The University Of Sheffield.

Ancillary Material: None

DEPARTMENT OF COMPUTER SCIENCE **Autumn Semester 21/22**

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.

THIS PAGE IS BLANK

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: Many politicians are corrupt in some 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 have used. [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) What are the stages of processing commonly followed within natural language generation (NLG) systems? For each stage, please explain its purposes. [25%]

   b) 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. [25%]

   c) Explain three metrics to evaluate the quality of binary (negative/positive) sentiment analysis systems. Give their intuitions and show their formulae. [25%]

   d) Explain the intuition behind using a Naive Bayes classifier for text classification. Give the general classifier equation as part of your answer. What are the main components in this classifier? [25%]

3. a) Differentiate *subjectivity* from *sentiment*. How are the tasks of Subjectivity Classification and Sentiment Analysis related? [20%]

   b) Discuss the relevance of automatic techniques for sentiment analysis for marketing purposes. [20%]

   c) Explain the weighted lexical-based approach for Sentiment Analysis. Given the following sentences and opinion lexicon, apply this approach to classify *each* sentence in S1-S4 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. [30%]

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.

   d) 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. 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. [30%]

> "I have just bought the new iPhone 13. It is a bit heavier than the iPhone 12, 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. Mark Jobs, 12/11/2021."

4. *Relation extraction* is one of the main tasks in the sub-area of text processing known as *information extraction*.

   a) Briefly explain what the task of relation extraction is and illustrate your answer with an example. [10%]

   b) Relation extraction is sometimes split into two sub-tasks, *relation detection* and *relation classification*. Explain what relation detection and relation classification are, making clear how they differ, and illustrate your answer with an example (you may use the same example as in the preceeding part, but are not required to do so). [10%]

   c) Various linguistic features of natural language make relation extraction hard. Identify **three** such features and give an example of each. [20%]

   d) Supervised learning approaches to relation extraction have been quite successful but have the drawback of requiring substantial amounts of manually annotated training data. Two approaches that have been devised to address this problem are the *distant supervision approach* to relation extraction and the *bootstrapping approach* to relation extraction.

      (i) Briefly explain how the *distant supervision approach* to relation extraction works, give an example of how it works and briefly identify the strengths and weaknesses of this approach. [30%]

      (ii) Briefly explain how the *bootstrapping approach* to relation extraction works, give an example of how it works and briefly identify the strengths and weaknesses of this approach. [30%]

END OF QUESTION PAPER