Scalable Data Engineering, Exercise 4

Henry Haustein

Task 1

(a) True

(b) False, they can't be proven algorithmically.

(c) True

Task 2

The bigrams are

- firstname: f, fi, ir, rs, st, tn, na, am, me, e
- lastname: l, la, as, st, tn, na, am, me, e
- street: s, st, tr, re, ee, et, t
- name: n, na, am, me, e
- address: a, ad, dd, dr, re, es, ss, s
- forename: f, fo, or, re, en, na, am, me, e

Then the similarities are

	firstname	**lastname**	**street**
name	$\frac{2\cdot 4}{10+5} = 0.533$	$\frac{2\cdot 2\cdot 4}{9+5} = 0.571$	$\frac{2\cdot 0}{7+5} = 0$
address	$\frac{2\cdot 0}{10+8} = 0$	$\frac{2\cdot 0}{9+8} = 0$	$\frac{2\cdot 1}{7+8} = 0.133$
forename	$\frac{2\cdot 5}{10+9} = 0.526$	$\frac{2\cdot 4}{9+9} = 0.444$	$\frac{2\cdot 1}{7+9} = 0.125$

Stable Marriage Algorithm:

- firstname proposes to name, agrees: (firstname, name)
- lastname proposes to name, agrees + leaves: (lastname, name)
- firstname proposes to forename, agrees: (lastname, name), (firstname, forename)
- street proposes to adress, agrees: (lastname, name), (firstname, forename), (street, adress)

Task 3

(a)

(b) Nation ⇔ Country, (Region ⇔ Continent)

(c) SQL:

```
ALTER TABLE mondial.country ADD COLUMN cid UUID

UPDATE mondial.country SET cid = gen_random_uuid()
```

(d) SQL:

```
ALTER TABLE supplier ADD COLUMN s_countrykey UUID

UPDATE supplier SET s_countrykey = c.uuid
FROM nation AS n, mondial.country AS c
WHERE
  lower(n.name) = lower(c.name) AND
  supplier.s_nationkey = nation.n_nationkey
```