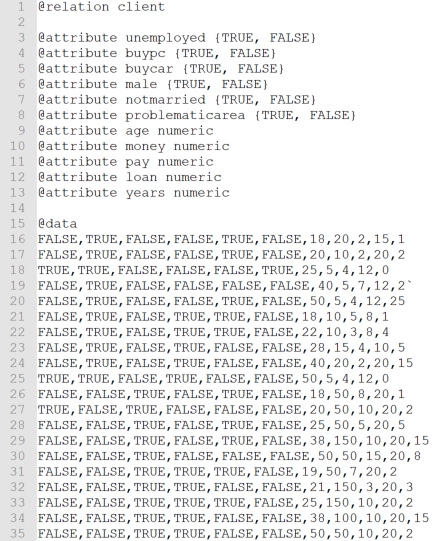
**TP 01 : Prise en main de WEKA**

**Yuteng WANG Wenxu ZHAO Liwei XU**

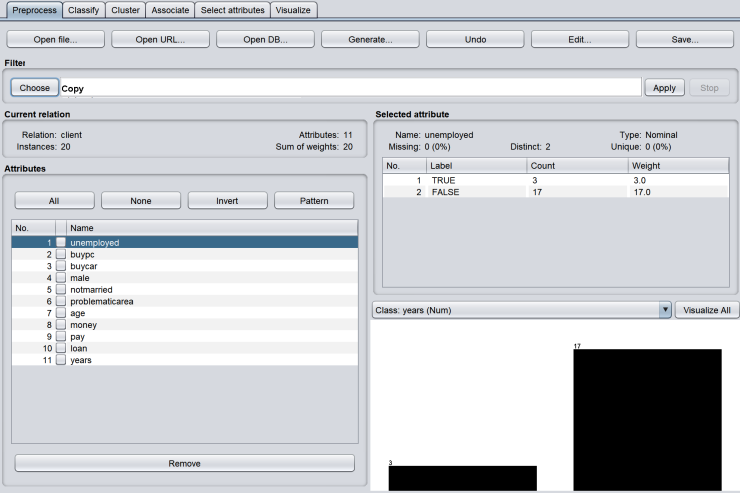
**Exercice 1** (Data Preprocessing)**.** Dans cet exercice, il s’agit de traiter des infor

mations sur les prêts japonais

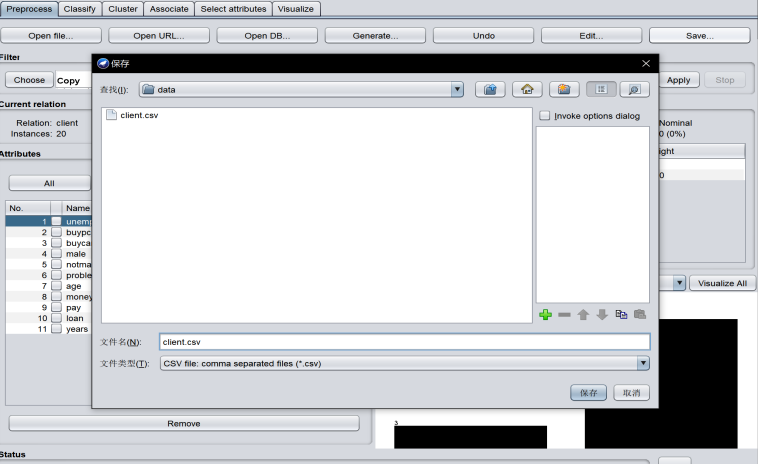
1. Construire le fichier *ARFF (Attribute-Relation File Format)*  correspondant ：



1. ouvrier le fichier avec *Weka* et le convertir en un fichier *CSV*

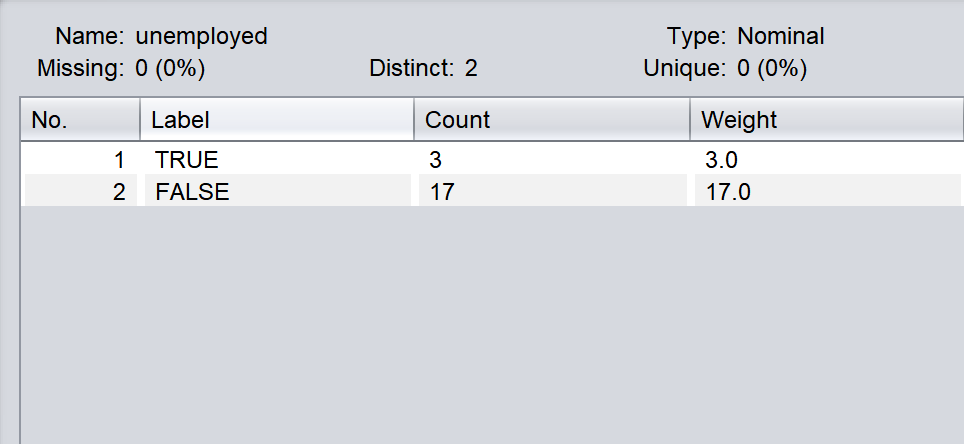


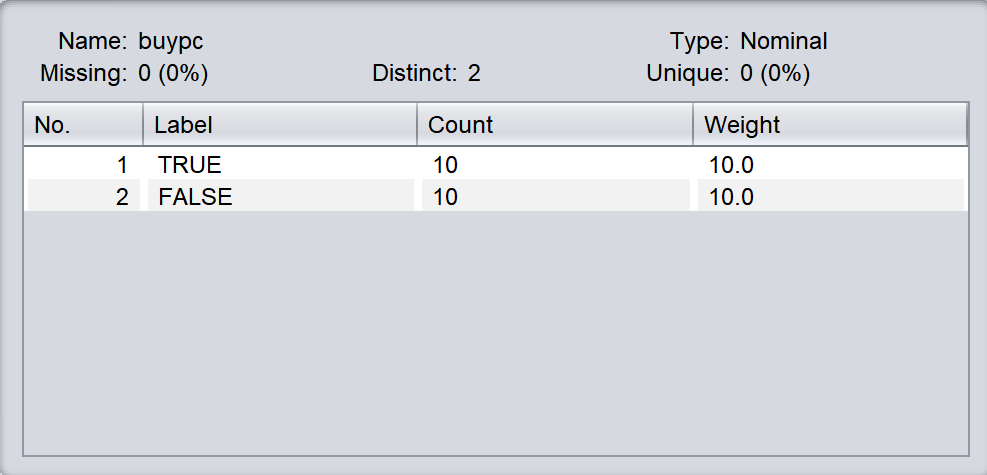
ouvrier le fichier

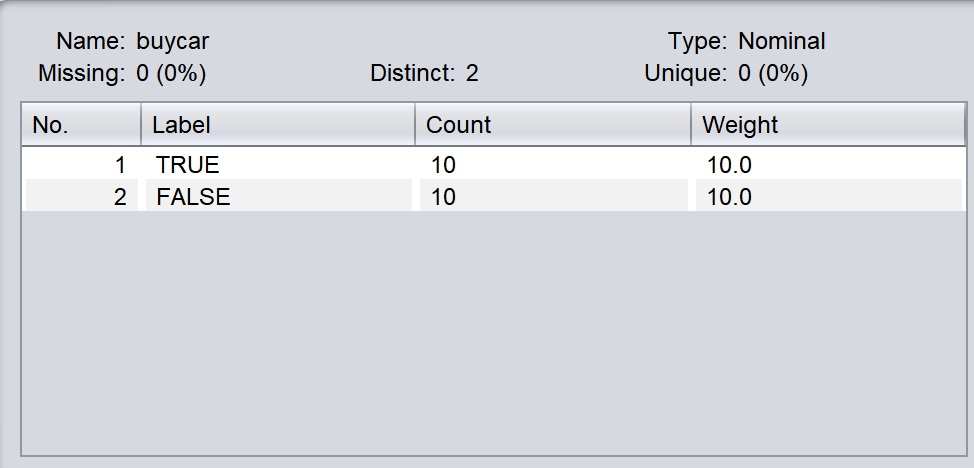


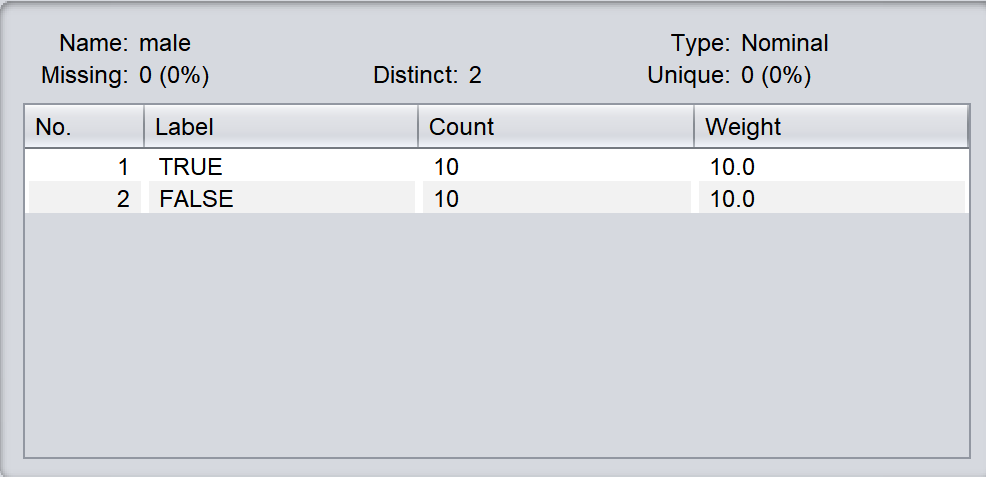
convertir en un fichier CSV

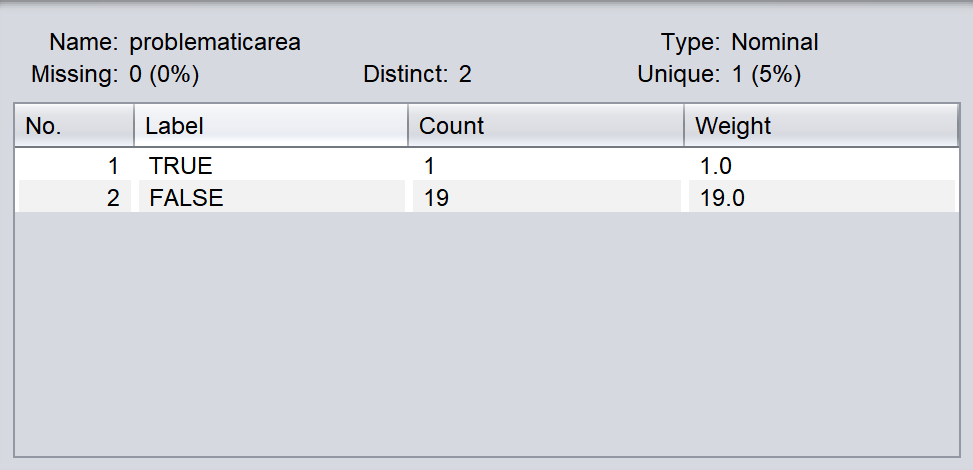
1. Examiner et visualiser les données :
   1. Type et propriétés des attributs

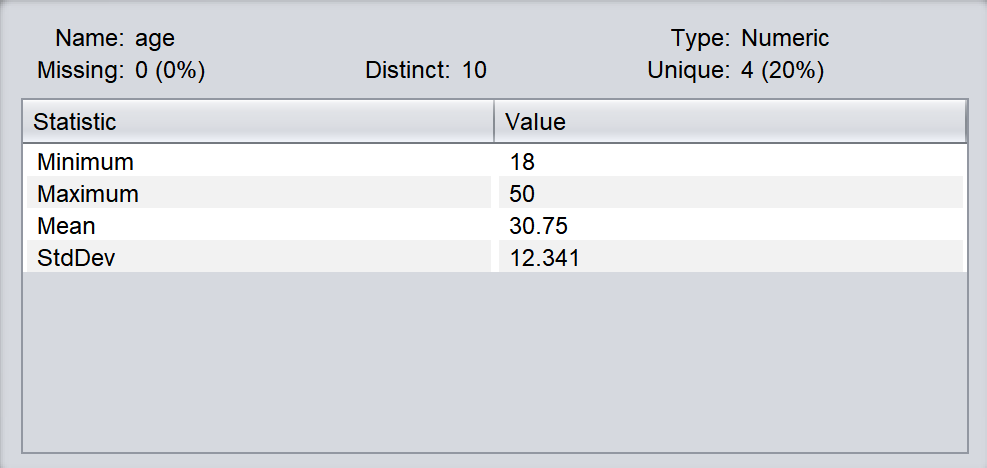


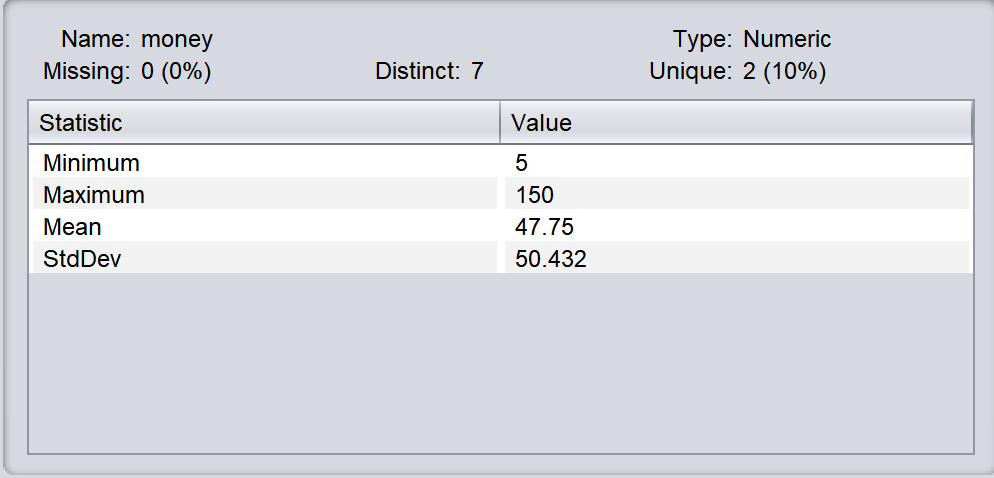


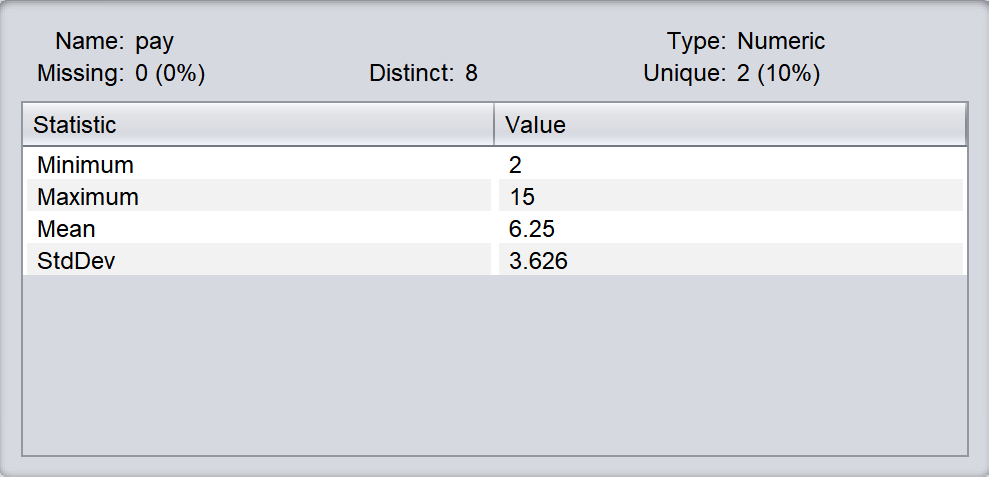


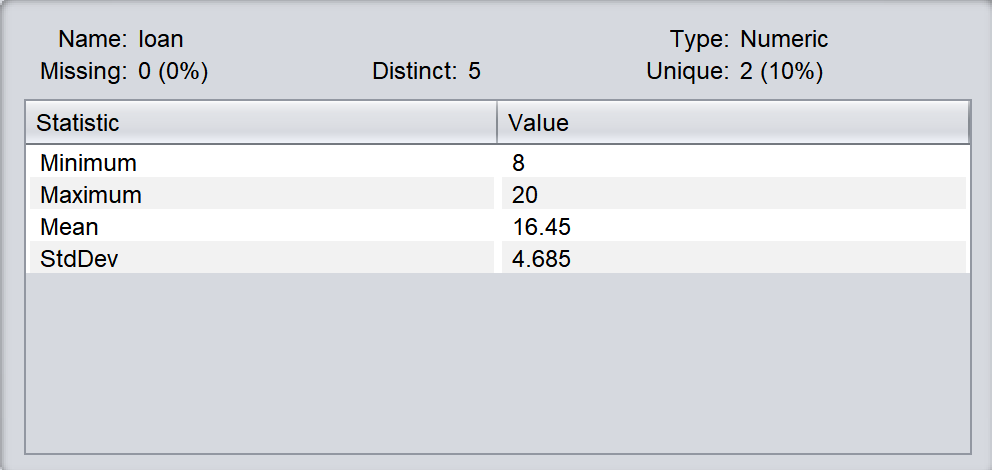


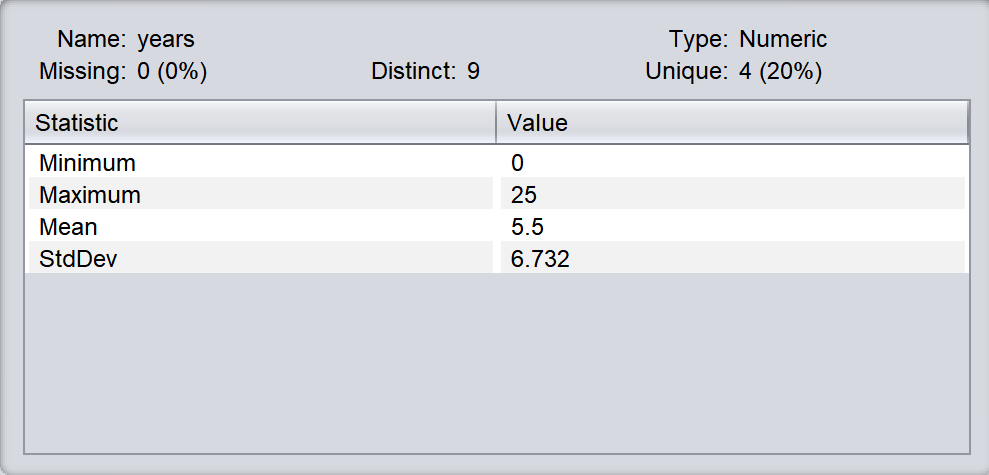


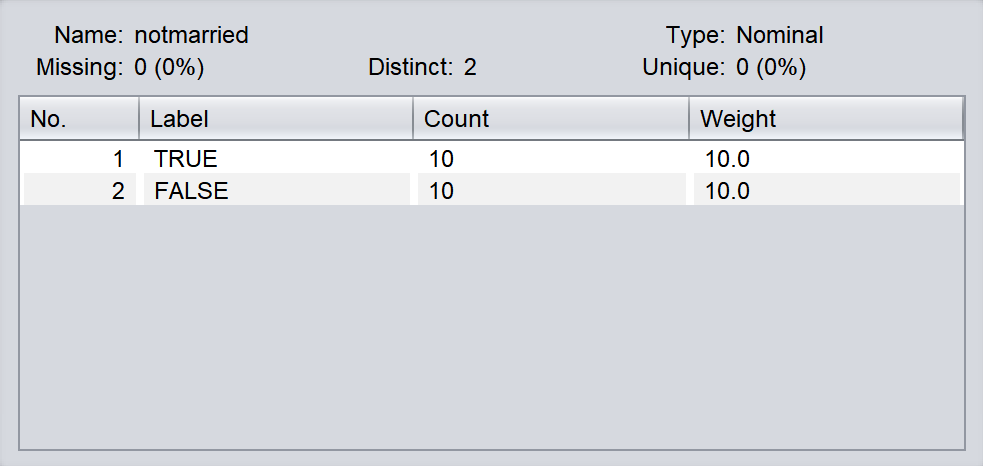




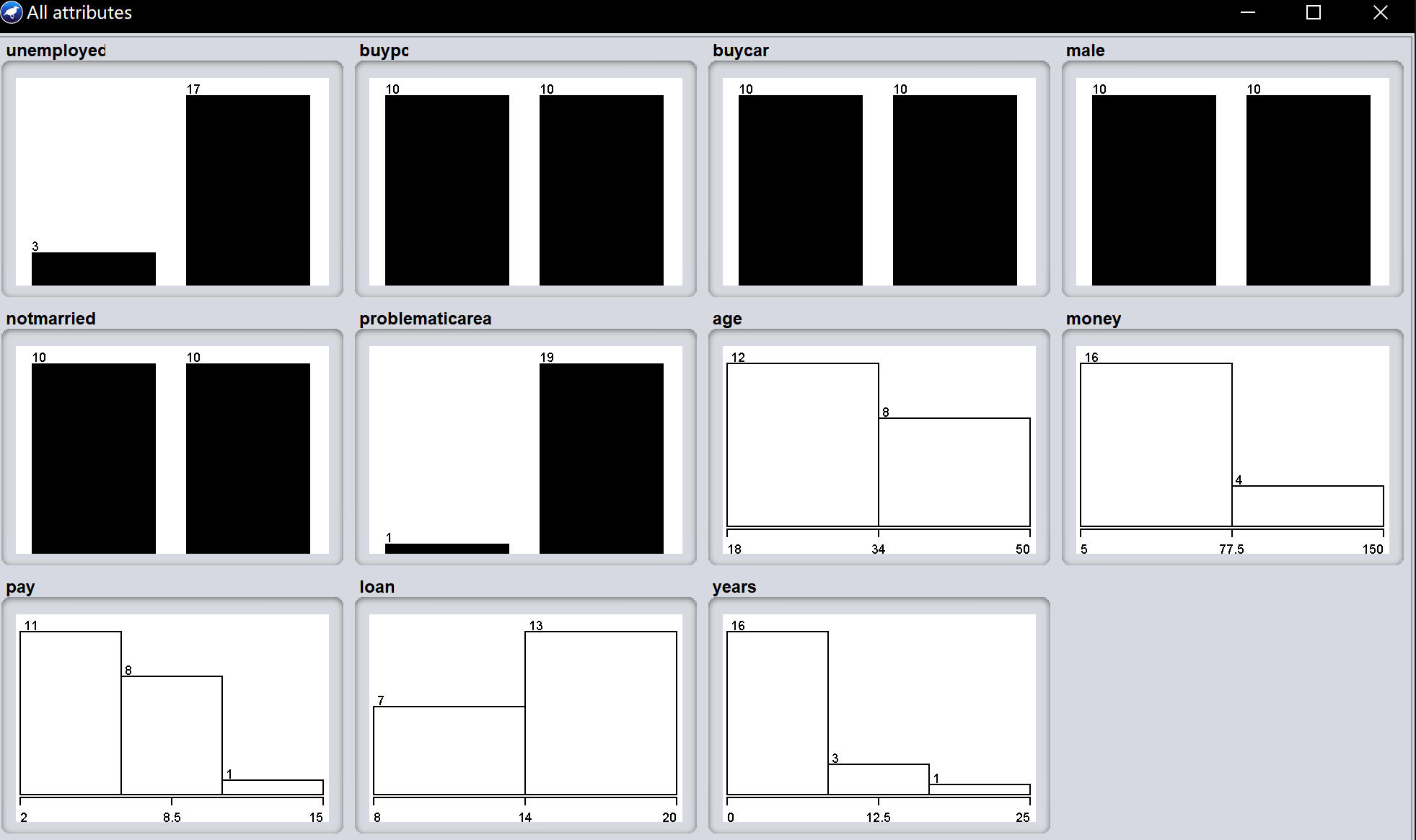






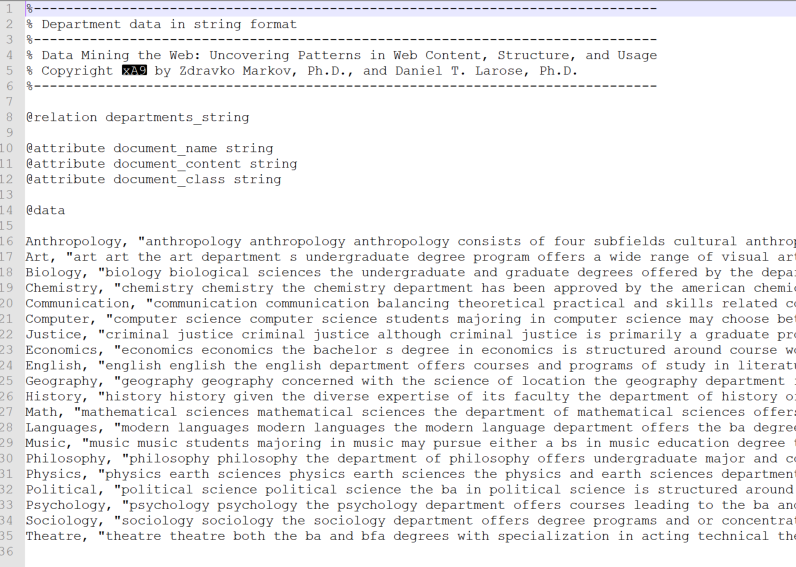


3.2 Distribution par rapport aux classes



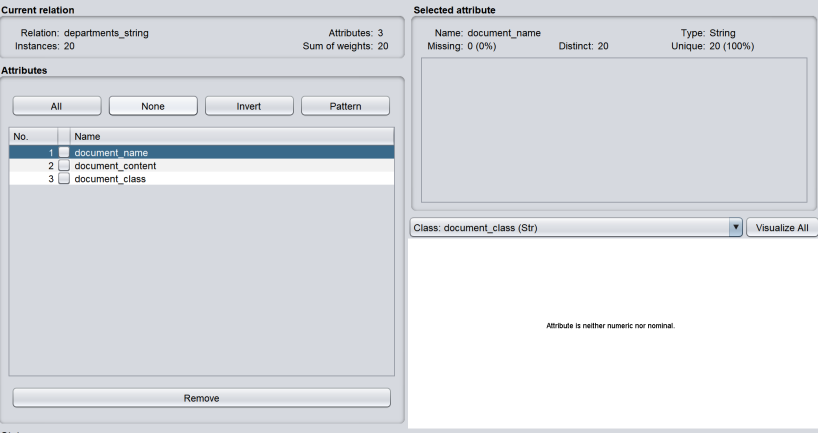
**Exercice 2** (Data Transformation)**.** Dans cet exercice, il s’agit de：

1. Visualiser *doc\_string.arff* à l’aide d’un éditeur texte et l’examiner.

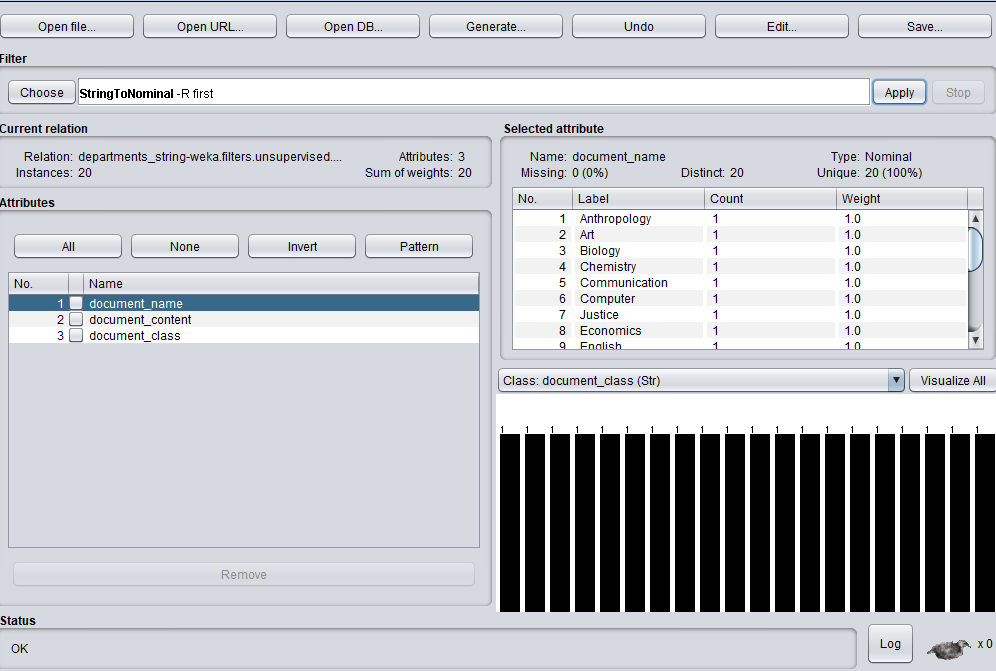


3. Ouvrir le fichier à l’aide de Weka et convertir les premier et dernier attributs

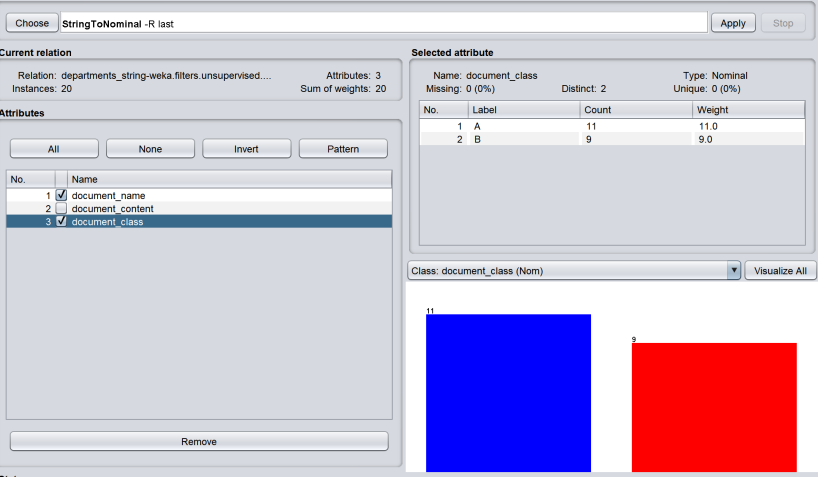
en attributs nominaux *(filters/unsupervised/attribute/StringToNominal)*.



Ouvrir le fichier



convertir le premier attribut en filtre ***StringToNominal***

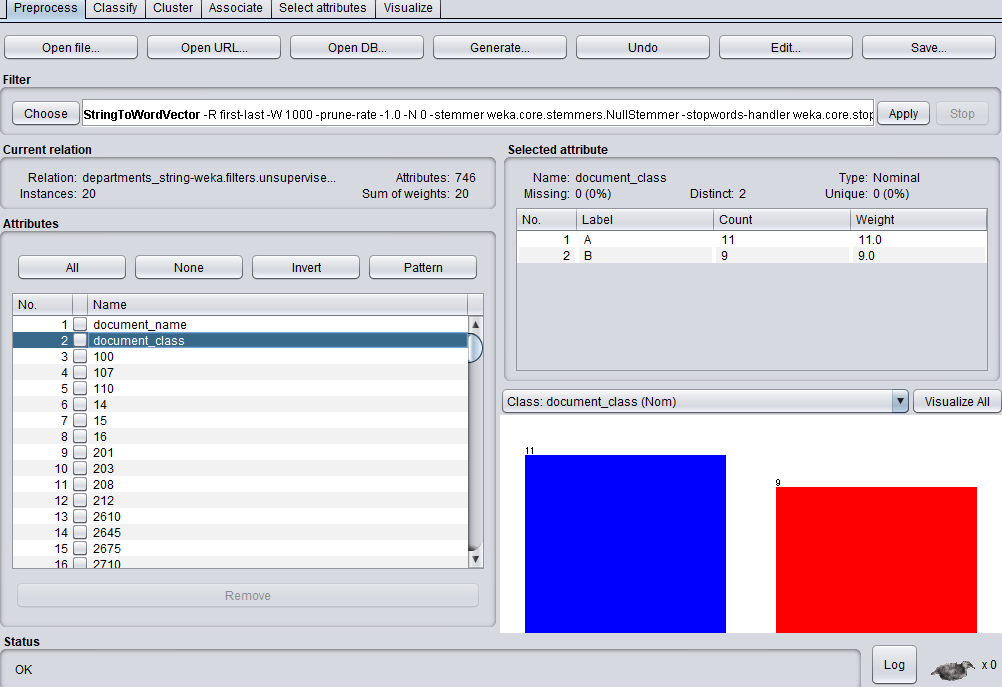


convertir le dernier attribut en filtre ***StringToNominal***

4. Convertir le texte contenu dans le deuxième attribut en un vecteur de mots

*(filters/unsupervised/attribute/StringToWordVector)*. Chaque mot devient un

attribut numérique.



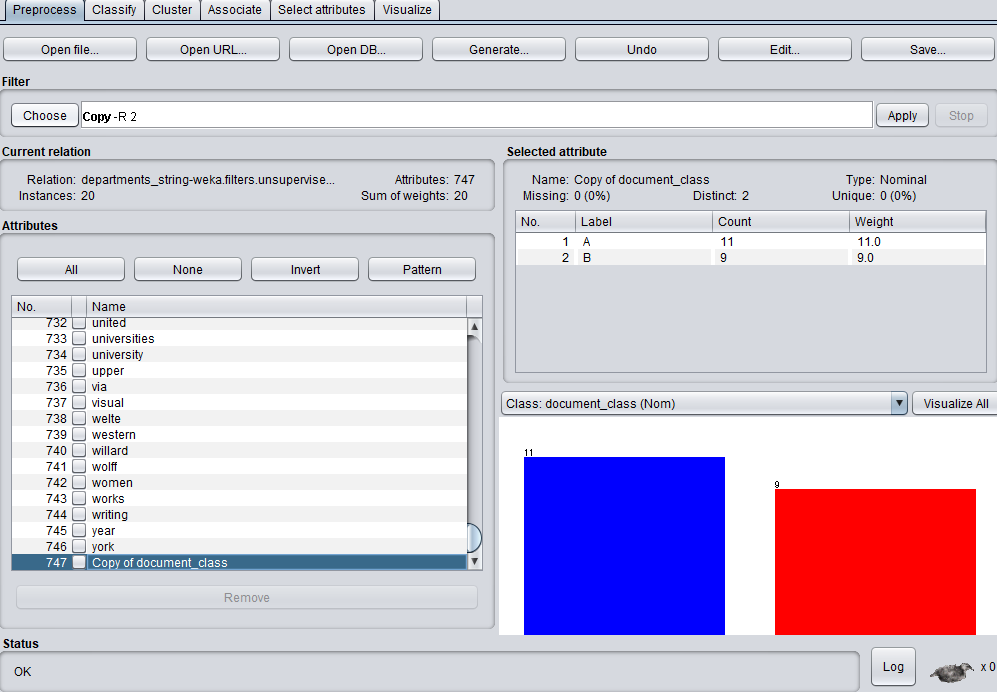
Convertir le texte contenu dans le deuxième attribut en filtre ***StringToWordVector***

5. Remettre l’attribut *Class* en fin de la liste des attributs :

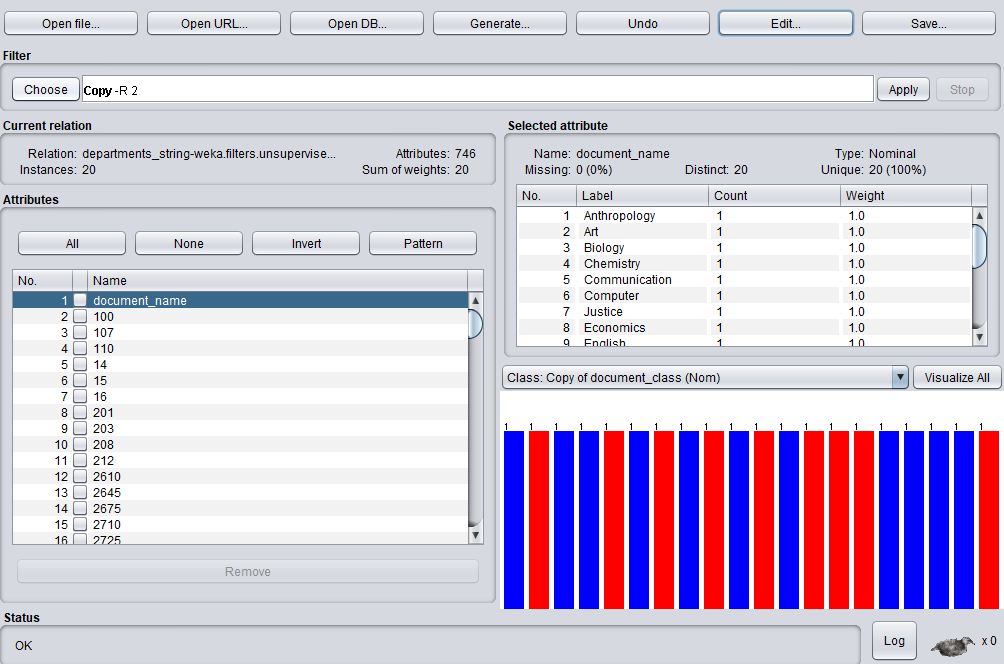
— *filters/unsupervised/attribute/Copy*.

— Mettre *index = 2*.

— Supprimer le deuxième attribut.

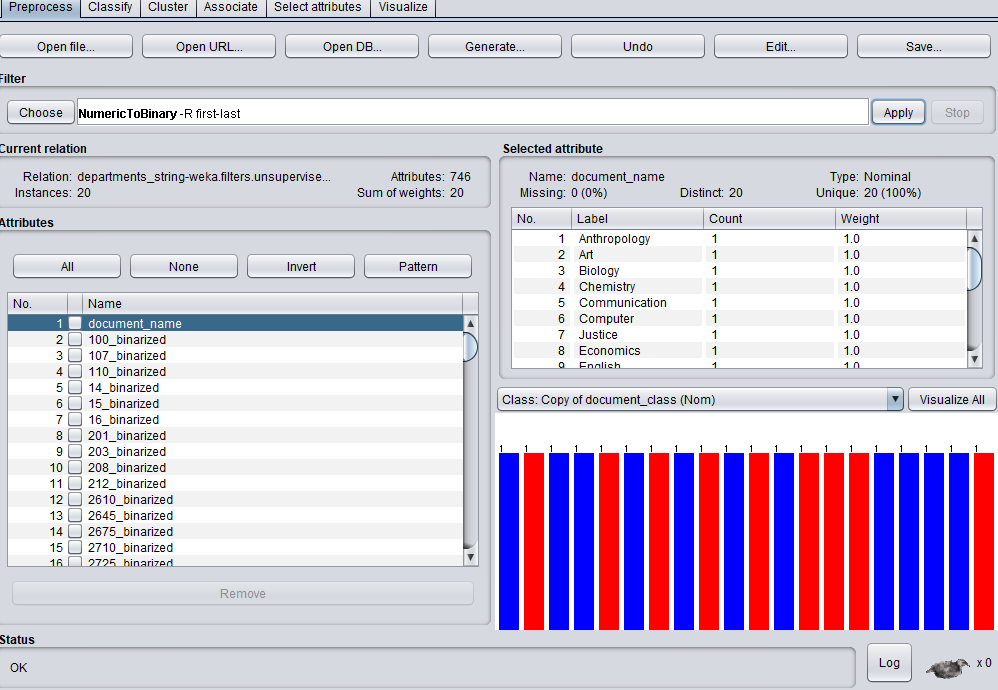


Index = 2 et attribut fin de la liste est le copy



Supprimer le deuxième attribut.*class*

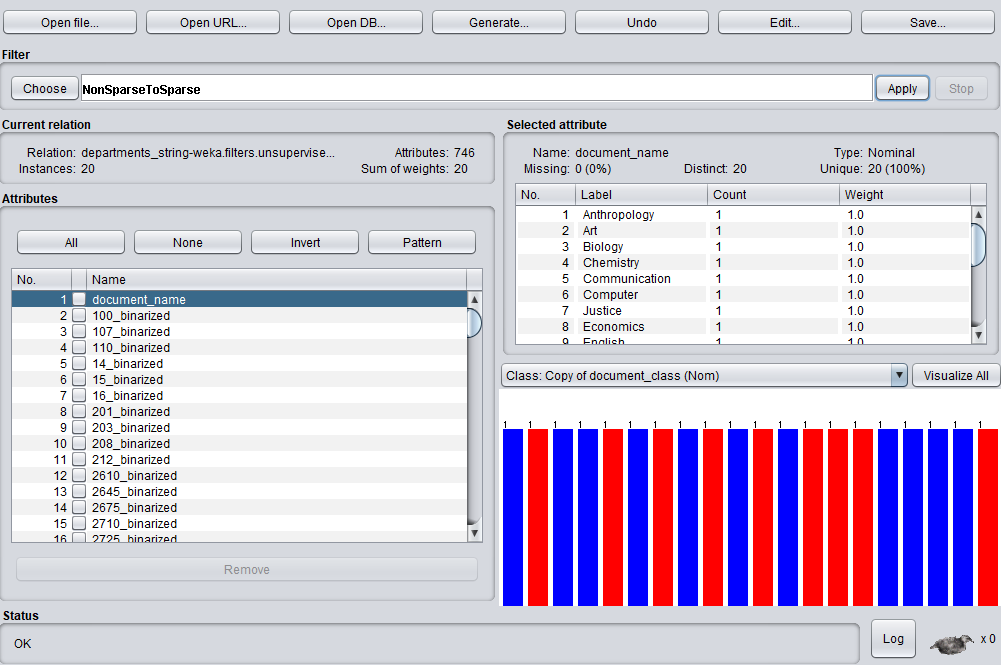
1. Changer le format numérique des attributs en format binaire et sauvegarder.



Changer le format numérique des attributs en format binaire

1. Sauvegarder une autre version *sparse* du fichier

*(filters/unsupervised/instance/NonSparseToSparse)*.



*NonSparseToSparse*

**Exercice 3** (Sélection d’attributs)**.** L’objectif de cet exercice est de trouver un sous

ensemble d’attributs permettant de mieux discriminer entre les classes : **attributs**

**pertinents**. On appelle cette opération sélection d’attributs (Future Selection).

2.Examiner les trois attributs suivants : *accounting*, *science* et *sciences*. Remplir

les tableaux de distribution suivants :

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | science | | sciences | | accounting | |
|  | 0 | 1 | 0 | 1 | 0 | 1 |
| A | 4 | 7 | 8 | 3 | 10 | 1 |
| B | 9 | 0 | 9 | 0 | 9 | 0 |

3. Extraire des tableaux précédents des règles sous la forme :

***Si*** *attribut = valeur\_attribut* ***Alors*** *classe = valeur\_classe*

***Erreur =*** *valeur\_erreur*

***Couverture =*** *valeur\_couverture*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | science | | sciences | | accounting | |
|  | 0 | 1 | 0 | 1 | 0 | 1 |
| A | 9/13 | 0 | 9/17 | 0 | 9/19 | 0 |
| B | 4/13 | 1 | 8/17 | 1 | 10/19 | 1 |

*Table erreur*

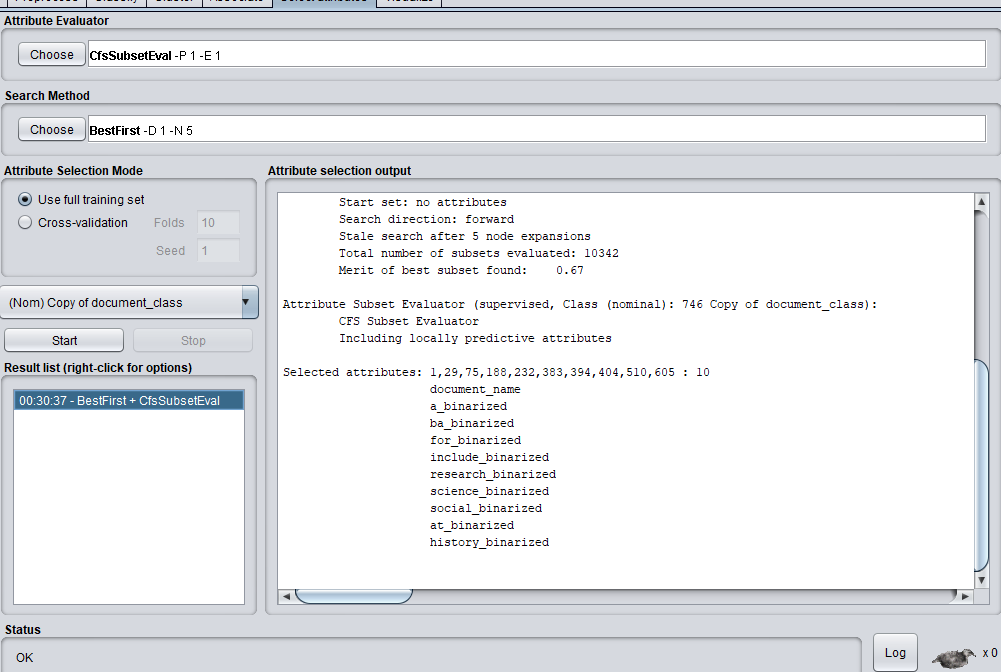
4. Parmi les attributs de la question précédente, déduire quels sont les attributs

pertinents.

Science

5. Découvrir l’onglet *Select attributes*. Explorer les paramètres par défaut (*At*

*tribute evaluator*, *Search method*).



6. En utilisant les paramètres par défaut, effectuer et comparer les résultats :

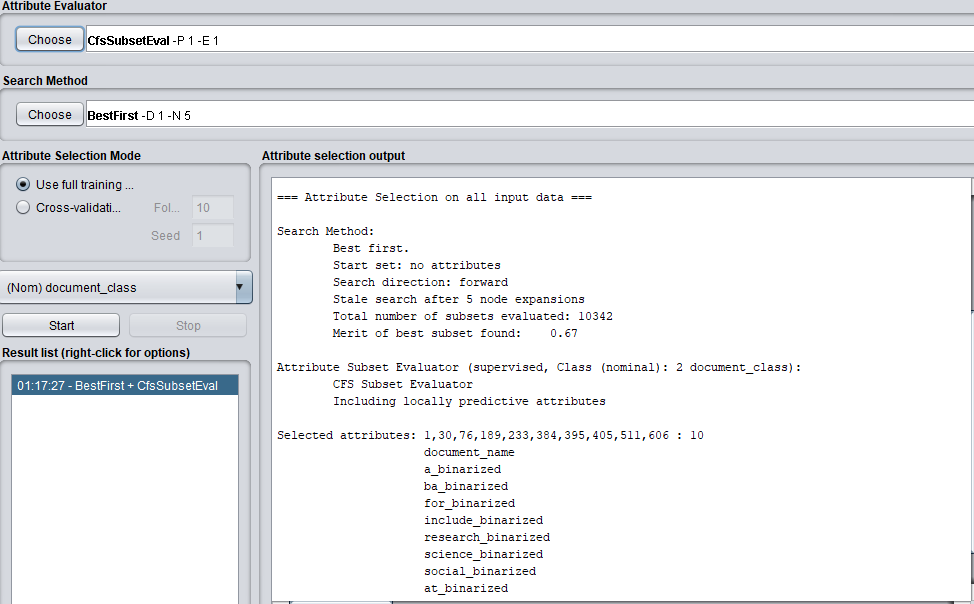
— Une sélection d’attributs en gardant l’attribut *document\_name*.

— Une sélection d’attributs en éliminant l’attribut *document\_name*.

***CfsSubsetEval*** est une méthode d'évaluation de sous-ensembles de caractéristiques (attributs) qui prédit la corrélation (ou redondance) entre chaque caractéristique et la capacité prédictive de la caractéristique pour la classification

La méthode de recherche ***BestFirst*** effectue une ascension gourmande avec retour en arrière

1.



**Une sélection d’attributs en gardant l’attribut *document\_name*.**

Selected attributes: 1,30,76,189,233,384,395,405,511,606 : 10

document\_name

a\_binarized

ba\_binarized

for\_binarized

include\_binarized

research\_binarized

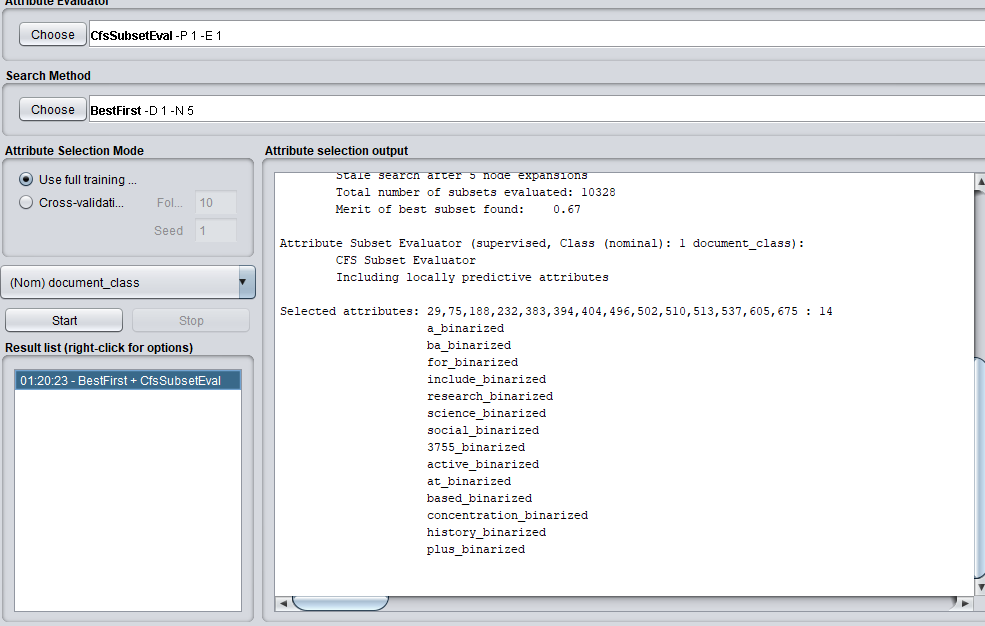
science\_binarized

social\_binarized

at\_binarized

history\_binarized

2.



**Une sélection d’attributs en éliminant l’attribut *document\_name***

*Selected attributes: 29,75,188,232,383,394,404,496,502,510,513,537,605,675 : 14*

*a\_binarized*

*ba\_binarized*

*for\_binarized*

*include\_binarized*

*research\_binarized*

*science\_binarized*

*social\_binarized*

*3755\_binarized*

*active\_binarized*

*at\_binarized*

*based\_binarized*

*concentration\_binarized*

*history\_binarized*

*plus\_binarized*

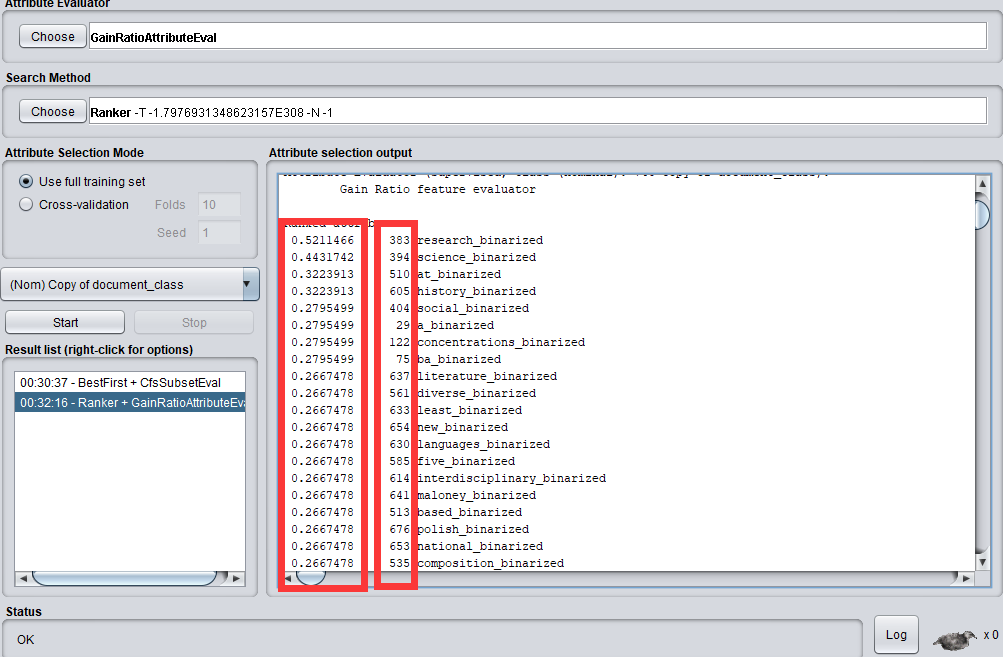
7. Classement d’attributs (*ranking*) : en utilisant *(Attribute evaluator = Gain*

*RatioAttributeEval)* et *(Search method = Ranker)*, e\_x001D\_ectuer un classement

d’attributs. Chercher une explication dans les distributions des valeurs de

classe par rapport aux valeurs d’attributs.

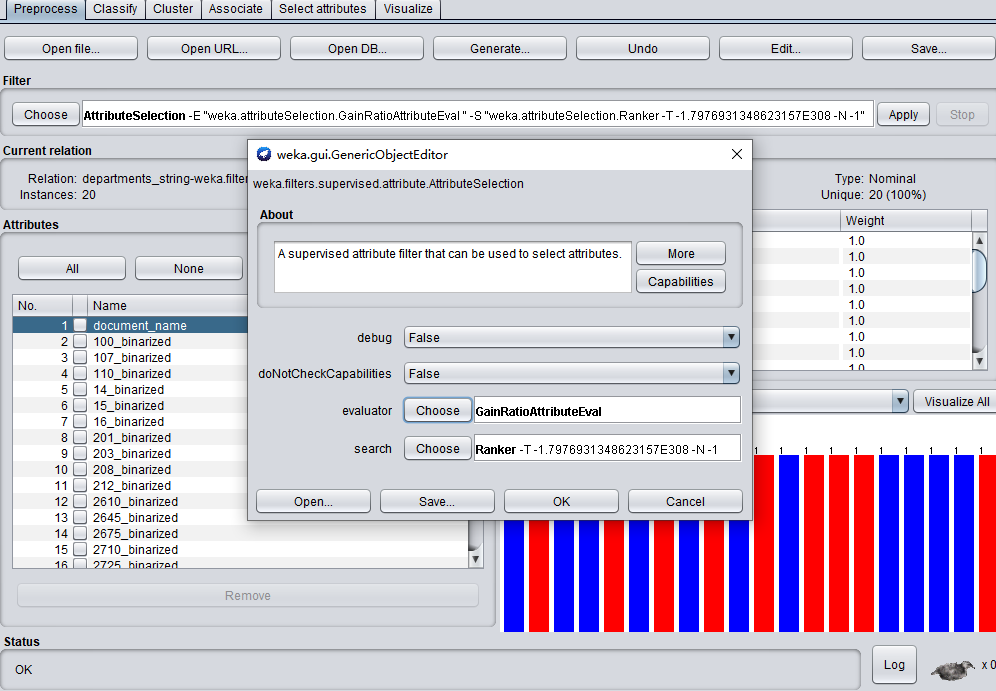
La méthode du ***Ranker*** classe une seule propriété par ordre décroissant et supprime ensuite les propriétés les moins bien classées.

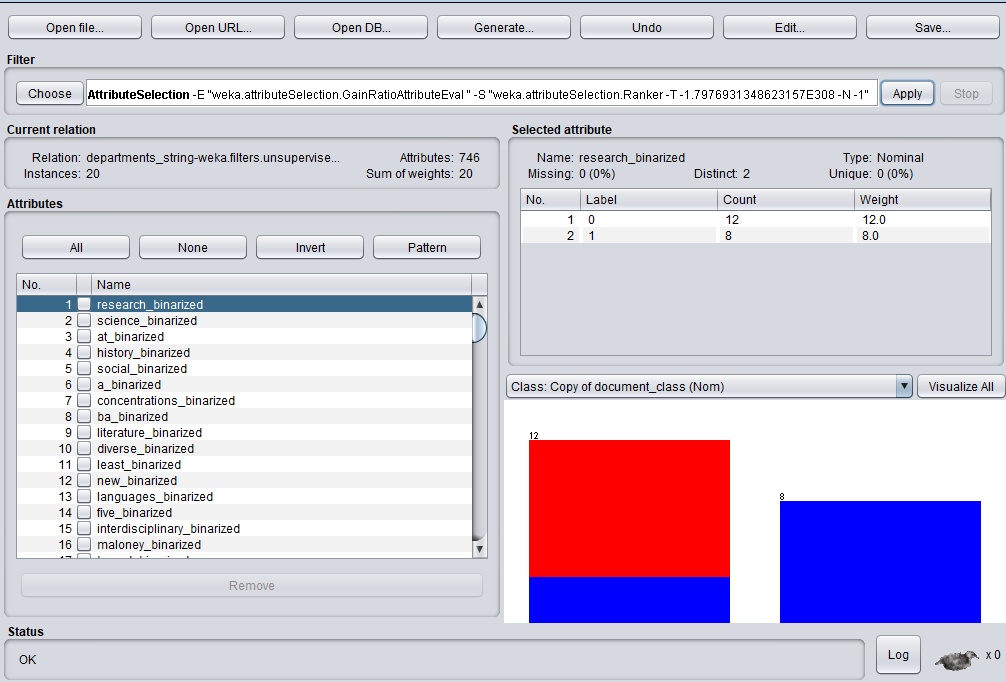


les distributions des valeurs de classe par rapport aux valeurs d’attributs

8. Effectuer la même opération en utilisant le filtre

(*filters/supervised/attribute/AttributeSelection*).





Le rank of these attributes