Bases de Données II, Partie I, Charleroi, 11 janvier 2016

NOM + PRÉNOM :	
Orientation + Année :	

Cette partie de l'examen contient 4 questions.

Les prix et volumes de fruits disponibles auprès de différents magasins sont stockés dans un document XML. Les prix sont exprimés en euro/kg ou eurocent/kg, et les volumes (attribut `stock`) en tonne. Par exemple, vers la fin du fichier, on observe que `Colruyt` vend la Jonagold au prix de 145 eurocent/kg et dispose d'un stocke de 2 tonne de cette variété.

La première partie du fichier liste des variétés de fruits, avec leur mois de cueillette. La Jonagold est une variété de pomme cueillie en octobre.

La DTD est incluse au début du document XML de la figure 1.

Pour les questions 1 à 3, évitez, si possible, l'usage des axes suivants : parent, ancestor, following-sibling, preceding-sibling, following et preceding.

Pour les questions 1 à 3, il n'est pas permis d'utiliser des fonctions d'agrégation, telles que count, max, min...

Question 1 Écrivez une expression XPath (aussi simple que possible) qui rend les variétés de fruits qui sont cueillies en août et que l'on ne trouve dans aucun magasin. Pour le document de la figure 1, la réponse est comme suit :

```
<vnom sorte="prune">Mirabelle de Nancy</vnom>
<vnom sorte="pomme">James Grieve</vnom>
```

.../5

```
<?xml version="1.0" encoding="utf-8"?>
<!DOCTYPE VenteDeFruits [
<!ELEMENT VenteDeFruits (fruits, ventes)>
<!ELEMENT fruits (variete)*>
<!ELEMENT variete (vnom, cueillette)>
<!ELEMENT ventes (magasin)*>
<!ELEMENT magasin (mnom, gamme)>
<!ELEMENT gamme (fruit)*>
<!ELEMENT vnom (#PCDATA)>
<!ELEMENT mnom (#PCDATA)>
<!ELEMENT cueillette (#PCDATA)>
<!ELEMENT fruit (#PCDATA)>
<!ATTLIST vnom sorte CDATA #REQUIRED>
<!ATTLIST fruit unit (euro|eurocent) #REQUIRED>
<!ATTLIST fruit prix CDATA #REQUIRED>
<!ATTLIST fruit stock CDATA #REQUIRED>
]>
<VenteDeFruits>
<fruits>
   <variete><vnom sorte="poire">Conference</vnom>
            <cueillette>octobre</cueillette>              </variete>
   <variete><vnom sorte="poire">Doyenne</vnom>
            <cueillette>octobre</cueillette>              </variete>
   <variete><vnom sorte="pomme">Jonathan</vnom>
            <cueillette>octobre</cueillette>              </variete>
   <variete><vnom sorte="pomme">Reinette de France</vnom>
            <cueillette>octobre</cueillette>              </variete>
   <variete><vnom sorte="prune">Mirabelle de Nancy</vnom>
            <cueillette>aout</cueillette>                 </variete>
   <variete><vnom sorte="pomme">James Grieve</vnom>
            <cueillette>aout</cueillette>                 </variete>
   <variete><vnom sorte="prune">Queen Victoria</vnom>
           <cueillette>aout</cueillette>                  </variete>
   <variete><vnom sorte="cerise">Bigarreau Sunburst</vnom>
            <cueillette>juillet</cueillette>              </variete>
</fruits>
<ventes>
   <magasin>
      <mnom>Aldi</mnom>
      <gamme><fruit unit="eurocent" prix="205" stock="11">Conference</fruit>
             <fruit unit="eurocent" prix="145" stock="2">Jonathan</fruit>
             <fruit unit="euro"     prix="2.40" stock="2">Reinette de France</fruit> </gamme>
   </magasin>
   <magasin>
      <mnom>Lidl</mnom>
      <gamme><fruit unit="euro" prix="1.90" stock="11">Conference</fruit>
             <fruit unit="euro" prix="2.45" stock="9">Queen Victoria</fruit>         </gamme>
   </magasin>
   <magasin>
      <mnom>Colruyt</mnom>
      <gamme><fruit unit="euro" prix="2.10" stock="3">Conference</fruit>
             <fruit unit="euro" prix="2.75" stock="7">Queen Victoria</fruit>
             <fruit unit="eurocent" prix="145" stock="2">Jonathan</fruit>            </gamme>
   </magasin>
</ventes>
</VenteDeFruits>
```

FIGURE 1 – Vente de fruits.

Question 2 Écrivez une expression XPath (aussi simple que possible) qui rend les variétés de fruits qui sont disponibles à un prix inférieur à 2 eur/kg (ou 200 eurocents/kg). Pour le document de la figure 1, la réponse est comme suit :

```
<vnom sorte="poire">Conference</vnom>
<vnom sorte="pomme">Jonathan</vnom>
```

.../5

Question 3 Écrivez une expression XPath (aussi simple que possible) qui rend le(s) magasin(s) avec le plus grand volume de la variété Conférence en stock. Pour rappel, l'usage de min et max n'est pas permis. Astuce : $\max A = \{x \in A \mid \forall y \in A(x \geq y)\}$. Pour le document de la figure 1, la réponse est comme suit :

```
<mnom>Aldi</mnom>
<mnom>Lidl</mnom>
```

.../5

Question 4 Écrivez un programme XSLT qui génère un document XML affichant les prix par variété de fruits, dans le format illustré par la figure 2. Seules les variétés que l'on sait acheter sont affichées. Les variétés sont regroupées par sorte de fruits (poire, pomme...). Ces sortes de fruits ne sont pas connues a priori, mais doivent être calculées à partir du fichier XML. Noter que toutes les sortes de fruits seront affichées, même les sortes pour lesquelles il n'existe pas de vente (comme cerise). Tous les prix sont affichés en eurocent.

La position des blancs et retours à la ligne n'a pas d'importance. Le programme ne peut pas contenir des `xsl:for-each` ou `xsl:if`.

```
<ventes>
   <poire>
      <variete nom="Conference">
         <vente magasin="Aldi" prix="205" />
         <vente magasin="Lidl" prix="190" />
         <vente magasin="Colruyt" prix="210" />
      </variete>
   </poire>
   <pomme>
      <variete nom="Jonathan">
         <vente magasin="Aldi" prix="145" />
         <vente magasin="Colruyt" prix="145" />
      </variete>
      <variete nom="Reinette de France">
         <vente magasin="Aldi" prix="240" />
      </variete>
    </pomme>
    <prune>
       <variete nom="Queen Victoria">
          <vente magasin="Lidl" prix="245" />
          <vente magasin="Colruyt" prix="275" />
       </variete>
   </prune>
   <cerise />
</ventes>
```

FIGURE 2 – Output du programme XSLT.

.../15

Bases de Données II, Partie II, Charleroi, 11 janvier 2016

Cahier fermé. Cette partie de l'examen contient 4 questions.

NOM + PRÉNOM :	
Orientation + Année :	

Question 1 Situez chaque terme dans le cursus et expliquez de façon succincte mais précise.

Generalization error.

. . ./4

Recall.

. . ./4

Question 2 Une faiblesse de l'algorithme ID3 est que celui-ci a tendance à favoriser des attributs avec un grand nombre de valeurs distinctes. Comparer, par exemple, l'attribut `Gendre` avec deux valeurs (Femme et Homme) et l'attribut `Revenue` avec dix-neuf valeurs $[10, 20]$, $[20, 30]$,..., $[190, 200]$. L'algorithme C4.5 utilise le *gain ratio* pour faire face à cette faiblesse.

1. Expliquez plus en détail l'origine de cette faiblesse de ID3.
2. Détaillez le *gain ratio* et expliquez son usage.

. . ./8

Question 3 Expliquez la figure suivante de façon détaillée et précise.

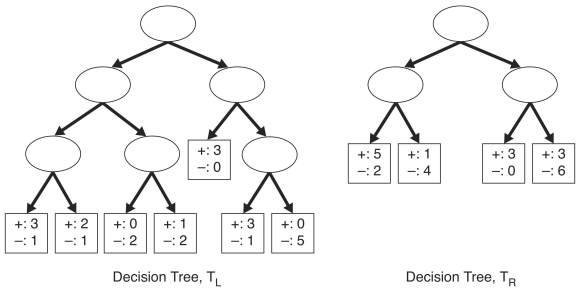

Figure 4.27. Example of two decision trees generated from the same training data.

- Que signifient les symboles $+$ et $-$?
- Comment l'arbre T_R est-il calculé à partir de T_L ?
- Soit e le *training error*. Donnez les valeurs de $e(\mathsf{T}_\mathsf{L})$ et de $e(\mathsf{T}_\mathsf{R})$.
- Comment peut-on rectifier e pour tenir compte de la complexité du modèle ?

. . ./8

Question 4 Détaillez l'exécution de l'algorithme FP-Growth sur les données suivantes, avec un seuil de support de 0.40.

TID	Produits
1	{Boursin, vin, sucre, lait}
2	{Boursin, vin, sucre}
3	{Boursin, vin, sucre}
4	{Boursin, vin, lait}
5	{vin}

. . ./8