The University Of Sheffield.

Data Provided: None

DEPARTMENT OF COMPUTER SCIENCE Autumn Semester 2020-21

TEXT PROCESSING: QUIZ 1

Answer ALL questions. Questions 1 and 2 are worth 50 marks each. The paper, therefore, is out of 100.

THIS PAGE IS BLANK

1. 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. [10%]

   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. [15%]

   "I have just bought the new iPhone 12. It is a bit heavier than the iPhone 11, 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. Michael Jordan, 12/08/2020."

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? [10%]

2. a) 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 the concept "independence assumption" used by a Naive Bayesian Classifier. [4%]

(iii) 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. [8%]

(iv) 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.

[10%]

b) What are the stages of processing commonly followed within natural language generation (NLG) systems? For each stage, please explain its purposes. [8%]

c) Explain three metrics to evaluate the quality of binary (negative/positive) sentiment analysis systems. Give their intuitions and show their formulae. [5%]

d) An NLG system needs to take care of details of language such as morphological details. How does inflectional morphology differ from derivational morphology? Explain with examples from the English language. [10%]

END OF QUESTION PAPER