TP 8-9 - Challenge de classification

L'objectif de ce TP est de répondre à un challenge de classification par Programmation Logique Inductive (PLI). Il s'agit de découvrir de bonnes hypothèses à partir de données tabulaires et d'exemples positifs/négatifs.

1 Données

La base MONDIAL collecte des informations géographiques issues du *CIA World Fact Book* et de *Wikipédia* (entre autres). Cette base décrit les pays, provinces, villes, mers, lacs, montagnes, etc. Elle donne des attributs numériques telles que population, PIB, date d'indépendance, etc. Elle donne aussi des relations entre entités telles que langues parlées, capitales, pays voisins, etc. Dans le cadre de ce TP, on se limite aux informations suivantes :

- les pays avec leur type de gouvernement ;
- les villes, les continents, les mers, les langues ;
- les relations entre ces entités : voisinage et dépendance entre pays, continent et langues parlées par un pays, pays d'appartenance et proximité d'une mer pour les villes.

Toutes les données sont disponibles sur le site `https://www.dbis.informatik.uni-goettingen.de/Mondial/`, et ce dans divers formats (SQL, XML, RDF). Pour vous faciliter les choses, les données sont disponibles sur Teams au format CSV. Les entêtes de ces fichiers sont affichés ci-dessous.

Tâche 1 *Extraire les informations utiles depuis les fichiers disponibles sur le site et les représenter sous forme de faits Prolog. Cela suppose de modéliser les données sous forme de types et de prédicats. La production des faits doit bien sûr être automatisée.*

Plusieurs techniques sont possibles pour transformer les données CSV en faits Prolog :

- construire un workflow Knime ;
- programmer un script (ex., en Python) ;
- importer les données CSV dans une base de données (ex., phpMyAdmin) puis générer les faits Prolog par des requêtes SQL avec interpolation de chaines de caractères.

2 Classes à apprendre

Les classes à apprendre – qui consituent autant de challenges à résoudre – sont fournies sous la forme de fichiers CSV à deux colonnes `class_country_X.csv` sur Teams. La première colonne contient les instances à classer (des pays) et la deuxième colonne contient des booléens indiquant l'appartenance ou non de chaque instance à la classe X. Dans ces fichiers, les pays sont représentées par leur nom.

Tâche 2 *Pour autant de classes que possible, tenter de produire avec cplint/Aleph un ensemble d'hypothèses "expliquant" ce que les positifs ont en commun. Cela suppose de définir le langage d'hypothèses (prédicats, types et modes des arguments) et de générer des exemples Aleph à partir des fichiers de classes. Voici la liste des classes X par ordre croissant de difficulté : 1, 7, 3, 4, 8, 2, 6, 9, 10, 5.* Vous pouvez aussi fabriquer de nouveaux challenges à échanger entre vous !

Voici les entêtes des différents fichiers disponibles.

```
-- 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"
...
```