COMP38211 – exam feedback 2019/2020

Q1a. This was quite an easy and straightforward question. Almost everyone answered it correctly, with two (wrong) answers referring to a physical location ("how to get to London").

Q1b. In addition to obvious (correct) answers (domain-specific tokenisation, stop-words, language models), some (partial) answers referred to document classification (to determine the domain) without pointing how that would be done (certainly not manually – as a couple of answers suggested). There were few answers that (wrongly) confused 'domain of the document' with the URL domain.

Q1c. While stemming was easy to describe, some answers were quite superficial and short (e.g. nor mentioning over/under-stemming). It was expected that the answers would compare what would happen when stemming was performed to what would have happened if it wasn't – would stemming lower precision/recall? The question specifically asked for justification of the answers for the two statements, rather than just saying which one was true/false. Some wrong answers claimed that either of the metrics would stay constant. Please also note the wording: 'never lowers' (rather than 'increases').

Q1d. This question required a combination of 'bookwork' (describing tf*idf) and some deeper discussions – comparison and contrasting between Luhn's hypothesis and td*idf. The key part (which was missed by the majority) was that the former focuses on characterisation of the entire dataset, whereas the latter on indexing of individual documents. That's where the differences came from. Only describing tf*idf would yield only half of the marks. Some answers failed to describe how tf*idf was used for indexing.

Q1e. The first part of the question (issues with BoW) was relatively easy and straightforward, although not all answers pointed to the key issues (losing context, multi-word terms, negation, etc.). Some answers discussed the sparsity of the VSM, but then were unable to discuss how n-grams could help with that - claims that n-grams reduce the dimensionality are wrong. Many (almost all) answers failed to mention/describe how n-gram language models could model word sequences (using n-gram conditional probabilities and counts) – and that was explicitly the part of the question.

Q1f. This question was somehow a continuation of the Q1e (with unigrams). Some answers only answered the practical part (often correctly), without answering the first part (how to rank documents, what is the P(q|d), how are these estimated – MLE, the mixed model etc.). As a consequence, some of the answers failed to properly calculate the query probabilities, despite properly estimating term probabilities – these should be multiplied, not added!

Q1g. This was a big disappointment. The majority of answers calculated the overall precision and recall values (at the point where *all* 10 documents are retrieved), rather than the precision and recall values at k (for k = 2, 3, 4, 8). No marks were earned for just overall precision/recall. A small number of answers calculated the recall values not out of all 5 relevant documents, but out of the number that was maximally *possible* at the given k (e.g. at k = 3, the system could return only up to 3 (relevant) documents). If properly explained and argued for, these answers were accepted.

Q2a. This question was quite straightforward and was answered well by most students, with only one or two students losing a mark for not demonstrating graph merging (i.e., having duplicated nodes).

Q2b. Many students (around half of the cohort) lost marks for two common reasons: (1) not representing the prod:SFPNYogurt a food:Food; triple in their graphs; and (2) not representing Maltodextrin as a blank node connected to prod:SFPNYogurt (via a food:containsIngredient predicate); this blank node should then be the subject of two literal triples: one with food:Ingredient as the object, the other with "Maltodextrin" as the object.

Q2c. Majority of the students answered this question correctly. Some students lost marks for missing the rdfs:subPropertyOf predicate (for defining that “technique” is a type of the “method” property) and/or the rdfs:domain predicate (for defining that “technique” applies only to “Ingredients”).

Q2d. This was a bookwork type of question that was answered by majority of the students quite well. Some students lost marks for simply not knowing what 1-, 2- and 3-star data is (but knowing that 4- and 5-star data is RDF-compliant).

Q2e. This question was answered well by most students, signifying that students seem to have learned well how to interpret SPARQL queries. Many students however somewhat carelessly mentioned that the constraint on publication date is that it should be after the 1st January 2019 (rather than “on or after the 1st January 2019”). These answers were accepted anyway as this question is not meant to test whether students understand comparison operators such as “>=”.

Q2f. This question was answered quite poorly overall. While many students know which Dublin Core (DC) terms should be used in which HTML lines, they did not make use of the “prefix”, “property” and “rel” attributes in conjunction with the DC terms.