University College Dublin
An Coláiste Ollscoile, Baile Átha Cliath

SEMESTER I EXAMINATION – 2015/2016

COMP 47240

Using Text Analytics to Discover Meaning

Prof. S. Dobson

Prof. P. Cunningham

Prof. M. Keane*

Time allowed: 2 hours

Instructions for candidates

Answer any FOUR questions.

All questions carry equal marks.

The paper is marked out of 100%.

Use of calculators is prohibited.

.

Instructions for invigilators

Use of calculators is prohibited.

1. Text Analytics typically begins with the pre-processing of each text, in some selected corpus of texts, to prepare it for subsequent processing.

   Describe five of the typical pre-processing steps that are carried out during in this initial stage of processing and show how each pre-processing step might modify a text fragment.

   In describing each pre-processing step, explain why it is used and describe some of the benefits it brings to subsequent processing.

   [5 x 5%]
   [25% overall]

2. 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%]

   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%]

   Finally, illustrate each of these techniques with an example from the text analytics literature. [2 x 5%]
   [25% overall]

3. Describe, in general, what evaluation tries to achieve when it is carried out on a text analytics system. [5%]

   Evaluation in text analytics depends on four key ideas (Ground Truth, Precision, Recall and the F-measure). Explain what each of these four ideas involve, giving examples of how each one may be used to evaluate a text analytics system.

   [4 x 5%]
   [25% overall]

4. Three main methods are used to identify sentiments in texts: human ratings, sentiment lexicons and sentiment classifiers. Describe each of these methods in detail and describe how they are used. Also critically evaluate the adequacy of each of these methods as solutions to the sentiment-identification problem. [3 x 8.333%]
[25% overall]

5. Describe the three broad families of similarity techniques used in text analytics. Describe one instance of each family, illustrated with an example and identify when it is appropriate to use one technique-family rather than another, and any issues that can arise in the use of such techniques. [3 x 8.333%]
[25% overall]