TP 8-9 - Classification challenge

The goal of this lab session is to address a classification challenge using Inductive Logic Programming (ILP). The challenge is about discovering good hypotheses from tabular data and positive/negative examples.

1 Data

The Mondial database collects geographical information from the *CIA World Fact Book* and from *Wikipedia* (among others). This database describes countries, provinces, cities, seas, lakes, mountains, etc. It gives numerical attributes such as population, GDP, date of independence, etc. It also gives relationships between entities such as languages spoken, capitals, neighboring countries, etc. For the purposes of this lab session, we will limit ourselves to the following information :

- countries and their type of government ;
- cities, continents, seas, languages ;
- relationships between these entities : neighborhood and dependence between countries, continent and languages spoken for a country, country of belonging and proximity to a sea for cities.

The data is available on the website `https://www.dbis.informatik.uni-goettingen.de/Mondial/`, in various formats (SQL, XML, RDF). For convenience, the data is available on Teams as CSV files. Their headers are shown below.

Task 1 *Extract the useful information from the data files, and represent them as Prolog facts. This implies modeling the data in the form of types and predicates. The production of the facts must of course be automated.*

Several techniques are possible to transform CSV data into into Prolog facts :

- build a Knime workflow ;
- program a script (e.g., in Python) ;
- import the CSV data into a database (ex., phpMyAdmin) then generate the Prolog facts by SQL queries with string interpolation.

2 Goal

The classes that must be learned – each one constituting a challenge to be solved – are provided as two-column CSV files `class_country_X.csv` on Teams. The first column contains the instances to be classified (countries) and the second column contains booleans indicating whether or not each instance belongs to class X. In these files, the countries are represented by their name.

Task 2 *For as many classes as possible, try to produce a set of hypotheses with cplint/Aleph ”explaining” what the positives have in common. This involves defining the hypothesis language (predicates, type and mode of arguments) and then generating Aleph examples from the class files. Here is the list of the different classes in increasing order of difficulty : 1, 7, 3, 4, 8, 2, 6, 9, 10, 5.* You can also build new challenges to share with the rest of the class !

Here are the headers of the available files.

```
-- countries.csv --
"country","population","government","capital"
"Vanuatu","236299","parliamentary republic","Port Vila (Vanuatu)"
...
```

```
-- cities.csv --
"city","population"
"Kathmandu (Nepal)","1003285"
...

-- continents.csv --
"continent"
"Africa"
...

-- languages.csv --
"language"
"Serbo-Croatian"
...

-- seas.csv --
"sea"
"Greenland Sea"
...

-- hasDependency.csv --
"country","dependency (country)"
"Spain","Ceuta"
...

-- class_country_1.csv --
"country","class"
"Vanuatu","false"
...
"Indonesia","true"
...
```

```
-- hasCountry.csv --
"city","country"
"Kathmandu (Nepal)","Nepal"
...

-- hasContinent.csv --
"country","continent"
"Vanuatu","Australia/Oceania"
...

-- hasLanguage.csv --
"country","language"
"Afghanistan","Afghan Persian"
...

-- bordersSea.csv --
"city","sea"
"Georgetown (Pulau Pinang, Malaysia)","Malakka Strait"
...

-- hasNeighbour.csv --
"country1","country2"
"Afghanistan","China"
...
```