The University Of Sheffield.

Data Provided: None

DEPARTMENT OF COMPUTER SCIENCE **Autumn Semester 2014-2015**

TEXT PROCESSING **2 hours and 30 minutes**

Answer the question in Section A, and THREE questions from Section B.

All questions carry equal weight. Figures in square brackets indicate the percentage of available marks allocated to each part of a question.

SECTION A

1. a) Two Information Retrieval systems, System 1 and System 2, each return a ranked list of 10 documents they believe to be relevant for a particular query. It is known that this collection has 12 relevant documents. The following table shows whether each document returned by each system is actually relevant (✓) or not (×) to the query.

Document	System 1	System 2
d1	✓	×
d2	×	✓
d3	✓	×
d4	✓	✓
d5	✓	✓
d6	×	×
d7	×	✓
d8	✓	✓
d9	×	✓
d10	×	✓

Compute the overall precision of each System, showing the equations as part of your answer. Then, compute the precision at two cutoff points: top 3 and top 5. Finally, discuss the differences between overall precision and precision at different cutoff points when comparing two or more Information Retrieval systems. Use your solution to exemplify your answer. [30%]

b) What is the noisy channel model? Give a diagram of the model as part of your answer. Suggest a text processing context where the noisy channel model has been used. [30%]

c) Explain two metrics to evaluate the quality of binary (negative/positive) sentiment analysis systems. Show their formulae as part of your answer. [20%]

d) LZ77 is a popular compression method, used in common compression utilities such as $gzip$. The following shows some possible LZ77 encoder output (assuming the encoding representation presented in the lectures of the Text Processing module):

$$\langle 0,0,b\rangle\langle 0,0,a\rangle\langle 0,0,d\rangle\langle 3,3,b\rangle\langle 1,3,a\rangle\langle 1,3,d\rangle\langle 1,3,a\rangle\langle 11,2,a\rangle$$

Sketch how LZ77 works. State the output that would be produced by decoding the above representation, showing how your answer is derived. [20%]

SECTION B

2. In the context of Information Retrieval, given the following documents:

Document 1: Your dataset is corrupt. Corrupted data does not hash!!!

Document 2: Your data system will transfer corrupted data files to trash.

Document 3: Most politicians are corrupt in many developing countries.

and the query:

Query 1: hashing corrupted data

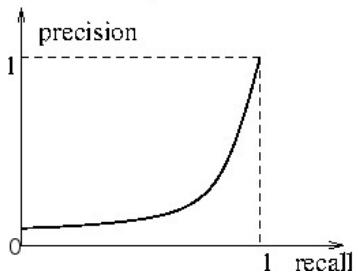
a) Apply the following term manipulations on document terms: *stoplist removal*, *capitalisation* and *lemmatisation*, showing the transformed documents. Explain each of these manipulations. Provide the stoplist used, making sure it includes punctuation. [20%]

b) Explain what is meant by an *inverted index* and why such indices are important in the context of Information Retrieval. Show how Document 1, Document 2 and Document 3 would be represented using an inverted index which includes term frequency information. This inverted index should not have more than 10 words. [20%]

c) Using *term frequency* (TF) to weight terms, represent the documents and query as vectors. Produce rankings of Document 1, Document 2 and Document 3 according to their relevance to Query 1 using two metrics: Cosine Similarity and Euclidean Distance. Show which document is ranked first according to each of these metrics. [30%]

d) Define the precision and recall measures in Information Retrieval. Is Graph A a possible precision/recall graph? Is a curve of this shape likely when evaluating the results of a realistic Information Retrieval system such as Google? Explain your answers. [20%]

Graph A



e) Discuss the advantages and disadvantages of *boolean* versus *ranked* approaches to Information Retrieval. [10%]

3. a) When applied to translating from French to English, the IBM approach to Statistical Machine Translation can be expressed by the following equation:

$$E^* = \underset{E}{\operatorname{argmax}}\ P(E) \cdot P(F|E)$$

Explain what this equation means, and indicate the role played by the components $P(E)$ and $P(F|E)$ in the process of translation. [20%]

b) Show how the equation given in 3(a) is derived using Bayes Rule. What is the benefit of this approach as compared to one attempting to use the probability $P(E|F)$ directly? [30%]

c) Consider the following text processing techniques studied in the context of Information Retrieval: capitalisation, stop-word removal and stemming. Discuss whether or not each of these techniques could be useful in the context of Phrase-based Statistical Machine Translation and why. Would each of these techniques be equally applicable to the source and target language data? At which stage of the process would they be applied? Give examples of words to support your answer. [25%]

d) Ensuring that output is grammatical and fluent is one of the main goals in machine translation. Explain how this problem is addressed in Phrase-based Statistical Machine Translation approaches. Your explanation should specify what type of data is necessary to ensure fluency in the building of a Phrase-based Statistical Machine Translation system. Discuss how you would collect such data for a given language, say Spanish. Would you pre-process the data in any way? Cite and explain two pre-processing techniques. [25%]

4. a) Text compression techniques are important because growth in volume of text continually threatens to outstrip increases in storage, bandwidth and processing capacity. Briefly explain the differences between:

   (i) **symbolwise** and **dictionary** text compression methods; [10%]

   (ii) **modelling** versus **coding** steps; [10%]

   (iii) **static**, **semi-static** and **adaptive** techniques for text compression. [10%]

   b) Sketch the algorithm for Huffman coding, i.e. for generating variable-length codes for a set of symbols, such as the letters of an alphabet. What does it mean to say that the codes produced are *prefix-free*, and why do they have this property? [20%]

   c) We want to compress a large corpus of text of the (fictitious) language *Fontele*. The writing script of Fontele employs only the six letters found in the language name (`f,o,n,t,e,l`) and the symbol ␣, used as a 'space' between words. Corpus analysis shows that the probabilities of these seven characters are as follows:

Symbol	Probability
`e`	0.3
`f`	0.04
`l`	0.26
`n`	0.2
`t`	0.04
`o`	0.1
␣	0.06

   (i) Show how to construct a Huffman code tree for Fontele, given the above probabilities for its characters. Use your code tree to assign a binary code for each character. [20%]

   (ii) Given the code you have generated in 4(c)(i), what is the average bits-per-character rate that you could expect to achieve, if the code was used to compress a large corpus of Fontele text? How does this compare to a minimal fixed length binary encoding of this character set? [20%]

   (iii) Use your code to encode the message "`telefone␣noel`" and show the resulting binary representation. Compare the average bits-per-character rate achieved for this message to the expected rate that you computed in 4(c)(ii), and suggest an explanation for any difference observed between the two values. [10%]

5. a) Consider the two sentences:

   - *My new phone works well, is very pretty and much faster than the old one.*
   - *My new phone has 32GB of memory and plays videos.*

   What is the first step to detect the sentiment in these two sentences? Should both these sentences be addressed in the same way by Sentiment Analysis approaches? If not, explain a common approach to select only relevant sentences for Sentiment Analysis. [20%]

   b) Given the following sentences S1 to S4 and opinion lexicon of adjectives, apply the weighted lexical-based approach to classify EACH sentence as **positive**, **negative** or **objective**. Show the final emotion score for each sentence, and also how it was generated. In addition to using the lexicon, make sure you consider any general rules that have an impact on the final decision. Explain these rules when they are applied. [15%]

Lexicon:	
awesome	5
boring	-3
brilliant	2
funny	3
happy	4
horrible	-5

   (S1) He is brilliant and funny.
   (S2) I am not happy with this outcome.
   (S3) I am feeling AWESOME today, despite the horrible comments from my supervisor.
   (S4) He is extremely brilliant but boring, boring, very boring.

   c) According to Bing Liu's model, an **opinion** is said to be a quintuple $(o_j, f_{jk}, so_{ijkl}, h_i, t_l)$. Explain each of these elements and exemplify them with respect to the following text. Identify the features present in the text, and for each indicate its sentiment value as either *positive* or *negative*. Discuss two language processing challenges in automating the identification of such elements. [25%]

   "I have just bought the new iPhone 5. It is a bit heavier than the iPhone 4, but it is much faster. The camera lenses are also much better, taking higher resolution pictures. The only big disadvantage is the cost: it is the most expensive phone in the market. Lucia Specia, 12/08/2014."

d) Assume a lexicon-based approach to binary Sentiment Analysis. A manually created initial lexicon is available which contains only three positive words:

- good
- nice
- excellent

and three negative words:

- bad
- terrible
- poor

This lexicon needs to be expanded in order for the approach to be effective in a realistic task. Explain two alternative methods to expand this lexicon automatically. Which of these methods should result in the larger lexicon and why? [20%]

e) Explain the intuition behind using a Naive Bayes classifier for Sentiment Analysis. Give the general classifier equation as part of your answer. What are the main components in this classifier? Give two types of features that could be used and provide examples for these types of features. [20%]

END OF QUESTION PAPER