57 candidates answered this question. The highest mark was 29/30 (96.66%), the lowest 7/30 (23.33%). The average mark was 21.88/30 (72.93%) and SD 7.1. Several of the questions were generally the subject of over-answering and verbosity.

Q1a and Q1b: were answered verbosely by many, with duplication of material. More examples than were asked for were given in numerous cases for Q1a. Criteria were often given for Q1a (which did not call for them) and repeated for Q1b (their proper place).

Q1ci was generally well answered, the following being among the more noteworthy aspects:
- There was quite a bit of verbosity/over-answering, for example in describing what tokenisation, case-folding, stemming, etc., are. This was not called for, explanation of decisions was called for.
- The token 'surgeon' would not be stemmed by any reasonable stemmer and, if lemmatising, it is clearly already a lemma (dictionary headword).
- A process should apply consistently unless stated otherwise (several instances of stemming being decided upon but not applied consistently, with no accompanying explanation of why).
- If one removes hyphens, this does not automatically mean that they are replaced with whitespace.

[Several candidates referred to 'Norman Springs' as a person, although this did not affect their mark, as the reference is to an event in 1991. As movie buffs will know, Gale Storm is indeed the stage name of a real actress in IMDB's top 100. Again, it was not at all necessary to know this, I simply record this for your information.]

Q1cii was somewhat over-answered by many, who chose to, e.g., set out the binary term-document incidence matrix as well as the inverted index. Several candidates specified document frequency although this is not part of a basic inverted index, which simply gives a mapping from an index term to a postings list. Some candidates did not apply the decisions they said they had taken in Q1ci (or: were vague there about whether or not they had actually taken a decision).

Q1ciii was also generally well answered. Marks were not deducted here if, e.g., the inverted index had been incorrectly constructed, as that would be counting an error twice. That is, the Boolean query was assessed in relation to the candidate's inverted index, there was no ideal answer to be marked against. There was again some over-answering as some candidates specified the order in which parts of the query would be processed. This was irrelevant as efficiency was not in question here.

Q1d was fully answered by only a few. Most candidates referred to under- and over-stemming, but other aspects were generally not addressed.

Q1e proved a good distinguisher, as several candidates failed to correctly characterise IDF. Numerous answers focussed on the effect of log() and confused this with the core concept of IDF. Several candidates wrongly thought DF/IDF related to the number of times a term appears in a document.

Q1f was also a reasonable distinguisher, as several candidates thought that the higher the score, the less similar two documents were.

Q1fi was well answered, although some under-answered or were vague: "the norm is used carry out normalisation". If you can match the question and your answer to this pattern: "What does a graphic equalizer do? It equalises the graphics." (Alas Smith and Jones), then you should re-write your answer to be more informative.

Q1g 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. Nevertheless, I was able to award full marks for some of these, because I could see, e.g., that a) they had the correct initial values, b) iteration 1 had been correctly applied, so the algorithm was understood and c) what I took to be an evident slip had been made in iteration 2, typically failing to add one value coming from some node, although the value for that outgoing link had been correctly calculated. I also awarded not-quite-full marks where someone had assumed a wrong initial node value (typically 1 instead of 1/5 = 0.20) but then done everything else correctly. One candidate noted they "could not remember the formula". This was not necessary and would in fact have been inappropriate, as the formula is more complex than the basic algorithm asked for, as it considers factors for random surfing and random jumping. I had noted in the revision lecture slides that I would not ask anyone to remember a formula, if needed I would supply it.

Q1h was the subject of some confused or vague answers. In particular, the action to be taken was not well handled by many.

Q1i was also a good distinguisher. Answers generally fell into one of two camps: ones that paid appropriate attention to text data, context, and the number/type of tokens (for example, whether a token was in lower or upper case, a digit, a comma, etc.) when specifying patterns and ones that gave general, vague patterns that did not match anything (or anything much) in the text. For example, a DATE pattern that specified "DD/MM/YY" would not match anything in the text, but one that specified a sequence of, e.g., "1 or 2 digits, 'th', a capitalised token, a comma, followed by 4 digits" would be doing well. A pattern that recognised that month names could be listed in a resource and so replaced 'capitalised token' with month_name (say) would be doing even better. Many candidates specified that a PERSON could be preceded by a title, but either left it at that (no indication of how the actual name would be
In part (g), many answers failed to clarify what the role of RDFS and OWL was and how they can be used to different steps.

Part (f) was answered correctly but not always with the full explanation as what the query aimed to address in complicated!

(e.g. RDFa) are exclusively for human consumption, or that some formats are difficult to parse as XML is serialisations (e.g. there were claims that certain serialisations cannot represent linked data completely, or that some

Part € was mainly answered reasonably, although there were some errors and misunderstanding of the role of a COMPANY entity might end with the token 'Company' or that an ORGANISATION might end with the token 'Agency' was not often taken advantage of in proposed patterns. 'Hydrogen' is not a SPECIES. An ARTEFACT covers more than something with 'tm' after it, for example, a particular variety of aircraft would certainly count as an ARTEFACT.

Still, there were some excellent discussions on when to use more expressive schemas and what the overheads are.

In general, over-answering and verbosity were features of especially the earlier parts of this question. In several cases, this approached brain-dumping. It is absolutely wasteful of time to repeat the question at the beginning of an answer. It is wasteful to write, e.g., "The first point in answer to this question that I would like to make is that the process of tokenisation..." when, e.g., "Firstly, tokenisation..." would concisely do. Such verbosity appeared to be connected, in part, to writing as one would speak. While I can appreciate the desire to cover as many bases as possible in the hope of obtaining a full mark, candidates who expand into verbosity risk running out of time. Conciseness under exam conditions is preferable (I leave it as an exercise for the reader to determine whether 'concision' (9 characters) would be more concise than 'conciseness' (11 characters) in this sentence).

Q2. SS

The average mark for question 2 was of about 60% having 40 students attempting it.

The questions that were better answered by the students were the bookwork questions, for which sensible answers could be built using the material provided on the lecture notes. Questions that required views and opinions that would go beyond the lecture notes, but which were could be found on the suggested reading material were answered well by a smaller fraction of the students. This suggests that a significant number of the students did not read from the suggested reading material.

Parts a) and b) require an answer containing arguments defending the advantages of Web services and possibly arguments pointing out the disadvantages of Web services compared to previous approaches as well as in general. While these have been discussed in class and arguments can be found from the lecture notes and suggested reading material, some students were unable to provide convincing, complete and clear arguments. Some answers only contained a single argument being rephrased several times as to look as if there were several arguments.

Part c) requires application of knowledge and problem solving skills. Most students answered this question well, even though very few obtained full marks, being penalized for lack of completeness.

Part d) was a simple bookwork question that was well answered by most students.

Part e) requires application of knowledge on the impact of quality of service on Web users. Most students answered this question very well, indicating that students were able to construct a sensible answer by drawing from their own experience.

Parts f), g) and h) combine bookwork and independent thinking on the topic of cloud computing and services, requiring from the students analysis and memory. Answers to these questions lacked, in many cases, completeness, clarity, as well as analysis.

Q3. GN

Around 40% of students took this question, with an average mark of 58%. Parts (a)-© were straightforward bookwork, but were surprisingly not correctly answered by many students – in some cases the answers contained only examples (e.g. file formats) without specifying what they are to illustrate.

Part (d) was easy for listing most triples contained in the example, but some students struggled with the last expression which contained three triples (using a blank node).

Part € was mainly answered reasonably, although there were some errors and misunderstanding of the role of serialisations (e.g. there were claims that certain serialisations cannot represent linked data completely, or that some (e.g. RDFa) are exclusively for human consumption, or that some formats are difficult to parse as XML is complicated!).

Part (f) was answered correctly but not always with the full explanation as what the query aimed to address in different steps.

In part (g), many answers failed to clarify what the role of RDFS and OWL was and how they can be used to introduce “structure” in the data.

Still, there were some excellent discussions on when to use more expressive schemas and what the overheads are. Similarly, there were some interesting and useful insights in possible impact of linked data on a company in part (h).

In some cases, reasonable discussions included what type of data should be published. 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, not to linked data specifically (e.g. security concerns).