The University Of Sheffield.

Data Provided:
None

DEPARTMENT OF COMPUTER SCIENCE

Autumn Semester 2022/23

2 hours

TEXT PROCESSING

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.

THIS PAGE IS BLANK

1. a) Sentiment Analysis (SA) can be performed at different levels of granularity. Describe three different levels on which sentiment analysis can be performed, and explain which level you think gives the most precise analysis. [15%]

b) Explain the weighted lexical-based approach for Sentiment Analysis. [15%]

c) Specify the quintuple of Bing Liu's model for Sentiment Analysis and explain each of its elements. Exemplify the model with respect to the example below. Identify all the features present in the text and, for each of them, indicate the sentiment value as *positive* or *negative*.

> "I have just arrived in a late night flight from Lisbon with Ryanair. The flight was 7 hours delayed, which was a terrible experience. The lack of drinkable water on-board also made the in-flight experience awful, mainly for elderly and children. The main drawback was the ticket price, that was very expensive for a low-budget company. The only good thing was the staff: they were very helpful and managed to keep everyone calm. William MacS, 12/10/2022."

[20%]

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

(i) Give the mathematical formulation of the Naive Bayes classifier and explain in detail how such a classifier can be trained and used to predict the polarity class (positive or negative) of a subjective text. [15%]

(ii) Explain what assumption makes the Naive Bayes classifier *naive*. Discuss the validity of this assumption for NLP classification tasks such as Sentiment Analysis. [10%]

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

Doc	Words	Class
1	Amazing movie, the perfect way to make a sequel.	Positive
2	Hypnotic, surrealist, and most of all, maybe the most beautiful movie of the year.	Positive
3	Beautiful film. Well-paced; never felt it was overly long.	Positive
4	Visually stunning and amazing. A bit long, perhaps, but never boring.	Positive
5	Great plot but bad acting. Too long, boring in the middle.	Negative
6	Very boring, not entertaining, too artsy, plot holes galore, too long.	Negative
7	Visually beautiful but way too long and the soundtrack was annoying	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
9	Beautiful sets but too long and sooo boring.	???

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

[25%]

2. a) The table below shows a confusion matrix.

	Predicted +	Predicted -
True +	100	40
True -	60	300

(i) Calculate the following measures: overall accuracy; precision; recall; F-Measure. [10%]

(ii) Explain under what circumstances using accuracy as a measure will cause issues. [5%]

b) For the processing pipeline commonly followed within natural language generation (NLG) systems

(i) Describe the stages of processing within the pipeline. For each stage, please explain its purposes. [15%]

(ii) Explain which aspect of a typical NLG pipeline is most important and why? [5%]

c) (i) Explain what are *Cue phrases* and *Rhetorical Relations*. [5%]

(ii) List four Rhetorical Relations and provide an example sentence that contains at least two Rhetorical Relations from your list. [10%]

d) Briefly discuss two baseline approaches for *Word Sense Disambiguation*. [15%]

e) Using the *Lesk algorithm* show how the word *bank* in Sentence 1 is disambiguated and which sense is chosen:

> **Sentence 1**: That bank is shaped by the currents and water channels and holds deposits of sand.

given the following two WordNet senses:

bank[1]	Gloss:	a financial institution that accepts deposits and channels the money into lending activities
	Examples:	"he cashed a check at the bank", "that bank holds the mortgage on my home"
bank[2]	Gloss:	sloping land (especially the slope beside a body of water)
	Examples:	"they pulled the canoe up on the bank", "he sat on the bank of the river and watched the currents"

[15%]

f) Calculate the *word similarity* between *knowledge* and *information* and *knowledge* and *digital* using cosine similarity, and determine which one is closer in meaning to *knowledge*, assuming the raw counts from the following shortened table:

	retrieval	data	computer
information	25	8	2
digital	5	4	7
knowledge	5	12	6

[20%]

3. a) One of the main tasks of *Information Extraction* is that of *Relation Extraction*. Provide a concise explanation of what relation extraction is, and demonstrate it with an example. [10%]

b) Moreover, as relation extraction may be subdivided into two tasks, *relation detection* and *relation classification*, very briefly explain what each of these two tasks are, using an example to demonstrate your explanation and clarifying how they differ. [10%]

c) As supervised learning approaches to relation extraction require extremely large amounts of manually annotated training data, alternative approaches that address this problem have been devised, such as a *bootstrapping approach* to relation extraction. Explain VERY BRIEFLY the bootstrapping approach to relation extraction, including its main advantages and disadvantages, and demonstrating how it works with an example. [30%]

d) In the context of Information Retrieval, given the following documents:

Document 1: Summer scent!! Find your summer scent.

Document 2: You will find all lovely summer scents at the Summer Province Market.

Document 3: Marketing is growing, in the Summer Scently Provincial Markets.

and the query:

Query 1: summer scent province market

(i) 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%]

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

(iii) 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 the Euclidean Distance as metric. Show which document is ranked first according to this metric. [20%]

END OF QUESTION PAPER