Data Mining Course Project 2021

Problem Description

We have a database of **patients**. A patient is represented as an entity. She/he has an id, and a set of <name,value> pairs that describe the characteristics of the patient.

Patients have a set of **conditions**. A condition is also an entity that has an id and a set of attribute <name,value> pairs. A condition is a medical problem that a patient has. For instance, allergy to wool is a condition. Eye dryness is another condition. Blood pressure of more than 20% is a third condition.

To cope with the medical conditions, doctors have designed **therapies**. Each therapy is also an entity with an id and a set of attribute name-value pairs that describe the characteristics of the therapy. Note that conditions may be temporary or permanent. For instance a flu is temporary or a thyroid may be permanent.

When a patient has a condition, doctors prescribe a therapy. A trial is a therapy that has been applied to a specific patient for a specific condition. The trial is a tuple <t,p,date, params, success>, where t is a therapy, p is a patient, date is the time it was applied, and params is a set of parameters (i.e., pairs of attribute name-value) that indicate how the therapy was applied. Therapies are not always fully successful. They may be even completely unsuccessful. This is why the success field is used to record how successful the trial was. This means that when a patient has a medical problem (i.e., a condition), he/she is associated with a sequence of trials for that problem. Note that the order is important because the success of a trial may depend on what has been tried in the past. Also the success of a trial depends on each patient (what other condition he/she has). It is possible that for a given condition, a patient has no trials. This means that a condition has been identified to the patient but no therapy has so far been tried to deal with the condition.

The medical history of a patient is a list of the trials he/she has gone through and the respective conditions these trials were done.

You are asked to create a system to help doctors suggest the next most prominent therapy for a patient to try a specific condition he or she may have, i.e., a trial.
In other words the input is:

1. A set P of patients, their conditions, and the ordered list of trials each patient has done done for each of his/her conditions (i.e, his/her medical history)
2. A specific patient p_q, his/her conditions, the ordered list of trials he/she has done for each of these conditions (i.e, his/her medical history).
3. A condition c_q

The output is:

1. a therapy th_{ans}

Task

You are asked to design a method that implements a solution to the above problem, materialize it into a program, test its performance and effectiveness and produce a report in which you describe all the above.

You are also asked to create a test dataset for the program. The test dataset should contain a large number of elements. When you provide the dataset, you also provide 3 test cases (without the expected answer)

Last but not least, you are asked to provide the results of 10 test cases (runs of your program) each one containing the input and the output your program produces. The dataset on which you run these test cases and the test cases you will have to try will be provided by the instructor after the delivery of the 1st phase of the project.

Delivery Schedule & Deadlines

5th January 2022 for the dataset, 3 test cases

After this phase, the instructor will provide you with a dataset and a number of test cases.

19th January 2022 for the project (code, report, results)

Dataset

The dataset that you produce to test your data (and is given as input to the program) should be in a JSON format. In particular:

```
{
Conditions : [
        {" id": "Cond1",
        "name": "High Blood Pressure",
        "type": "Blood Pressure", // Attribute specifying what kind of condition it is
        …
        },
        {" id": "Cond2",
        "name": "Heart Arrhythmia",
        …
        },
        ...
],
Therapies: [
        {" id": "Th1",
        "name": "Acetoxybenzoic Acid (Aspirin)",
        "type": "Acid",          // Attribute specifying what kind of therapy it is
        …
        },
        { "id": "Th2",
        "name": "Cough Syrup Quibron",
        "type": "Sugar",
```

```
        …
        },
        ...
],
Patients: [
        { "id": 1,
        "name": "John",
        "conditions": [
                {"id": "pc1",                   // The specific condition of the patient
                "diagnosed": 20210915,          // The time that it was diagnosed
                "cured": 20210915,              // The time that it was cured. NULL if not
                "kind": "Cond1",                // The id of the condition
                },
                ...
        "trials": [
                {"id": tr1,
                "start": 20210915,          // The moment it started
                "end": 20211215,            // The moment it ended
                "condition": "pc1",         // The id of the condition for which it was done
                "therapy": "Th1",           // What therapy was applied
                "successful": "10%",        // How successful was the trial? You can use this
                                            //      as a number between 0 and 100
                                            //      meaning (no % symbol)

                …
                },
                {"id": tr353,
                "start": 20211216,
                "end": 20211229,
                "condition": "pc1",
                "therapy": "Th2",
                "successful": "100%",
                …
                },
                ...
        ]
        …
        },
        {" id": "2",
        "name": "Mary",
        …
        },
        ...
]
}
```

*The items in **bold** are mandatory to exist.*

For the creation of the dataset it is good to look a little real. So you can use real values.

- For patients, you can use this dataset to get names. https://github.com/philipperemy/name-dataset
- For the conditions, use this page to select values that look realistic. The more you use the better. https://www.nhsinform.scot/illnesses-and-conditions/a-to-z
- For therapies, you can use data from here: https://en.wikipedia.org/wiki/List_of_therapies

You can create a script that collects these values or find values elsewhere. Pay attention to the " " (empty space) character.

Project input / output

A test case (or a query patient) is a patient and a condition. The patient is supposed to exist in the dataset (meaning that we already know the medical history of the patient and her/his characteristics. This means that to run the program you need to provide 3 arguments:

1. The dataset
2. The patient id
3. The condition

The output of the program will be an ordered list of 5 recommended therapies (the therapy id and the respective therapy name), with the first being the one most highly advised to perform, and the 5th the least advised.

An example of running the program is:

myprogram dataset.json JohnID headacheID

where JohnID is the id of the patient John, and headacheID is the id of the condition headache. The dataset.json is the json file that contains the data.

Report

The report should contain the following sections:

1) Introduction & Motivation
2) Related Work (max 1 page. You briefly describe the methods you will use, what they do and what is their role. E.g. you describe what clustering does and what techniques exist for clustering.
3) Problem Statement (See the formal model as we defined It in the Rec Systems Lecture for an example). Most of the things in the problem statement are already said in the description of the problem. Do not get distracted and add irrelevant information here like how important the task is etc. These are all statements that go in the introduction.
4) Solution (The actual solution in detail. Note that there is no need for code or specific software component tools description here. Also, you do not explain things already known by the theory, e.g., do not start elaborating on what clustering is and how useful it is. These are all counting negatively.)
5) Implementation (Description of what tools you have used to implement the solution you described above). The difference of this section from the solution is that the solution part describes mathematical formulas, and algorithms. This section says only how the algorithms and the formulas were turned into a system.
6) Dataset (Description of the dataset structure and how you created it)

7) Experimental Evaluation (Perform the necessary steps to illustrate that the method is good – or is not good. You can do this through a user evaluation and through comparison with some base line method. It is up to you to select the base line method. Then you can compare the results and comment on what you observe. You should also care not only about the quality but also about the scalability, i.e., time, related to the size of the data.

[Report Format]
The report should be written in latex using the following template: https://github.com/EDBT2021/Template **The first page should contain apart from the name of the author, the year of studies (e.g. 1st or 2nd), the program (CS, DS, EIT DS, etc.)**
Many people often ask how many pages the report should be. There is no answer to this question, There have been excellent reports that are short and long that are very bad. In general you should aim to include all the information about your project. If someone reads your report you should be able to understand and reproduce your program. Then you know that your report is in a good state. Here are two good reports from previous years. They are on a different topic (and maybe they are a little too long), but they are very good reports and give you a good idea on what is to be expected. [Paper1] [Paper2]

Delivery Components

The project is delivered in different phases
1. Dataset & 3 test cases.
2. Final Delivery (includes report, code, experiments, result dataset, etc)

Delayed assignments:

As per the course syllabus, delivery of the project in the June/July period entails a 10% penalty, and delivering in Aug/Sep a 20% penalty.

Delivery Method

Create a directory on google drive called "YourFirstname YourLastname YourMatricola", in this order and separated with spaces. Share the specific folder with the professor velgias@unitn.it before the delivery deadline of the first phase. The directory should contain the following subdirectories:

1) doc: containing the report
2) src: containing the source code of the project
3) data: containing the dataset
4) bin: the binary file of the execution of the program alongside the results it produces through the various runs you have performed.
5) results: containing the results of the results your program produces on the instructor's test dataset for the provided by the instructor test cases.
6) A README.txt file in which you explain how one can compile your program and how to run it.