INFS3200 Advanced Database Systems

Tutorial 7: Data Quality Management

Semester 1, 2021

Question 1: Explain the following data quality dimensions with examples.

- Integrity
- Accuracy
- Representational consistency
- Completeness
- Timeliness
- Accessibility

What data quality problems, e.g., inaccuracy, inconsistency, duplication, etc. can you find from the following tables?

Surname	**Firstname**	**Age**	**Major**	**Degree**	**Sex**
Barratt	John	22	Maths	BSc	Male
Burns	Robert	24	CS	BSc	Male
Carter	Laura	20	Physics	MSc	Female
Davies	Michael	12	CS	BSc	Male

(a) Student Table

Surname	**Firstname**	**DoB**	**Driving test passed**
Smith	J.	17/12/85	12/12/05
Smith	Jack	17/12/85	12/12/2005
Smith	Jock	17/12/95	12/12/2005

(b) Driving Test Table

Question 2: In a company, the *reference table* consists of the clean data on the surname and given name of its employees. Given a set of input records, the duplicates can be identified by checking against the reference table. In other words, two input records are regarded as duplicates if they are linked to the same tuple in the reference table. Suppose that we use Edit distance as the similarity measure and use dynamic programming matrix to calculate Edit distance between two strings.

(a) Considering surname only, are there any duplicates in the following input records?

(b) When both surname and given name are considered (with the same importance), are there any duplicates in the following input records?

ID	Surname	Given name
1	Duboice	Nicholas
2	Hansson	William
3	Wang	Sean
4	Dubosh	Nick
5	Garnette	Kevin
6	Pitt	William
7	Hanschen	Williams
8	Aniston	Jennifer
9	Zhou	Xuan

Reference

ID	Surname	Given name	Expense
001	Duboise	Nicholas	130
002	Duboch	Nicolas	30
003	Hanson	Williams	47

Input records

---ooo000O000ooo---