

COMP90049 Knowledge Technologies

Final exam, Semester 1, 2018

[illegible]

Part I : String/Text Processing

[43 marks in total]

1. Describe two (or more) steps that we would typically perform in the Tokenisation process for an Information Retrieval collection, according to the discussion in this subject. [4 marks]

2. It has been claimed that there are three primary types of “information need” in a web search context: “informational”, “navigational”, and “transactional”. Briefly describe each, optionally with the aid of an example. [3 marks]

3. In the context of Information Retrieval:
 - (a) Explain how Data retrieval is different to Information Retrieval. [2 marks]
 - (b) Give an example of a method or source of information that we might incorporate in our engine, that is specific to Web-scale Information Retrieval. [1 marks]

4. ...And more questions to add up to the marks stated above. (-:

Part II: Data Mining/Machine Learning [42 marks in total]

For these questions, we have a training dataset comprised of the following 6 instances, 3 attributes, and two classes F and T, and a single test instance labelled with ?:

ele	fed	aust	CLASS
1	1	1	F
1	0	0	F
1	1	0	T
1	1	0	T
1	1	1	T
1	1	1	T
0	0	0	?

5. Classify the given test instance using the method of Naive Bayes, as described in this subject. [4 marks]

6. Explain why 1-Nearest Neighbour will give a different prediction to 3-Nearest Neighbour on this test instance. (Note that it is not required to show all of your workings for this question.) [2 marks]

7. Consider the method of Random Forests:
 - (a) Briefly explain how a Random Forest would be constructed on the training data above. [4 marks]
 - (b) Is there any evidence that a Random Forest would label the given test instance differently to a regular Decision Tree? [3 marks]

8. Exclude the CLASS labels from the dataset, and cluster all 7 instances using the method of k -means. Apply the Manhattan Distance as a similarity measure; use the second (1,0,0) and third (1,1,0) instances as seeds. [4 marks]

9. ...And more questions to add up to the marks stated above. :-)