The University Of Sheffield.

Data Provided: None

DEPARTMENT OF COMPUTER SCIENCE **Autumn Semester 2016-2017**

TEXT PROCESSING **2 hours**

Answer THREE questions.

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

1. 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 *stemming*, showing the transformed documents. Explain each of these manipulations. Include in your answer the stoplist you used, making sure it includes punctuation, but no content words. [20%]

   b) Show how Document 1, Document 2 and Document 3 would be represented using an *inverted index* which includes term frequency information. [10%]

   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) Explain the intuition behind using TF.IDF (*term frequency inverse document frequency*) to weight terms in documents. Include the formula (or formulae) for computing TF.IDF values as part of your answer. For the ranking in the previous question using cosine similarity, discuss whether and how using TF.IDF to weight terms instead of TF only would change the results (assume here that the document collection consists solely of Documents 1 – 3). [20%]

   e) Explain the metrics Precision, Recall and F-measure in the context of evaluating an Information Retrieval system against a gold-standard set. Discuss why it is not feasible to compute recall in the context of searches performed on very large collections of documents, such as the Web. [20%]

2. a) Explain the differences between *direct*, *transfer-based* and *interlingual* approaches to machine translation. Give the main advantage and disadvantage of each of these approaches. [15%]

   b) (i) What is the noisy channel model and how can it be applied to machine translation? [15%]

   (ii) State the fundamental probabalistic equation formalising the noisy channel model for machine translation and explain how it relates to that model. Show how the equation can be rewritten using Bayes Theorem and then simplified. Be sure to state in words what each of the terms in the equation is. [15%]

   (iii) The simplified equation of 2(b)(ii) has three components that need to be implemented to build a working machine translation system. Name each of these components and describe briefly what its role in the translation system is. [15%]

   c) Explain in a general way how word alignments are learnt from a parallel corpus in IBM model 1. Full mathematical details are not necessary. [20%]

   d) Explain briefly how the BLEU measure, which is used to automatically evaluate the quality of machine translated texts, is calculated. [20%]

3. a) Differentiate *subjectivity* from *sentiment*. How are the tasks of Subjectivity Classification and Sentiment Analysis related? [10%]

   b) Give Bing Liu's model for an **opinion**. Explain each of the elements in the model and exemplify them with respect to the following text, which is adapted from a *TripAdvisor* review of a restaurant in Sheffield. 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 and illustrate these challenges with reference to the example text. [30%]

> "I went with my girlfriend on a Friday night, and was greeted in a friendly way by the waitress. It is simply decorated and clean, but for my personal taste was a bit too bright, and could do with a bit more colour. It is fantastic you can take your own wine and there is no uncorking fee. We was welcomed very well by the staff and I liked it that she explained the specials board to us and explained what each dish was. For starters we had the meat balls... It was amazing !! The sauce was so tasty! For our main course we had a sea food mixture with a sauce ... We felt it was a little expensive for what it was and was nice but could have been a few pounds cheaper." Trevor M., posted 12/10/2015

   c) Explain the graded lexicon-based approach for Sentiment Analysis. Given the following sentences and opinion lexicon, apply this approach to classify *each* sentence in S1-S3 as **positive**, **negative** or **objective**. Show the final emotion score for each sentence and also how this score was generated. Give any general rules that you used to calculate this score as part of your answer. Explain these rules when they are applied. [25%]

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.

d) A second approach to Sentiment Analysis is the corpus-based supervised learning approach.

(i) Explain the corpus-based supervised learning approach to Sentiment Analysis in general terms, i.e. in terms of inputs, outputs and processes involved. [5%]

(ii) Explain how a Naive Bayes classifier can be trained and then used to predict the polarity class (positive or negative) of a subjective text. Be sure to give the mathematical formulation of the Naive Bayes classifier. [10%]

(iii) Suppose you are given the following set of labelled examples as training data:

Doc	Words	Class
1	A sensitive, moving, brilliant work	Positive
2	An edgy thriller that delivers a surprising punch	Positive
3	A sensitive, insightful, beautiful film	Positive
4	Neither revelatory nor truly edgy – merely crassly flamboyant and comedically labored	Negative
5	Unlikable, uninteresting, unfunny, and completely, utterly inept	Negative
6	A sometimes incisive and sensitive portrait that is undercut by its awkward structure and ...	Negative
7	It's a sometimes interesting remake that doesn't compare to the brilliant original	Negative

Using as features just the adjectives (underlined words in the examples), how would a Naive Bayes sentiment analyser trained on these examples classify the sentiment of the new, unseen text show below?

Doc	Words	Class
8	A sensitive comedy that is moving and surprising	???

Show how you derived your answer. You may assume standard pre-processing is carried out, i.e. tokenisation, lowercasing and punctuation removal. You do not need to smooth feature counts.

[20%]

4. a) (i) Explain how the LZ77 compression method works. [30%]

(ii) Assuming the encoding representation used in class (i.e. in the lectures of the Text Processing module), show what output would be produced by the LZ77 decoder for the following representation. Show how your answer is derived.

<0,0,y> <0,0,a> <0,0,b> <2,1,-> <0,0,d> <5,5,o> <1,4,o>

[15%]

b) We want to compress a large corpus of text of the (fictitious) language *Sosumi*. The writing script of Sosumi uses only the letters {s, o, u, m, i, d} and the symbol ∼ (which is used as a 'space' between words). Corpus analysis shows that the probabilities of these seven characters are as follows:

Symbol	Probability
s	0.12
o	0.23
u	0.05
m	0.25
i	0.08
d	0.09
∼	0.18

(i) Sketch the algorithm for Huffman coding. Illustrate your answer by constructing a code for Sosumi, based on the above character probabilities. [30%]

(ii) Use your Huffman code from 4(b)(i) to encode the message: "modo∼mi∼sumo" How does the bits-per-character rate achieved on this message compare to a minimal fixed length binary encoding of the same character set? [5%]

c) What is a *canonical* Huffman code? Show how a canonical Huffman code can be derived from the Huffman code that you created for Sosumi in 4(b)(i). What are the advantages of using a canonical Huffman code? [20%]

END OF QUESTION PAPER