University College Dublin
An Coláiste Ollscoile, Baile Átha Cliath

2022/2023 AUTUMN TRIMESTER EXAMINATIONS

COMP47600

Text Analytics

Module Coordinator: Professor Mark Keane

Student Number

Seat Number

Time Allowed: 120 minutes

Materials Permitted in the Exam Venue:
None

Materials to be Supplied to Students:
None

Instructions to Students:

Answer any FOUR questions.
All questions carry equal marks.
Total marks available 100.
Use of calculators is prohibited.

1. Several different approaches have been used to find temporal regularities in text data. Describe three of the approaches that have been used, illustrating each with an example from the literature, with a critical evaluation of each. **[3 x 8.33 marks]**

[25 marks overall]

2. Latent Methods (e.g., Latent Semantic Indexing and Latent Semantic Analysis) are used in text analytics to find hidden or latent associations between words. Describe how LSI/LSA are computed when applied to a term-document matrix. **[10 marks]** Evaluate these methods on some of the issues that arise in their use. **[5 marks]** How do these methods differ from more recent attempts to define word embeddings (e.g., in word2vec or in BERT)? **[10 marks]**

[25 marks overall]

3. Log-Likelihood Ratios (LLRs) are used in text analytics to find patterns of significant words in comparisons between different texts or text-corpora. What formula is used to compute LLRs? **[5 marks]** Describe four different ways in which LLRs have been applied to textual data, giving a sample study for each, along with a critical evaluation of that study. **[4 x 5 marks]**

[25 marks overall]

4. In machine learning, a fundamental distinction is often made between supervised and unsupervised methods. Describe the main differences between these two broad classes of methods. **[5 marks]** Then, give detailed accounts of one example of each class (i.e., provide two specific technique descriptions, one that is supervised and one that is unsupervised). **[2 x 5 marks]** Finally, illustrate each of these techniques with an example from the text analytics literature. **[2 x 5 marks]**

[25 marks overall]

5. Text analytics typically begins with the pre-processing of each text-item, in some selected corpus, to prepare it for subsequent processing. Describe five of the main pre-processing steps that are carried out during this pre-processing stage and show how each step modifies a given text fragment. In describing each pre-processing step, explain why it is used and the benefits that follow from its use. **[5 x 5 marks]**

[25 marks overall]

oOo