# UG Exam Performance Feedback
# Third Year
# 2016/2017 Semester 2

COMP38120 Documents, Services and Data on the Web

Sandra Sampaio
Jock McNaught
Goran Nenadic

## Comments

Examiner's comments on Documents (Question 1)

General comments

Comments are made in respect of unmoderated marks. 60 candidates answered Question 1. The mean mark was 17.53/30, the highest mark 26/30 and the lowest 6/30. Several answers were characterised by overanswering for the marks available; several were characterised by underanswering for the marks available or being too general, not entering into specifics and not addressing the question.

Specific comments (not all parts deserve comment)

1a) lower marks were obtained where no detail was given about how one could classify queries and/or what difficulties might arise.

1c) i) and ii) proved to be excellent distinguishers, and served to separate those who had apparently memorised aspects of the Boolean model (for 1d) from those who could apply Boolean logic with an understanding of the nature of term distribution in document collections. Many candidates advanced highly impractical and highly inefficient/redundant solutions, or solutions that would have required going outside the constraints of the question, e.g., relying on direct manipulation of indexes/postings lists, wrongly assuming an Extended Boolean Model was available. Several candidates confused '*' in the EBM with interpretations of this operator in other environments: typically, e.g., in Westlaw, '*' matches a single character only: c*t would match cut, cat, cot; c** would match cut, cap, can, cup, cop, cot,…, etc. Several candidates proposed searching for the space character, which revealed a serious misunderstanding of tokenisation for indexing: whitespace is typically used as one of the delimiters for tokenisation in English, and is therefore not stored as a token itself.

1d) was mostly answered well, although several candidates did not enter into many specifics.

1e) gave rise to several bare assertions whose impact on indexing was not addressed. The open source nature of the tokeniser was not explored by many.

1f) was another excellent distinguisher revealing that, although many candidates could reproduce the IDF formula and could describe the nature of a stop word, they had not fully grasped the impact for ranking for both solutions.

1g i) was in general well tackled, although several candidates did not appear to have a good grasp of how to apply cosine to already normalised vectors.

1h) gave rise to some excellent answers, showing very good evidence of deep learning, of reading around in the topic and of the ability to synthesise knowledge. Several answers remained very general and/or were over-brief for the marks available, or were experiential without referring to techniques underpinning semantic search.

Question 2

This was a very popular question, chosen by the majority of the students. It was composed of 8 parts and included bookwork, application of technique and original thought question styles. The results were positive in general, with the overall average mark being 61%.

The following summarizes what has been observed from students' answers to Question 2:

* A large number of the students failed to fully answer Bookwork questions indicating lack of attention when reading the question, for example, failing to illustrate answers with examples, when explicitly required by the question and failing to provide a precise definition for the concept being explicitly asked about.

* A large number of the students failed to fully answer Application of Technique questions, providing only partial calculations/solutions/comments. Sometimes, using a model different to the one being explicitly requested in the question.

* A small minority of the students was able to correctly and completely answer the original thought questions.

It seems that the quality of the revision done by a significant number of the students is quite poor. Perhaps only based on past exam papers, and perhaps only based on solutions to past exam papers provided by a small group of

students which make the effort to properly solve the questions and which share their answers with all the other students to have them merely memorizing the answers. And this can also be noticed when students provide answers that to not correspond to the question being asked.

Question 3

A third of students took this question and they answered it reasonably well overall (average mark of 63%, with std. dev. of 15%).

The first three parts were mostly straightforward: part (a) was answered correctly in most cases (although not everyone provided examples), as was part (b). Part (c) was done by almost all students.

Part (d) was less well answered, with a number of answers not specifying what RDF(S) is used for. Some answers on RDF(S) shortcoming were vague. However, most answers were correct when defining the required class and property.

Part (e) was generally OK when it comes to the SPARQL part, but there were (again) some issues with the casting operator (it has to interpret the string in the query as being of the date type).

Finally, in part (f), some answers failed to specify benefits of and issues with ontologies and/or SKOS, mainly discussing examples of resources for say health-related issues or environment(!). SKOS was generally discarded (with no justification) as a potential representation model, failing to recognise the importance of probably existing vocabularies that local authorities already are likely to have. Weak answers for the final part were providing vague comments on citizens being able to easily (?) and efficiently access the data (?), and local authorities struggling to deal with broken links between data (?). Most answers failed to highlight the benefits of linked data in terms of data integration and provision of data in a single, open format.