Assignment 2: ASQL_ICs_DI

Student: Koray Poyraz
Studentnr: 5367646

Question 1

see file db.sql

Integrity constraints were taken into account when creating the relations.

Department
Building is not unique because it might be that a building could have multiple departments, so building could have his own relation but due to not having specific attributes related to a building its not needed.

Student and Instructor
These relations have set null on delete of a tuple in department because we don't want to delete tuples from student and instructor when a department gets deleted, it might be that a student or instructor will join another department.

Advisor
If a student is deleted or updated then those tuples in the advisor are deleted and updated as well. If a instructor is deleted then we set the value to null so you can have in 1 glance all students who don't have a advisor. Also, it has cascade on update.

Course
I didn't make title a unique attribute because it might be that a department wants to give the same course with same title but a different course id. We could make title and dept_name unique but that would not give the possibility to have the same course with different id, same title within the same department but different credits.

Prereq
When a course is been deleted or updated then the effect will be cascaded to the tuples in prereq relation.

Teaches
When we remove a section then there is no need to have the related teaches tuples.

Section
If there is a new classroom for a section then the update will reflect the change or delete the related tuples on delete. If a course is been deleted then there is no need to have the related tuples in section.

Takes
When a student is deleted or updated the change will be reflected on the related tuples in takes. If a section is deleted than the related tuples in takes will be deleted as well and on update the change will be cascaded.

Question 2

see file inserts.sql

- 1st turn on

pragma foreign_keys = on;

- 2nd add departments

insert into department values ('CS', 'Computer Science', 10000);

insert into department values ('DS', 'Data Science', 10000);

select * from department;

- Add 4 students to student relation

insert into student (name, dept_name, tot_cred) values ('Jan', 'CS', 10);

insert into student (name, dept_name, tot_cred) values ('Kees', 'DS', 8);

insert into student (name, dept_name, tot_cred) values ('Peter', 'CS', 7);

insert into student (name, dept_name, tot_cred) values ('Klaas', 'DS', 9);

select * from student;

- Add 4 instructors to instructor relation

insert into instructor (name, dept_name, salary) VALUES ('Daalon', 'CS', 2800);

insert into instructor (name, dept_name, salary) VALUES ('Henk', 'DS', 2800);

insert into instructor (name, dept_name, salary) VALUES ('Jeffrey', 'CS', 3500);

insert into instructor (name, dept_name, salary) VALUES ('Rick', 'DS', 2800);

select * from instructor;

- Add 4 advisors to advisor relation

insert into advisor (s_id, i_id) VALUES (1, 2);

insert into advisor (s_id, i_id) VALUES (2, 1);

insert into advisor (s_id, i_id) VALUES (3, 4);

insert into advisor (s_id, i_id) VALUES (4, 3);

select * from advisor;

- Display student names and their advisor name

select s1.name student_name, s2.name advisor_name from advisor s inner join student s1 on s.s_id = s1.ID inner join instructor s2 on s.s_id = s2.ID;

Question 3

R(A,B,C,D,E,F,G)

Functional dependencies

- ABD -> EG
- C -> DG
- E -> FG
- AB -> C
- G -> F

Closure

- ABCDEFG+ -> (ABCDEFG)
  - ABD can determine EG so we can discard EG (Reflexivity)
  - and because E and G determines F we can also discard F (Transitivity)
- ABCD+ -> (ABCDEFG)
  - C can determine DG so we can discard DG (Reflexivity)
- ABC+ -> (ABCFEGD)
  - AB can determine C so we can discard C (Reflexivity)
- AB+ -> (ABCDGEF)

The proper subset of AB+ is {A},{B}, A and B are not super key
so we can say that AB is the candidate key.

Candidate key

- AB

Question 4

Functional dependency violation

First name -> Gender

In [1]:
```python
import pandas as pd
df_data = pd.read_csv('../assignment_two/fdExample.csv')
```

In [3]:
```python
print('Violation of functional dependency:')

df_name_gender = df_data[['First name', 'Gender']].rename(columns={'First name': 'first_name'})

# Version 1: shows for each name a set of different genders (key x values)
# Below I have a code which shows the first name -> genders values which are violated, total is 52 first names
# the code could be more simplified by just using the unique() and intersection() methods
# but I found it important to show both columns values

violated_values = df_name_gender\
    .groupby('first_name')\
    .apply(lambda x: ','.join(x['Gender']))\
    .reset_index()\
    .rename(columns={0:'count_diff_gender'})

count_names = 0
for row in violated_values.values:
    set_val = set(row[1].split(','))
    if len(set_val) > 1:
        count_names += 1
        # currently I only need count_names as extra info together with version 2
        # but (see latest code line of this code cell)
        # you can uncomment the code line below to see the unique names and
        # their related multi genders printed
        # print(f' First name: {row[0]} \t - Gender: {set_val}')

# print(f'Violated: {count_names}')


# Version 2: shows all tuples including the row number, the name and gender which violate
key_map = {}

for index, x in enumerate(df_name_gender.values):
    if x[0] not in key_map:
        key_map[x[0]] = [f'{index}:{x[1]}']
    else:
        idx_of_name = key_map[x[0]]
        idx_of_name.append(f'{index}:{x[1]}')

count = 0

for name in key_map:
    genders = map(lambda x: x.split(':')[1], key_map[name])
    is_violated = len(set(genders)) > 1

    if is_violated:
        for gender in key_map[name]:
            count += 1
            gender_idx = gender.split(':')
            print(f'Row number: [{gender_idx[0]}] \t name: {name} \t gender: {gender_idx[1]}')

        print('-------------------------------------------------------------------')

print(f'Number of first names which violate:\t {count_names}')
print(f'Number of tuples which violate:\t {count}')
```

```
Violation of functional dependency:
Row number: [1]          name: David     gender: M
Row number: [144]        name: David     gender: M
Row number: [187]        name: David     gender: M
Row number: [201]        name: David     gender: M
Row number: [202]        name: David     gender: M
Row number: [203]        name: David     gender: M
Row number: [300]        name: David     gender: M
Row number: [326]        name: David     gender: M
Row number: [346]        name: David     gender: M
Row number: [357]        name: David     gender: M
Row number: [396]        name: David     gender: M
Row number: [460]        name: David     gender: M
Row number: [546]        name: David     gender: M
Row number: [601]        name: David     gender: M
Row number: [652]        name: David     gender: M
Row number: [653]        name: David     gender: M
Row number: [768]        name: David     gender: M
Row number: [799]        name: David     gender: M
Row number: [845]        name: David     gender: M
Row number: [874]        name: David     gender: M
Row number: [881]        name: David     gender: M
Row number: [917]        name: David     gender: M
Row number: [968]        name: David     gender: M
Row number: [969]        name: David     gender: M
Row number: [1086]       name: David     gender: M
Row number: [1089]       name: David     gender: M
Row number: [1289]       name: David     gender: M
Row number: [1290]       name: David     gender: M
Row number: [1314]       name: David     gender: M
Row number: [1532]       name: David     gender: M
Row number: [1547]       name: David     gender: M
Row number: [1569]       name: David     gender: M
Row number: [1626]       name: David     gender: M
Row number: [1673]       name: David     gender: M
Row number: [1691]       name: David     gender: M
Row number: [1702]       name: David     gender: M
Row number: [1761]       name: David     gender: M
Row number: [1764]       name: David     gender: M
Row number: [1896]       name: David     gender: M
Row number: [1958]       name: David     gender: M
Row number: [1990]       name: David     gender: M
Row number: [2007]       name: David     gender: M
Row number: [2061]       name: David     gender: M
Row number: [2070]       name: David     gender: M
Row number: [2130]       name: David     gender: M
Row number: [2173]       name: David     gender: M
Row number: [2212]       name: David     gender: M
Row number: [2253]       name: David     gender: M
Row number: [2517]       name: David     gender: M
Row number: [2592]       name: David     gender: M
Row number: [2603]       name: David     gender: M
Row number: [2649]       name: David     gender: M
Row number: [2755]       name: David     gender: M
Row number: [2940]       name: David     gender: M
Row number: [3013]       name: David     gender: M
Row number: [3100]       name: David     gender: M
Row number: [3103]       name: David     gender: M
Row number: [3119]       name: David     gender: M
Row number: [3237]       name: David     gender: M
Row number: [3240]       name: David     gender: M
Row number: [3355]       name: David     gender: M
Row number: [3530]       name: David     gender: M
Row number: [3557]       name: David     gender: M
Row number: [3584]       name: David     gender: M
Row number: [3601]       name: David     gender: M
Row number: [3627]       name: David     gender: M
Row number: [3640]       name: David     gender: M
Row number: [3644]       name: David     gender: M
Row number: [3688]       name: David     gender: M
Row number: [3805]       name: David     gender: M
Row number: [3904]       name: David     gender: M
Row number: [4077]       name: David     gender: M
Row number: [4185]       name: David     gender: M
Row number: [4233]       name: David     gender: M
Row number: [4238]       name: David     gender: M
Row number: [4293]       name: David     gender: M
Row number: [4386]       name: David     gender: M
Row number: [4453]       name: David     gender: F
Row number: [4517]       name: David     gender: M
Row number: [4533]       name: David     gender: M
Row number: [4606]       name: David     gender: M
Row number: [4618]       name: David     gender: M
Row number: [4650]       name: David     gender: M
Row number: [4734]       name: David     gender: M
Row number: [4769]       name: David     gender: M
Row number: [4770]       name: David     gender: M
Row number: [4962]       name: David     gender: M
Row number: [5033]       name: David     gender: M
Row number: [5094]       name: David     gender: M
Row number: [5141]       name: David     gender: M
Row number: [5221]       name: David     gender: M
Row number: [5361]       name: David     gender: M
Row number: [5431]       name: David     gender: M
Row number: [5476]       name: David     gender: M
Row number: [5620]       name: David     gender: M
Row number: [5691]       name: David     gender: M
Row number: [5792]       name: David     gender: M
Row number: [5840]       name: David     gender: M
Row number: [6019]       name: David     gender: M
Row number: [6042]       name: David     gender: M
Row number: [6138]       name: David     gender: M
Row number: [6207]       name: David     gender: M
Row number: [6269]       name: David     gender: M
Row number: [6270]       name: David     gender: M
Row number: [6355]       name: David     gender: M
Row number: [6387]       name: David     gender: M
Row number: [6403]       name: David     gender: M
Row number: [6408]       name: David     gender: M
Row number: [6480]       name: David     gender: M
Row number: [6539]       name: David     gender: M
Row number: [6541]       name: David     gender: M
Row number: [6551]       name: David     gender: M
Row number: [6554]       name: David     gender: M
Row number: [6560]       name: David     gender: M
Row number: [6574]       name: David     gender: M
Row number: [6604]       name: David     gender: M
Row number: [6619]       name: David     gender: M
Row number: [6756]       name: David     gender: M
Row number: [6821]       name: David     gender: M
Row number: [6862]       name: David     gender: M
Row number: [6897]       name: David     gender: M
Row number: [6965]       name: David     gender: M
Row number: [7073]       name: David     gender: M
Row number: [7087]       name: David     gender: M
Row number: [7104]       name: David     gender: M
Row number: [7111]       name: David     gender: M
Row number: [7231]       name: David     gender: M
Row number: [7260]       name: David     gender: M
Row number: [7302]       name: David     gender: M
Row number: [7350]       name: David     gender: M
Row number: [7380]       name: David     gender: M
Row number: [7394]       name: David     gender: M
Row number: [7456]       name: David     gender: M
Row number: [7558]       name: David     gender: M
Row number: [7567]       name: David     gender: M
Row number: [7643]       name: David     gender: M
Row number: [7644]       name: David     gender: M
Row number: [7822]       name: David     gender: M
Row number: [7840]       name: David     gender: M
Row number: [7862]       name: David     gender: M
Row number: [7909]       name: David     gender: M
Row number: [7997]       name: David     gender: M
Row number: [8007]       name: David     gender: M
Row number: [8039]       name: David     gender: M
Row number: [8409]       name: David     gender: M
Row number: [8512]       name: David     gender: M
Row number: [8541]       name: David     gender: M
Row number: [8548]       name: David     gender: M
Row number: [8634]       name: David     gender: M
Row number: [8683]       name: David     gender: M
Row number: [8779]       name: David     gender: M
Row number: [8797]       name: David     gender: M
Row number: [8943]       name: David     gender: M
Row number: [8987]       name: David     gender: M
Row number: [8992]       name: David     gender: M
Row number: [9010]       name: David     gender: M
Row number: [9098]       name: David     gender: M
-------------------------------------------------------------------
Row number: [86]         name: Jean      gender: M
Row number: [291]        name: Jean      gender: F
Row number: [1169]       name: Jean      gender: F
Row number: [1170]       name: Jean      gender: F
Row number: [3717]       name: Jean      gender: F
Row number: [4323]       name: Jean      gender: F
Row number: [4624]       name: Jean      gender: M
Row number: [5359]       name: Jean      gender: F
Row number: [5913]       name: Jean      gender: F
Row number: [6148]       name: Jean      gender: F
Row number: [6467]       name: Jean      gender: F
-------------------------------------------------------------------
Row number: [166]        name: Dana      gender: F
Row number: [1080]       name: Dana      gender: F
Row number: [1095]       name: Dana      gender: F
Row number: [1245]       name: Dana      gender: F
Row number: [2905]       name: Dana      gender: F
Row number: [6242]       name: Dana      gender: F
Row number: [6366]       name: Dana      gender: M
Row number: [7538]       name: Dana      gender: F
Row number: [7932]       name: Dana      gender: F
Row number: [8139]       name: Dana      gender: F
Row number: [8639]       name: Dana      gender: M
Row number: [8795]       name: Dana      gender: F
-------------------------------------------------------------------
Row number: [228]        name: Carey     gender: M
Row number: [245]        name: Carey     gender: M
Row number: [4012]       name: Carey     gender: F
Row number: [6260]       name: Carey     gender: F
-------------------------------------------------------------------
Row number: [231]        name: Dale      gender: M
Row number: [2095]       name: Dale      gender: M
Row number: [2447]       name: Dale      gender: M
Row number: [6363]       name: Dale      gender: F
Row number: [7328]       name: Dale      gender: M
Row number: [7831]       name: Dale      gender: M
Row number: [8193]       name: Dale      gender: M
Row number: [8794]       name: Dale      gender: M
-------------------------------------------------------------------
Row number: [301]        name: Erin      gender: F
Row number: [1474]       name: Erin      gender: F
Row number: [1528]       name: Erin      gender: M
Row number: [2543]       name: Erin      gender: F
Row number: [2547]       name: Erin      gender: F
Row number: [2891]       name: Erin      gender: F
Row number: [3825]       name: Erin      gender: F
Row number: [5721]       name: Erin      gender: F
Row number: [6414]       name: Erin      gender: F
Row number: [6531]       name: Erin      gender: F
Row number: [7090]       name: Erin      gender: F
Row number: [7649]       name: Erin      gender: F
Row number: [7878]       name: Erin      gender: F
Row number: [8483]       name: Erin      gender: F
Row number: [8884]       name: Erin      gender: F
-------------------------------------------------------------------
Row number: [397]        name: Corey     gender: M
Row number: [1545]       name: Corey     gender: M
Row number: [2446]       name: Corey     gender: M
Row number: [2486]       name: Corey     gender: M
Row number: [3543]       name: Corey     gender: M
Row number: [5091]       name: Corey     gender: F
Row number: [6226]       name: Corey     gender: M
Row number: [6667]       name: Corey     gender: M
Row number: [7930]       name: Corey     gender: M
Row number: [8037]       name: Corey     gender: M
-------------------------------------------------------------------
Row number: [427]        name: Stacy     gender: F
Row number: [756]        name: Stacy     gender: M
Row number: [1595]       name: Stacy     gender: F
Row number: [2327]       name: Stacy     gender: F
Row number: [2726]       name: Stacy     gender: M
Row number: [3387]       name: Stacy     gender: F
Row number: [3876]       name: Stacy     gender: M
Row number: [4748]       name: Stacy     gender: F
Row number: [7547]       name: Stacy     gender: F
Row number: [7552]       name: Stacy     gender: F
Row number: [8667]       name: Stacy     gender: F
-------------------------------------------------------------------
Row number: [432]        name: Jamie     gender: F
Row number: [887]        name: Jamie     gender: M
Row number: [4444]       name: Jamie     gender: F
Row number: [5518]       name: Jamie     gender: M
Row number: [6093]       name: Jamie     gender: M
Row number: [6419]       name: Jamie     gender: F
Row number: [7066]       name: Jamie     gender: F
Row number: [7141]       name: Jamie     gender: M
Row number: [8728]       name: Jamie     gender: F
-------------------------------------------------------------------
Row number: [476]        name: Alan      gender: M
Row number: [1140]       name: Alan      gender: M
Row number: [1344]       name: Alan      gender: M
Row number: [3529]       name: Alan      gender: M
Row number: [4062]       name: Alan      gender: M
Row number: [5687]       name: Alan      gender: M
Row number: [5924]       name: Alan      gender: M
Row number: [6262]       name: Alan      gender: F
Row number: [6676]       name: Alan      gender: M
Row number: [6783]       name: Alan      gender: M
Row number: [7042]       name: Alan      gender: M
Row number: [7517]       name: Alan      gender: M
Row number: [7750]       name: Alan      gender: M
Row number: [7765]       name: Alan      gender: M
Row number: [9036]       name: Alan      gender: M
-------------------------------------------------------------------
Row number: [510]        name: Jerry     gender: M
Row number: [2650]       name: Jerry     gender: M
Row number: [2847]       name: Jerry     gender: M
Row number: [3654]       name: Jerry     gender: M
Row number: [4018]       name: Jerry     gender: M
Row number: [5120]       name: Jerry     gender: F
Row number: [5485]       name: Jerry     gender: M
Row number: [6695]       name: Jerry     gender: M
Row number: [7663]       name: Jerry     gender: M
-------------------------------------------------------------------
Row number: [558]        name: Eric      gender: M
Row number: [940]        name: Eric      gender: M
Row number: [974]        name: Eric      gender: M
Row number: [1082]       name: Eric      gender: M
Row number: [1294]       name: Eric      gender: M
Row number: [1316]       name: Eric      gender: M
Row number: [1565]       name: Eric      gender: M
Row number: [2024]       name: Eric      gender: M
Row number: [2343]       name: Eric      gender: M
Row number: [2355]       name: Eric      gender: M
Row number: [2791]       name: Eric      gender: M
Row number: [3038]       name: Eric      gender: M
Row number: [3229]       name: Eric      gender: M
Row number: [3391]       name: Eric      gender: M
Row number: [3439]       name: Eric      gender: M
Row number: [3449]       name: Eric      gender: M
Row number: [3810]       name: Eric      gender: M
Row number: [3818]       name: Eric      gender: M
Row number: [3819]       name: Eric      gender: M
Row number: [4079]       name: Eric      gender: M
Row number: [4405]       name: Eric      gender: M
Row number: [5319]       name: Eric      gender: F
Row number: [5501]       name: Eric      gender: M
Row number: [5574]       name: Eric      gender: M
Row number: [5746]       name: Eric      gender: M
Row number: [5944]       name: Eric      gender: M
Row number: [6442]       name: Eric      gender: M
Row number: [6633]       name: Eric      gender: M
Row number: [6777]       name: Eric      gender: M
Row number: [6924]       name: Eric      gender: M
Row number: [7183]       name: Eric      gender: M
Row number: [7239]       name: Eric      gender: M
Row number: [7591]       name: Eric      gender: M
Row number: [7698]       name: Eric      gender: M
Row number: [7810]       name: Eric      gender: M
Row number: [8101]       name: Eric      gender: M
Row number: [8326]       name: Eric      gender: M
Row number: [8459]       name: Eric      gender: M
Row number: [8522]       name: Eric      gender: M
Row number: [8619]       name: Eric      gender: M
Row number: [8695]       name: Eric      gender: M
Row number: [8699]       name: Eric      gender: M
Row number: [8838]       name: Eric      gender: M
Row number: [8925]       name: Eric      gender: M
-------------------------------------------------------------------
Row number: [576]        name: Lee       gender: M
Row number: [3301]       name: Lee       gender: M
Row number: [3747]       name: Lee       gender: M
Row number: [3901]       name: Lee       gender: M
Row number: [4282]       name: Lee       gender: M
Row number: [5003]       name: Lee       gender: F
Row number: [5497]       name: Lee       gender: M
Row number: [5597]       name: Lee       gender: M
Row number: [7563]       name: Lee       gender: M
-------------------------------------------------------------------
Row number: [587]        name: Leslie    gender: F
Row number: [1511]       name: Leslie    gender: F
Row number: [2966]       name: Leslie    gender: F
Row number: [3390]       name: Leslie    gender: F
Row number: [3588]       name: Leslie    gender: F
Row number: [3662]       name: Leslie    gender: F
Row number: [4049]       name: Leslie    gender: F
Row number: [5332]       name: Leslie    gender: F
Row number: [6873]       name: Leslie    gender: F
Row number: [7156]       name: Leslie    gender: F
Row number: [7784]       name: Leslie    gender: M
Row number: [8501]       name: Leslie    gender: M
Row number: [8591]       name: Leslie    gender: F
Row number: [8776]       name: Leslie    gender: F
Row number: [8936]       name: Leslie    gender: F
-------------------------------------------------------------------
Row number: [670]        name: Rene      gender: M
Row number: [6708]       name: Rene      gender: M
Row number: [6976]       name: Rene      gender: M
Row number: [8825]       name: Rene      gender: F
-------------------------------------------------------------------
Row number: [703]        name: Donald    gender: M
Row number: [719]        name: Donald    gender: M
Row number: [827]        name: Donald    gender: M
Row number: [1189]       name: Donald    gender: M
Row number: [1604]       name: Donald    gender: M
Row number: [1605]       name: Donald    gender: M
Row number: [1897]       name: Donald    gender: M
Row number: [2195]       name: Donald    gender: M
Row number: [2429]       name: Donald    gender: M
Row number: [2479]       name: Donald    gender: M
Row number: [2728]       name: Donald    gender: M
Row number: [3164]       name: Donald    gender: M
Row number: [3494]       name: Donald    gender: M
Row number: [3505]       name: Donald    gender: M
Row number: [3761]       name: Donald    gender: F
Row number: [3777]       name: Donald    gender: M
Row number: [3824]       name: Donald    gender: M
Row number: [4751]       name: Donald    gender: M
Row number: [5675]       name: Donald    gender: M
Row number: [5729]       name: Donald    gender: M
Row number: [6047]       name: Donald    gender: M
Row number: [6063]       name: Donald    gender: M
Row number: [6504]       name: Donald    gender: M
Row number: [6530]       name: Donald    gender: M
Row number: [7344]       name: Donald    gender: M
Row number: [7568]       name: Donald    gender: M
Row number: [7648]       name: Donald    gender: M
Row number: [8173]       name: Donald    gender: M
Row number: [8452]       name: Donald    gender: M
Row number: [8617]       name: Donald    gender: M
Row number: [8618]       name: Donald    gender: M
Row number: [9004]       name: Donald    gender: M
-------------------------------------------------------------------
Row number: [755]        name: Terry     gender: F
Row number: [2318]       name: Terry     gender: F
Row number: [3047]       name: Terry     gender: M
Row number: [3460]       name: Terry     gender: M
-------------------------------------------------------------------
Row number: [854]        name: Taylor    gender: M
Row number: [2322]       name: Taylor    gender: F
Row number: [3228]       name: Taylor    gender: M
Row number: [5259]       name: Taylor    gender: M
-------------------------------------------------------------------
Row number: [891]        name: Shawn     gender: F
Row number: [1709]       name: Shawn     gender: M
Row number: [2736]       name: Shawn     gender: M
Row number: [3097]       name: Shawn     gender: M
Row number: [4407]       name: Shawn     gender: M
Row number: [4505]       name: Shawn     gender: M
Row number: [4713]       name: Shawn     gender: F
Row number: [6638]       name: Shawn     gender: M
Row number: [6835]       name: Shawn     gender: M
Row number: [7025]       name: Shawn     gender: M
Row number: [7554]       name: Shawn     gender: M
Row number: [7893]       name: Shawn     gender: F
-------------------------------------------------------------------
Row number: [901]        name: Frederick         gender: M
Row number: [2442]       name: Frederick         gender: M
Row number: [5290]       name: Frederick         gender: M
Row number: [8189]       name: Frederick         gender: F
-------------------------------------------------------------------
Row number: [955]        name: Andre     gender: M
Row number: [2287]       name: Andre     gender: M
Row number: [2701]       name: Andre     gender: M
Row number: [3570]       name: Andre     gender: M
Row number: [5010]       name: Andre     gender: F
Row number: [5043]       name: Andre     gender: M
Row number: [5368]       name: Andre     gender: M
Row number: [5665]       name: Andre     gender: M
Row number: [5825]       name: Andre     gender: M
Row number: [7846]       name: Andre     gender: M
Row number: [8563]       name: Andre     gender: M
-------------------------------------------------------------------
Row number: [1011]       name: Tracy     gender: F
Row number: [1306]       name: Tracy     gender: F
Row number: [1593]       name: Tracy     gender: F
Row number: [1674]       name: Tracy     gender: M
Row number: [2165]       name: Tracy     gender: F
Row number: [3461]       name: Tracy     gender: M
Row number: [4168]       name: Tracy     gender: F
Row number: [7513]       name: Tracy     gender: F
Row number: [8967]       name: Tracy     gender: M
-------------------------------------------------------------------
Row number: [1127]       name: Antwan    gender: M
Row number: [8625]       name: Antwan    gender: F
-------------------------------------------------------------------
Row number: [1177]       name: Jesus     gender: M
Row number: [1985]       name: Jesus     gender: F
Row number: [7227]       name: Jesus     gender: M
-------------------------------------------------------------------
Row number: [1328]       name: Juan      gender: M
Row number: [1781]       name: Juan      gender: M
Row number: [2536]       name: Juan      gender: M
Row number: [2895]       name: Juan      gender: M
Row number: [4255]       name: Juan      gender: M
Row number: [4942]       name: Juan      gender: M
Row number: [5006]       name: Juan      gender: M
Row number: [6318]       name: Juan      gender: M
Row number: [6594]       name: Juan      gender: M
Row number: [7050]       name: Juan      gender: M
Row number: [7229]       name: Juan      gender: M
Row number: [8995]       name: Juan      gender: F
-------------------------------------------------------------------
Row number: [1335]       name: Andrea    gender: F
Row number: [1384]       name: Andrea    gender: F
Row number: [1587]       name: Andrea    gender: F
Row number: [1981]       name: Andrea    gender: M
Row number: [2311]       name: Andrea    gender: F
Row number: [2904]       name: Andrea    gender: F
Row number: [3531]       name: Andrea    gender: F
Row number: [3583]       name: Andrea    gender: F
Row number: [6103]       name: Andrea    gender: F
Row number: [7341]       name: Andrea    gender: F
Row number: [7967]       name: Andrea    gender: F
-------------------------------------------------------------------
Row number: [1426]       name: Sun       gender: M
Row number: [1461]       name: Sun       gender: F
-------------------------------------------------------------------
Row number: [1452]       name: Stacey    gender: F
Row number: [1856]       name: Stacey    gender: F
Row number: [2644]       name: Stacey    gender: F
Row number: [2677]       name: Stacey    gender: F
Row number: [4225]       name: Stacey    gender: M
Row number: [4620]       name: Stacey    gender: F
Row number: [7314]       name: Stacey    gender: F
-------------------------------------------------------------------
Row number: [1648]       name: Torrie    gender: M
Row number: [8631]       name: Torrie    gender: F
-------------------------------------------------------------------
Row number: [1654]       name: Kelly     gender: F
Row number: [2864]       name: Kelly     gender: F
Row number: [3145]       name: Kelly     gender: F
Row number: [4174]       name: Kelly     gender: F
Row number: [5850]       name: Kelly     gender: M
Row number: [5872]       name: Kelly     gender: F
Row number: [6184]       name: Kelly     gender: F
Row number: [6757]       name: Kelly     gender: F
Row number: [8046]       name: Kelly     gender: M
Row number: [8349]       name: Kelly     gender: F
Row number: [8608]       name: Kelly     gender: M
Row number: [9077]       name: Kelly     gender: F
-------------------------------------------------------------------
Row number: [1657]       name: Sidney    gender: F
Row number: [4340]       name: Sidney    gender: M
Row number: [5741]       name: Sidney    gender: M
-------------------------------------------------------------------
Row number: [1738]       name: Shannon   gender: M
Row number: [2541]       name: Shannon   gender: F
Row number: [2672]       name: Shannon   gender: F
Row number: [5325]       name: Shannon   gender: F
Row number: [5905]       name: Shannon   gender: F
Row number: [6669]       name: Shannon   gender: M
Row number: [7244]       name: Shannon   gender: F
Row number: [7604]       name: Shannon   gender: F
Row number: [7742]       name: Shannon   gender: F
-------------------------------------------------------------------
Row number: [1961]       name: Danielle          gender: F
Row number: [2243]       name: Danielle          gender: F
Row number: [2605]       name: Danielle          gender: F
Row number: [2642]       name: Danielle          gender: F
Row number: [3562]       name: Danielle          gender: F
Row number: [4114]       name: Danielle          gender: F
Row number: [6650]       name: Danielle          gender: F
Row number: [7725]       name: Danielle          gender: F
Row number: [8869]       name: Danielle          gender: M
-------------------------------------------------------------------
Row number: [2148]       name: Carroll   gender: M
Row number: [3250]       name: Carroll   gender: M
Row number: [3523]       name: Carroll   gender: M
Row number: [5753]       name: Carroll   gender: M
Row number: [6516]       name: Carroll   gender: F
Row number: [6641]       name: Carroll   gender: M
-------------------------------------------------------------------
Row number: [2194]       name: Deon      gender: M
Row number: [6930]       name: Deon      gender: F
-------------------------------------------------------------------
Row number: [2450]       name: Randy     gender: F
Row number: [2669]       name: Randy     gender: M
Row number: [4596]       name: Randy     gender: M
Row number: [7195]       name: Randy     gender: M
Row number: [7337]       name: Randy     gender: M
-------------------------------------------------------------------
Row number: [2688]       name: Dominique         gender: M
Row number: [4697]       name: Dominique         gender: M
Row number: [7647]       name: Dominique         gender: F
-------------------------------------------------------------------
Row number: [2705]       name: M         gender: M
Row number: [4913]       name: M         gender: F
-------------------------------------------------------------------
Row number: [2887]       name: Alexis    gender: M
Row number: [6820]       name: Alexis    gender: F
-------------------------------------------------------------------
Row number: [2893]       name: Jaime     gender: M
Row number: [3968]       name: Jaime     gender: F
-------------------------------------------------------------------
Row number: [3265]       name: Jin       gender: M
Row number: [4446]       name: Jin       gender: F
-------------------------------------------------------------------
Row number: [3364]       name: Stefan    gender: F
Row number: [9050]       name: Stefan    gender: M
-------------------------------------------------------------------
Row number: [3504]       name: Angel     gender: F
Row number: [6920]       name: Angel     gender: F
Row number: [8427]       name: Angel     gender: M
-------------------------------------------------------------------
Row number: [3882]       name: Noel      gender: M
Row number: [6879]       name: Noel      gender: F
-------------------------------------------------------------------
Row number: [4205]       name: Kerry     gender: M
Row number: [5760]       name: Kerry     gender: M
Row number: [7435]       name: Kerry     gender: F
-------------------------------------------------------------------
Row number: [4276]       name: Hilary    gender: M
Row number: [8079]       name: Hilary    gender: M
Row number: [8871]       name: Hilary    gender: F
-------------------------------------------------------------------
Row number: [4743]       name: My        gender: M
Row number: [8226]       name: My        gender: F
-------------------------------------------------------------------
Row number: [5334]       name: Ronnie    gender: M
Row number: [7023]       name: Ronnie    gender: M
Row number: [8354]       name: Ronnie    gender: F
Row number: [8534]       name: Ronnie    gender: M
Row number: [8593]       name: Ronnie    gender: F
Row number: [9022]       name: Ronnie    gender: M
-------------------------------------------------------------------
Row number: [5527]       name: Yu        gender: F
Row number: [8971]       name: Yu        gender: M
-------------------------------------------------------------------
Row number: [6088]       name: Marion    gender: F
Row number: [8851]       name: Marion    gender: M
-------------------------------------------------------------------
Row number: [6263]       name: Hui       gender: M
Row number: [8979]       name: Hui       gender: F
-------------------------------------------------------------------
Row number: [7947]       name: Maxie     gender: M
Row number: [8675]       name: Maxie     gender: F
-------------------------------------------------------------------
Number of first names which violate:     52
Number of tuples which violate:  534
```

Question 5

Metanome

For this question I used the Metanome tool together with the FastFDs-1.2 and DepMiner-1.2 algorithm.

For FastFD (19 results), please see fd_algorithm_1.png and fd_algorithm_2.png

determinant	dependant
Abbreviation	Name
Unicode	Name
UnicodeDisplay	Name
Represents	Name
Name	AU_Abbreviation
Unicode	AU_Abbreviation
UnicodeDisplay	AU_Abbreviation
Represents	AU_Abbreviation
Name	Symbol
AU_Abbreviation	Symbol
Unicode	Symbol
UnicodeDisplay	Symbol
Represents	Symbol
UnicodeDisplay	Unicode
Represents	Unicode
Unicode	UnicodeDisplay
Represents	UnicodeDisplay
Unicode	Represents
UnicodeDisplay	Represents

For DepMiner (14 results), please see dm_algorithm_1.png and dm_algorithm_2.png

determinant	dependant
IAU_Abbreviation	Name
UnicodeDisplay	Name
Represents	Name
Unicode	Name
Represents	UnicodeDisplay
Unicode	UnicodeDisplay
UnicodeDisplay	IAU_Abbreviation
Represents	IAU_Abbreviation
Name	IAU_Abbreviation
Unicode	IAU_Abbreviation
UnicodeDisplay	Unicode
Represents	Unicode
UnicodeDisplay	Represents
Unicode	Represents

Algorithm FastFD (19 results) has more results than DepMiner (14 results).

Question 6

Computing Levenshtein distance

'*' = empty space

'+' = shows the minimum cost path

1

Computed

.	.	C	o	m	p	l	e	t	i	o	n
.	+0	1	2	3	4	5	6	7	8	9	10
C	1	+0	1	2	3	4	5	6	7	8	9
o	2	1	+0	1	2	3	4	5	6	7	8
m	3	2	1	+0	1	2	3	4	5	6	7
p	4	3	2	1	+0	1	2	3	4	5	6
u	5	4	3	2	+1	1	2	3	4	5	6
t	6	5	4	3	2	+2	2	2	3	4	5
a	7	6	5	4	3	3	+3	3	3	4	5
t	8	7	6	5	4	4	4	+3	4	4	5
i	9	8	7	6	5	5	5	4	+3	4	5
o	10	9	8	7	6	6	6	5	4	+3	4
n	11	10	9	8	7	7	7	6	5	4	+3

result = 3 cost

Handwritten

- Computation
- Comp*letion

d+s+s = 3 operations = 3 cost

2

Computed

.	.	C	o	m	p	l	e	t	i	o	n
.	+0	1	2	3	4	5	6	7	8	9	10
C	1	+0	1	2	3	4	5	6	7	8	9
o	2	1	+0	1	2	3	4	5	6	7	8
m	3	2	1	+0	1	2	3	4	5	6	7
p	4	3	2	1	+0	1	2	3	4	5	6
u	5	4	3	2	+1	2	3	4	5	6	7
t	6	5	4	3	2	+3	4	3	4	5	6
a	7	6	5	4	3	4	+5	4	5	6	7
t	8	7	6	5	4	5	6	+5	6	7	8
i	9	8	7	6	5	6	7	6	+5	6	7
o	10	9	8	7	6	7	8	7	6	+5	6
n	11	10	9	8	7	8	9	8	7	6	+5

result = 5 cost

Handwritten

- Computation
- Comp*letion

d+s+s = 3 operations = i + (2 x 2) = 5 cost

3

Computed

.	.	C	o	m	p	l	e	t	i	o	n
.	+0	1	2	3	4	5	6	7	8	9	10
C	1	+0	1	2	3	4	5	6	7	8	9
o	2	1	+0	1	2	3	4	5	6	7	8
m	3	2	1	+0	1	2	3	4	5	6	7
p	4	3	2	1	+0	+1	+2	3	4	5	6
u	5	4	3	2	1	2	+3	4	5	6	7
t	6	5	4	3	2	3	+4	3	4	5	6
a	7	6	5	4	3	4	+5	4	5	6	7
t	8	7	6	5	4	5	6	+5	6	7	8
i	9	8	7	6	5	6	7	6	+5	6	7
o	10	9	8	7	6	7	8	7	6	+5	6
n	11	10	9	8	7	8	9	8	7	6	+5

result = 5 cost

Handwritten

- Computation
- Comp*letion

d+s+s = 3 operations = i + (2 x 3) = 7 cost

(this is the result if we keep the empty space * between p and l of completion)

- Comp**utation
- Comple***tion

i+i+d+d+d = 5 operations = (i x 1) + (d x 1) = 5 cost

(this is the result if we add more empty space to use less cost)

When the update cost is > 3:
The algorithm is looking for the minimum cost out of the edit distances (insert, substitute, deletion) from the current position [i,j].
So, if the cost for substitution is high but the cost for e.g. insert/ delete in the edit distances at [i,j] position is lower, than insert/ delete operation is applied.

Question 7

Computing the gap distance

- (i) insertion cost = 1
- (o) open gap cost = 1
- (e) extend gap cost = 0.1

'Adv ances in Instrum entation and Control'

'Adv.________ Instrum.___________ Control'

'Adv[i][o][e][e][e][e][e][e][e][e] Instrum[i][o][e][e][e][e][e][e][e][e][e][e][e][e] Control'

- i = 1 + 1 = 2
- o = 1 + 1 = 2
- e = 8 + 12 = 20

cost = i+o+e = 2 + 2 + (20 x 0.1)
cost = 24

Question 8

Compute Jaccard bag similarity and Jaccard Distance

A = {1,1,2,2,5}; B = {1,2,2,2,5,5}; C = {1,2,3,4,5}

A and B

- J bag sim = 4/11 = 0.3636
- J distance = 7/11 = 0.6363

A and C

- J bag sim = 3/10 = 0.30
- J distance = 7/10 = 0.70

B and C

- J bag sim = 3/11 = 0.2727
- J distance = 8/11 = 0.7272

Question 9

Compute the Jaro and Jaro-Winkler similarity

- a r n a b
- u r b a n

Jaro sim

- c=2
- t=0
- l=5

0.6 = (1/3) * ((2/5)+(2/5)+((2-0)/2))

Jaro Winkler

- P=0.1
- L=0

Jw(S1, S2) = JaroSim + P x L x (1-JaroSim)

0.6 = 0,6 + 0,1 x 0 x (0,4)

They give the same result due to not having common prefix in both strings. If e.g. they would have a common prefix then Jaro Winkler would give a better matching result.

Question 10

Cardinality of the set that has 5-shingles

In [208…

```
sentence = 'Many problems can be expressed as finding similar sets'
k = 5

cardinality_result = len(sentence) + 1 - k

print(f'A set with {k}-shingles has {cardinality_result} cardinality')
```

```
A set with 5-shingles has 50 cardinality
```

Question 11

- D1 = {aa, bb, ab , ba}
- D2 = {aa, ac, ca, ba}
- D3 = {ab, ba, ca}

a - matrix representation

matrix representation of shingles documents relationship, the universal set is alphabetically ordered

.	D1	D2	D3
aa	1	1	0
ab	1	0	1
ac	0	1	0
ba	1	1	1
bb	1	0	0
ca	0	1	1

b - signature matrix

Permutation 1

index	P1	D1	D2	D3
1	aa	1	1	0
2	bb	1	0	0
3	ab	1	0	1
4	ba	1	1	1
5	ac	0	1	0
6	ca	0	1	1

Permutation 2

index	P2	D1	D2	D3
1	ca	0	1	1
2	ac	0	1	0
3	ba	1	1	1
4	ab	1	0	1
5	bb	1	0	0
6	aa	1	1	0

Permutation 3

index	P3	D1	D2	D3
1	ac	1	1	0
2	ca	1	0	1
3	ab	0	1	0
4	ba	1	1	1
5	bb	1	0	0
6	aa	0	1	1

signature matrix

.	.	.	.
P1	1	1	3
P2	3	1	1
P3	1	1	2

c - Jaccard similarity

Comparison of pair documents with their signatures by using Jaccard similarity

.	(D1, D2)	(D1, D3)	(D2, D3)
Col/Col	2/6 = 0,33	2/5 = 0,4	2/5 = 0,4
Sig/Sig	2/3 = 0,66	0/3 = 0	1/3 = 0,33

Info on col and sig:

- D1 and D2: on 'Col' similarity is 33% and on 'Sig' the agreement is 66%.
- D1 and D3: on 'Col' similarity is 40% and on 'Sig' the agreement is 0%.
- D2 and D3: on 'Col' similarity is 40% and on 'Sig' the agreement is 33%.