UNIVERSITY OF EDINBURGH

COLLEGE OF SCIENCE AND ENGINEERING

SCHOOL OF INFORMATICS

INFORMATICS 1 — DATA & ANALYSIS

Deadline: 4pm Thursday 23 March 2017

Submit to box outside ITO office in Forrest Hill

This paper is based on questions from past exams in Data & Analysis. It is being released on Thursday 9 March 2017 as a written coursework assignment. You have **two weeks** to complete this assignment. It will not necessarily take that long, but the time is there to help you schedule against other assignment loads from your different courses. The original exam time was two hours.

Questions 1 and 2 use only material already covered in the course so far. Question 3 requires material that will be covered in Lectures 15 and 16 during Week 8 of semester. The real exam is based on content from throughout the lecture course.

Submit your solutions on paper to the labelled box outside the ITO office in Forrest Hill by **4pm Thursday 23 March 2017**. Please ensure that all sheets you submit are firmly stapled together, and on the first page write your name, matriculation number, tutor name, tutorial group number, and the course code INF1-DA. If these are not clearly stated then your work will not reach your tutor and may not be marked.

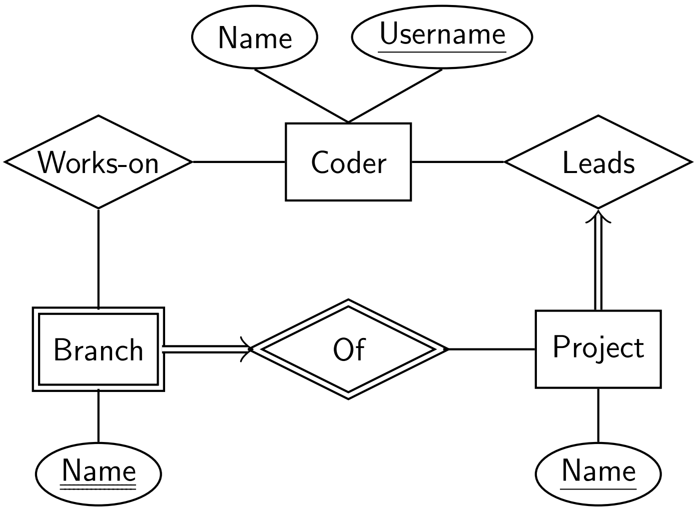
Your tutor will mark your work and return it to you in your Week 11 tutorial, with written and verbal feedback. However, these marks will not affect your final grade for Inf1-DA — this *formative* assessment is entirely for your feedback and learning. Because of this you can freely share help on the questions, discuss on *Piazza*, and talk about your work with other students. Please do.

INSTRUCTIONS TO CANDIDATES

1. **Note that ALL QUESTIONS ARE COMPULSORY.**
2. **DIFFERENT QUESTIONS MAY HAVE DIFFERENT NUMBERS OF TOTAL MARKS. Take note of this in allocating time to questions.**
3. **CALCULATORS MAY BE USED IN THIS EXAMINATION.**

1. [*This question is worth a total of 30 marks.*]

The following entity-relationship diagram captures information about a number of *coders* who work together on a wide range of software *projects*. A project may have a number of different *branches* under active development at any one time. Branches are named, often using standard descriptions such as “master”, “stable”, or “unstable”.

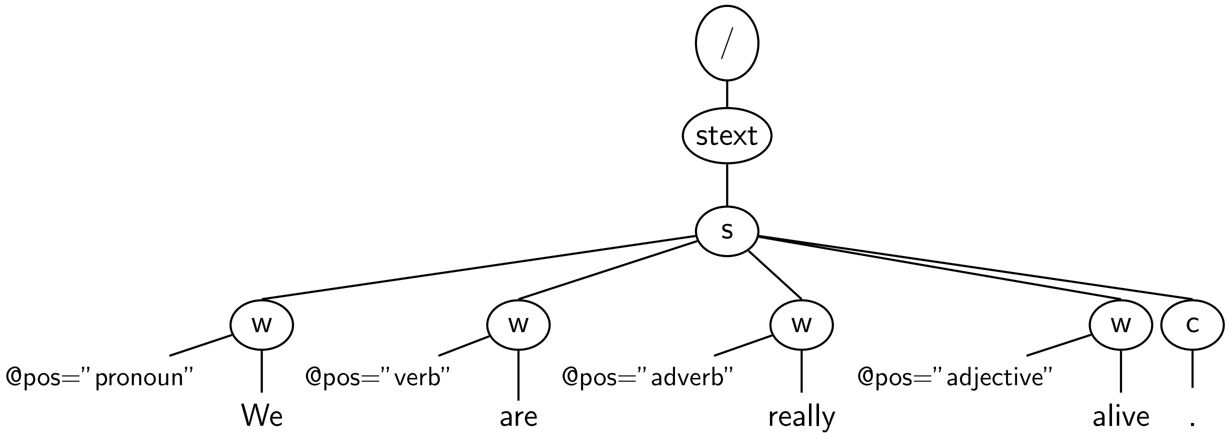


(a) What is the meaning of the double line around Branch? Why is this needed? What is the primary key for a Branch? [*6 marks*]

(b) Construct SQL data declarations for a set of tables to represent this entity-relationship diagram. Assume that usernames are limited to 64 characters and all other names fit within 200 characters. Make sure to use **not null** where necessary; however, you do not need to include **on delete** declarations. [*24 marks*]

2. [*This question is worth a total of 40 marks.*]

The following tree shows the XPath data model for a short XML document. It is a line of spoken text annotated using a simplified version of the British National Corpus mark-up scheme.



(a) Write out this tree as an XML document. [*9 marks*]

(b) This tree contains examples of all four of the main node types in XML v1.0. For example, the "/" at the top is a *root node*. Name the three other types of node in this tree, giving examples of each. [*3 marks*]

(c) In this mark-up scheme stext denotes spoken text, s indicates a sentence, w is a word and c punctuation. Every word is annotated with an appropriate part of speech (pos), taken from a long list of possibilities. A piece of spoken text may contain one or more sentences. Each sentence may contain words and punctuation but must begin with a word and end with punctuation.

Write a DTD to specify these constraints on the XML structure of a spoken text document. [*14 marks*]

(d) Write XPath expressions to return the following lists of text strings from any XML document that satisfies this mark-up scheme.

(i) All punctuation marks used.

(ii) All verbs.

(iii) Every adverb used in any sentence that uses an exclamation mark "!". [*10 marks*]

(e) The document above is just one spoken line. Standard resources for linguistic research like the British National Corpus bring together work from many sources. Building such corpora requires *balancing* and *sampling* to ensure that they are representative. Explain the meaning of balancing and sampling here. [*4 marks*]

3. [*This question is worth a total of 30 marks.*]

The non-existent start-up company *Find-a-Flick* has a website that makes film recommendations. The founders plan to monetise this some day by selling popcorn in flavours matched to individual movies through a deep-dive adaptive data-learning algorithm. For the moment, the site just suggests films based on keywords provided by a user.

For each of around 200,000 films, Find-a-Flick has a body of text built up from reviews, plot descriptions, and comments about the film. By counting occurrences of keywords in the text, the website makes recommendations of films to match user queries.

(a) The performance of an information retrieval system like Find-a-Flick can be evaluated in terms of its *recall* and *precision.* Informally, recall is the proportion of relevant results that are actually retrieved. Give a similar informal definition of precision.

(b) Which is more important for the Find-a-Flick service: recall or precision? Give a reason for your choice.

[*5 marks*]

The following table shows the keyword counts for text associated with three different films, computed to assist a search for "Exciting Scottish historical drama".

	Exciting	Scottish	Historical	Drama
Film A	30	15	0	10
Film B	2	2	4	1
Film C	0	0	4	3
Query	1	1	1	1

One way to identify which films are most relevant to the query is the *cosine similarity measure*, based on the *vector space model* of documents.

(c) Write out in full the formula for calculating the cosine of the angle α between the two four-dimensional vectors (x_1, x_2, x_3, x_4) and (y_1, y_2, y_3, y_4).

(d) Use this to rank these three films by relevance to the original query.

(e) The Find-a-Flick repository happens to have much more text about Film A than it does about either Film B or Film C — this is why the keyword counts are much higher for that film. Does this affect the ranking of Film A? If so, does it make it higher or lower? If not, why not?

[*15 marks*]

QUESTION CONTINUES ON NEXT PAGE

QUESTION CONTINUED FROM PREVIOUS PAGE

The Find-a-Flick team are testing out two possible information retrieval systems: *Hare* and *Tortoise*. These are being evaluated on a small test collection of just 4000 documents, with a single query for which there are 200 relevant documents. *Hare* returns 1200 documents from the collection, including 150 that are relevant; while *Tortoise* returns just 160, with 120 of them being relevant.

(f) Tabulate the results for each system and calculate their precision and recall on this test. Show your working.

(g) One way to combine precision and recall scores is to use their *harmonic mean*. Give the formula for this, and calculate its value for each of *Hare* and *Tortoise*.

[*10 marks*]