
      ii) part of speech tagger
      iii) sentence splitter
      iv) gazetteer/dictionary lookup
      v) tokenizer
      vi) PDF to text converter.

      (2 marks)
h) Explain why the following linguistic notions are of importance for, and how they can be used in, natural language processing:

- Head of a construction
- Subcategorisation.

(3 marks)
2.

a) 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, using Brill’s algorithm. In your answer, illustrate the format of a rule, and indicate the contents of the lexicon. (4 marks)

ii) How are the lexicon and the rules induced from a corpus of tagged text? (4 marks)

iii) Discuss to what extent there are any benefits to using a TBL-trained tagger in contrast to one trained using other machine-learning approaches. (2 marks)

b) Consider the following newswire text:

Aberdeen, Scotland, 21st February, 2013. /GlobalOilNews Daily/
Oil Rig Catches Fire In North Sea
By T. Henry Stoneturner
Ms. Senga MacBeth, a Dansk-ScotOil spokesperson, announced today at 07:00 GMT that its semi-submersible drilling rig, Forager-23, had caught fire at 20:15 GMT yesterday, some 200 km NE of Aberdeen. A 10-man fire and rescue team led by Capt. Magnus Thorhammer has arrived by Wave Hopper helicopter to tackle the blaze. Ms. MacBeth said that two Dansk-ScotOil employees, Olivier d’Harcourt-Bréville, Chief Geological Scientist and Martin “Muddy” Waters, the well-known drilling mud viscosity expert, had lost their lives in the initial explosion. According to MacBeth, a NorthSeaLines ferry, the Island Princess has taken 36 survivors on board and was last reported to be 80 km from Aberdeen Harbour. The share price of Dansk-ScotOil (LSE:DKSO) fell by 3% in early trading on the London Stock Exchange. A former Dansk-ScotOil employee, Max Grumpie, claimed the fire was caused by defects in the EzyFlowMaster device installed last year. Both Dansk-ScotOil and EzyFlow (UK) refused to comment. The Offshore Health and Safety Agency has launched an inquiry into the incident.

Instances of the following named entities are to be identified:

PERSON, LOCATION, DATE, TIME, ARTEFACT, TRANSPORT, COMPANY, ORGANISATION, DISTANCE, LENGTH_UNIT, PERSON>Title, JOB_TITLE

i) For a rule-based analyser, what entries would you add to a gazetteer (dictionary) to aid recognition of the desired named entities? Assume that a gazetteer entry specifies for a lookup token its named entity type and that this named entity type may or may not be different to those mentioned above. (1 mark)

[Question 2 continues on the following page]
ii) For a rule-based analyser, what patterns (including contextual clues) would you use to help you write rules to identify the maximum number of instances of the above named entities in the text? For each pattern you specify, state which instances it would match. You may specify patterns informally (e.g., one or more capitalised tokens + \{“city”, “river”\} = LOCATION). If you find it useful to introduce other named entity types to aid your analysis, do so. Note any problematic aspects of the text that may cause your patterns to recognise too much, too little, or nothing, in certain cases.

(5 marks)

c) Concerning machine-learning based named entity recognition:

i) State 4 features that would be useful for named entity detection in general news text.

(2 marks)

ii) State 4 features that would be useful for named entity classification in general news text.

(2 marks)
3.

a) Suggest an application where chunking would be the more appropriate technique over full parsing. 

   (1 mark)

b) Compare and contrast a naïve, top-down (goal-driven) parser and a naïve, bottom-up (data-driven) parser in terms of their advantages and disadvantages.

   (2 marks)

c) Compare and contrast use of a context-free grammar (CFG) and a probability context-free grammar (PCFG) in relation to sentence parsing.

   (2 marks)

d) Briefly explain how a PCFG can be constructed.

   (1 mark)

e) Explain the difference between a lexicalized PCFG and an unlexicalized PCFG?

   (1 mark)

f) Compare and contrast a phrase structure tree (parse tree) and a dependency tree, discussing why the latter is often preferred for applications in text mining.

   (2 marks)

g) Briefly explain the notion of feature structure, then discuss how the introduction of feature structures can help overcome shortcomings of pure context free grammars, taking the notion of agreement as a basis for discussion.

   (2 marks)

h) Briefly explain feature structure unification, giving one example showing unification success and one example showing failure of unification.

   (2 marks)

Answer EITHER (i) OR (j). Do not answer both.

EITHER

i) Consider the following grammar and lexicon:

   \[
   \begin{align*}
   S & \rightarrow NP \ VP \\
   NP & \rightarrow N \\
   VP & \rightarrow V \ NP \\
   V & \rightarrow \text{cuts} \\
   N & \rightarrow \text{cuts} \\
   N & \rightarrow \text{surprise} \\
   N & \rightarrow \text{government} \\
   N & \rightarrow \text{party}
   \end{align*}
   \]

   [Question 3 continues on the following page]
i) Show, by constructing two parse trees, that the string “government cuts surprise party” is ambiguous according to the above grammar.

(2 marks)

ii) Similarly, show that the string “government cuts surprise party” is ambiguous, by constructing two dependency graphs using the following dependency labels: subj (subject), obj (object) and nmod (noun modifier).

(2 marks)

iii) Assume that the Earley algorithm has been applied to parse the string “government cuts surprise party” with the grammar and lexicon shown above. The parser has paused at the state shown below as a chart. Reproduce this chart, showing the next six edges (representing scanning, prediction or completion steps) that would be produced by the algorithm.

(3 marks)

OR

j)

i) Identify the event(s) occurring in the following sentence, indicating for every event that you identify: Event_ID, Trigger, Event_Type, Theme and Cause.

We found that Y activates the expression of X.

(2 marks)

ii) The EventMine system involves a number of steps to extract events. Based on the following sentence, show step by step the event extraction process of EventMine, noting the features that would be used at each step:

The analysis showed IEXC295S was unable to significantly transactivate the c-sis/PDFG-B promoter.

(5 marks)
4. 

a) 

i) When two different distributions are compared using the \( t \) test, what assumption must be true for this test to be valid? 

(1 mark) 

ii) What is the advantage of \( t \) test over frequency based methods? 

(1 mark) 

iii) What is the advantage of \( \chi^2 \) (chi-square) test over \( t \) test? 

(1 mark) 

b) 

i) What is a likelihood ratio? 

(1 mark) 

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

(1 mark) 

iii) State the likelihood ratio \textit{dependence and independence hypotheses} when deciding upon the significance of a collocation. 

(2 marks) 

c) Which hypothesis test would you suggest for sparse data? 

(2 marks) 

d) \textit{For this question part, consult the provided formula sheet.} 

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

i) Compute the pointwise mutual information of the bigram \textit{black maria}. 

(3 marks) 

ii) What is the interpretation of pointwise mutual information? 

(1 mark) 

iii) State the \( t \) test null hypothesis and decide if the co-occurrence of \textit{black} and \textit{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)
[Question 4 continues from the previous page]

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

(1 mark)

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

(1 mark)

vi) 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.)  

(2 marks)
5.
a)  
   i) What text elements may convey subjective attitudes as understood in opinion mining or sentiment analysis? Give examples for the types of text element you identify.  
      (3 marks)

   ii) Describe two or more approaches to the construction of a subjectivity lexicon and discuss their strengths and weaknesses.  
      (4 marks)

   iii) To what extent can sentiment analysis be approached by the same methods as named entity recognition?  
      (3 marks)

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

<table>
<thead>
<tr>
<th>Annotator 1 Results</th>
<th>yes</th>
<th>no</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Annotator 2 Results</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>yes</td>
<td>350</td>
<td>10</td>
<td>360</td>
</tr>
<tr>
<td>no</td>
<td>20</td>
<td>120</td>
<td>140</td>
</tr>
<tr>
<td>total</td>
<td>370</td>
<td>130</td>
<td>500</td>
</tr>
</tbody>
</table>

Using the above values and the kappa statistic, calculate the inter-annotator agreement rate (IAA). Show your working.  
(3 marks)

Hint: The formula for the kappa statistic is: \( \frac{(P(A) - (PE))/(1 - P(E))} \), 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.


c) King (2007) noted of challenge evaluations that they “serve as a way of quantifying to what extent a system succeeds in producing the results it has been designed to produce, they tell us nothing of any other virtues or weaknesses”.

To what extent do you agree with this opinion, in relation to text mining, especially from the point of view of an end-user? Justify your position.  
(3 marks)

d) Discuss by reference to specific examples how the results of named entity recognition with concept mapping, and also of event mining, can be exploited in a document search application. Explain how, in such applications, users may be relieved of the need to use semantic labels in formulating a search.  
(4 marks)

END OF EXAMINATION
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

\[
t = \frac{\bar{x} - \mu}{s^2 / \sqrt{N}} \tag{2}
\]

where \( \bar{x} \) is the sample mean, \( s^2 \) the sample variance, \( N \) the sample size, and \( \mu \) the mean of the distribution.

<table>
<thead>
<tr>
<th>( \nu ) ( \cap \alpha )</th>
<th>0.05</th>
<th>0.025</th>
<th>0.01</th>
<th>0.005</th>
<th>0.0025</th>
<th>0.001</th>
<th>0.0005</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6.314</td>
<td>12.71</td>
<td>31.82</td>
<td>63.66</td>
<td>127.3</td>
<td>318.3</td>
<td>636.6</td>
</tr>
<tr>
<td>2</td>
<td>2.920</td>
<td>4.303</td>
<td>6.965</td>
<td>9.925</td>
<td>14.09</td>
<td>22.33</td>
<td>31.60</td>
</tr>
<tr>
<td>3</td>
<td>2.353</td>
<td>3.182</td>
<td>4.541</td>
<td>5.841</td>
<td>7.453</td>
<td>10.21</td>
<td>12.92</td>
</tr>
<tr>
<td>4</td>
<td>2.132</td>
<td>2.776</td>
<td>3.747</td>
<td>4.604</td>
<td>5.998</td>
<td>7.173</td>
<td>8.610</td>
</tr>
<tr>
<td>5</td>
<td>2.015</td>
<td>2.571</td>
<td>3.365</td>
<td>4.032</td>
<td>4.773</td>
<td>5.893</td>
<td>6.869</td>
</tr>
<tr>
<td>6</td>
<td>1.943</td>
<td>2.447</td>
<td>3.143</td>
<td>3.707</td>
<td>4.317</td>
<td>5.208</td>
<td>5.959</td>
</tr>
<tr>
<td>7</td>
<td>1.895</td>
<td>2.365</td>
<td>2.998</td>
<td>3.499</td>
<td>4.029</td>
<td>4.785</td>
<td>5.408</td>
</tr>
<tr>
<td>8</td>
<td>1.860</td>
<td>2.306</td>
<td>2.896</td>
<td>3.355</td>
<td>3.833</td>
<td>4.501</td>
<td>5.041</td>
</tr>
<tr>
<td>9</td>
<td>1.833</td>
<td>2.262</td>
<td>2.821</td>
<td>3.250</td>
<td>3.690</td>
<td>4.297</td>
<td>4.781</td>
</tr>
<tr>
<td>10</td>
<td>1.812</td>
<td>2.228</td>
<td>2.764</td>
<td>3.169</td>
<td>3.581</td>
<td>4.144</td>
<td>4.587</td>
</tr>
<tr>
<td>20</td>
<td>1.725</td>
<td>2.086</td>
<td>2.528</td>
<td>2.845</td>
<td>3.153</td>
<td>3.552</td>
<td>3.850</td>
</tr>
<tr>
<td>30</td>
<td>1.697</td>
<td>2.042</td>
<td>2.457</td>
<td>2.750</td>
<td>3.030</td>
<td>3.385</td>
<td>3.646</td>
</tr>
<tr>
<td>40</td>
<td>1.684</td>
<td>2.021</td>
<td>2.423</td>
<td>2.704</td>
<td>2.971</td>
<td>3.307</td>
<td>3.551</td>
</tr>
<tr>
<td>50</td>
<td>1.676</td>
<td>2.009</td>
<td>2.403</td>
<td>2.678</td>
<td>2.937</td>
<td>3.261</td>
<td>3.496</td>
</tr>
<tr>
<td>60</td>
<td>1.671</td>
<td>2.000</td>
<td>2.390</td>
<td>2.660</td>
<td>2.915</td>
<td>3.232</td>
<td>3.460</td>
</tr>
<tr>
<td>80</td>
<td>1.664</td>
<td>1.990</td>
<td>2.374</td>
<td>2.639</td>
<td>2.887</td>
<td>3.195</td>
<td>3.416</td>
</tr>
<tr>
<td>100</td>
<td>1.660</td>
<td>1.984</td>
<td>2.364</td>
<td>2.626</td>
<td>2.871</td>
<td>3.174</td>
<td>3.390</td>
</tr>
<tr>
<td>120</td>
<td>1.658</td>
<td>1.980</td>
<td>2.358</td>
<td>2.617</td>
<td>2.860</td>
<td>3.160</td>
<td>3.373</td>
</tr>
<tr>
<td>( \infty )</td>
<td>1.645</td>
<td>1.960</td>
<td>2.326</td>
<td>2.576</td>
<td>2.807</td>
<td>3.090</td>
<td>3.291</td>
</tr>
</tbody>
</table>

Table 1: \( t \) distribution table (\( \nu \): degrees of freedom)

Continued on next page
Continued from first page

$X^2$ statistic

$$X^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$  \hspace{1cm} (3)

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

<table>
<thead>
<tr>
<th>$\nu \setminus \alpha$</th>
<th>0.250</th>
<th>0.100</th>
<th>0.050</th>
<th>0.025</th>
<th>0.010</th>
<th>0.005</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.32330</td>
<td>2.70554</td>
<td>3.84146</td>
<td>5.02389</td>
<td>6.63490</td>
<td>7.87944</td>
</tr>
<tr>
<td>2</td>
<td>2.77259</td>
<td>4.60517</td>
<td>5.99146</td>
<td>7.37776</td>
<td>9.21034</td>
<td>10.58677</td>
</tr>
<tr>
<td>10</td>
<td>12.54886</td>
<td>15.98718</td>
<td>18.30704</td>
<td>20.48318</td>
<td>23.20925</td>
<td>25.1818</td>
</tr>
<tr>
<td>13</td>
<td>19.36886</td>
<td>23.54183</td>
<td>26.29623</td>
<td>28.84535</td>
<td>31.99934</td>
<td>34.26719</td>
</tr>
<tr>
<td>14</td>
<td>21.60489</td>
<td>25.98942</td>
<td>28.86930</td>
<td>31.52638</td>
<td>34.80531</td>
<td>37.15645</td>
</tr>
<tr>
<td>15</td>
<td>23.82769</td>
<td>28.41198</td>
<td>31.40443</td>
<td>34.16961</td>
<td>37.56623</td>
<td>39.99685</td>
</tr>
<tr>
<td>16</td>
<td>26.03927</td>
<td>30.81328</td>
<td>33.92444</td>
<td>36.78071</td>
<td>40.28936</td>
<td>42.79565</td>
</tr>
<tr>
<td>17</td>
<td>28.24115</td>
<td>33.19624</td>
<td>36.41503</td>
<td>39.36408</td>
<td>42.97982</td>
<td>45.55851</td>
</tr>
<tr>
<td>18</td>
<td>30.43457</td>
<td>35.56317</td>
<td>38.85714</td>
<td>41.23317</td>
<td>45.61648</td>
<td>48.28988</td>
</tr>
<tr>
<td>19</td>
<td>32.62049</td>
<td>37.91592</td>
<td>41.33714</td>
<td>44.46079</td>
<td>48.27824</td>
<td>50.99338</td>
</tr>
<tr>
<td>20</td>
<td>34.79974</td>
<td>40.25602</td>
<td>43.77297</td>
<td>46.97924</td>
<td>50.89218</td>
<td>53.67196</td>
</tr>
</tbody>
</table>

Table 2: $\chi^2$ distribution table ($\nu$: degrees of freedom)