Comments

Question 1

58 Candidates answered this question. The highest mark was 28/30, the lowest 7/30. Several question parts were generally the subject of over-answering and verbosity, however there were numerous other answers that were underdeveloped and general in nature. At least one candidate did not bring a calculator to the exam.

Q1a was generally well answered, although some wrote much more than necessary for the marks available. Some apparently misunderstood the question and wrote about document processing aspects, and did not address classification of needs.

Q1b i) and Q1b ii) were good distinguishers, in that several candidates revealed lack of knowledge of the role of DF (document frequency) in ordering queries for processing. Some thought that alphabetic ordering was important. Others indulged in hedging their answers: giving an examiner several choices of answer does not convince him that a solid understanding has been achieved. Some thought that the answer to (ii) was simply to process the bracketed expressions first. There was a query raised about a mistake (missing open parenthesis). This was promptly answered as soon as communicated by the invigilator and examinations office.

Q1b iii) was answered very well by some, however minimally by several who confined their answer to bullet points (despite the revision session advice not to do so, unless it is the last moments of the exam). Answering in bullets serves mainly to convince an examiner that basic bookwork (slide contents) has been memorised, but that there is no ability to construct an argument.

Q1ci was generally well answered, although there were several underdeveloped answers for the marks available. Some thought that a tokenizer trained on business news would have no problems when being used for other domains. Some thought that the Porter stemmer would be useful for languages other than English (despite it being a stemmer entirely customised for English). Others correctly commented on disadvantages and advantages of stemming and noted that documents could not only be from different domains but also from different languages. Many answers regarding spell-checking assumed this should only apply to queries. Some stated that they would assume that documents would be properly spelled to begin with. One need only think of blogs, emails, etc., to see the invalidity of such a position. Having just examined some 20 final year reports, where a professional approach to writing is assumed by default, I can unfortunately confirm on the basis of this sample that judicious use of a spell-checker is a foreign notion to many students. Some who advocated stemming of documents were concerned that spell-checking of the results would be a disadvantage leading to meaningless forms, but did not consider that spell-checking could take place before stemming. This may have been due to ignoring the word ‘options’ and treating the list as a strictly ordered list of processes.

Q1d i) was generally well answered, although some confused term frequency with collection frequency, or did not give sufficient explanation of the concept to reveal that they understood it.

Q1d ii) Some candidates failed to appreciate the nature of IDF.

Q1e was generally well answered, although some gave only one way, and a few thought that the higher the cosine score, the less similar two documents were.

Q1f was also well answered by many, although some did not attempt it. In several cases, candidates failed to arrive at a correct value for node B after 2 iterations, however I was able to award partial marks for some. Several misunderstood the PageRank algorithm and retained values on nodes to add to incoming values. A query was raised about the double-headed arrow. This is a normal graphical means of expressing a link going in one direction and another going in the opposite direction.

Q1g was the subject of some confused or vague answers. In particular, the action to be taken was not well handled by many.

Q1h attracted some very good answers that demonstrated deep learning and moreover the ability to bring in contributions from other parts of the unit. Several answers remained very general/vague and did not show much knowledge of advanced techniques related to the unit’s topics. Many answers were underdeveloped for the marks available: I hypothesise that this was due to one of two things, or indeed both: i) a lack of reading around on one’s own initiative, of investigating recommended material (over-reliance on slides); ii) inability to produce an argument of the type expected from a final year student, demonstrating critical analysis and awareness of key issues. On several occasions, I could not decipher handwriting due to illegible words.

In answers, several students referred to topics of course workshops and brought in their experiences from these and the lab, all of which was very welcome. Some made very good references to aspects/applications beyond the immediate concerns of the course, equally welcome.

Q2 - Services

This year, Question 2 was very popular, having 66% of the students solved it.

The average mark corresponds to 56% of the question and the median mark was 16.7 out of 30. Out of 40 students, only 8 scored below 40%, and 12 scored 70% and above.

While the question parts that required calculation were the ones that had higher quality answers; the ones that required discussion were, in many cases, answered poorly, showing lack of depth in understanding of the fundamentals and analytical skills. The answers for those questions were in many cases vague, sometimes
senseless and repetitive.

Q3 Around 45% of students took this question, with an average mark of 59%.

A) was answered correctly in most cases (bookwork), although many answers failed to identify how links are established across different resources

b) was also straightforward bookwork, there were surprisingly a number of vague answers missing the main point of reusability. Part© was generally OK as most students clearly applied the linked data concepts in these settings.

D) was also straightforward, although a number of answers did not take into account that RDFa should be also targeting machines.

E) was mainly answered reasonably but rarely for full marks, as there were several issues as what the (main) role of RDFs and OWL is.

F) was answered by and large correctly, but not always with the full explanation as what each different option has as pros/cons. A couple of answers confused storage options and serialisation formats.

G) was correctly answered virtually by everyone.

Finally, in part (g), many answers failed to specifically relate to linked data in this context, e.g. what possibilities there are for linking, integration, machine-supported access etc. Rarely the point was made that only one dataset is not enough to bring the benefits clearly (i.e. if the rest of the data is not linked). Weak answers were mostly focused on either trivial aspects (e.g. staff will need to learn linked data) or points that are applicable generally to publishing any data in any format, not to linked data specifically (e.g. security concerns; storage is needed). Still, some good points on why Excel would not work well.