Glossaire

- **Apprentissage Artificiel** : un sous-domaine de l'intelligence artificielle qui s'intéresse à conférer aux machines le pouvoir de raisonner comme des êtres humains.
- **Attribut** : dans le contexte de la classification automatique ce terme indique souvent le « mot » présent dans un texte.
- **C4.5** : un algorithme de la famille des arbres de décision.
- **Exemple** : un *exemple* de l'espace d'apprentissage dans le contexte d'une classification automatique de documents est l'ensemble composé d'un document et de sa catégorie.
- **La classification/catégorisation automatique** : une approche basée sur des algorithmes d'apprentissage qui sert à classer automatiquement des documents dans des catégories en se basant sur leur contenu.
- **Classifieur** : n'importe quel algorithme d'apprentissage utilisé pour catégoriser des documents.
- **Corpus d'apprentissage** : l'ensemble des documents utilisés pour entrainer la machine.
- **DIINAR.1** : Une ressource lexicale complète pour l'arabe disposant d'un ensemble d'outils qui servent au traitement automatique de la langue arabe. Pour plus d'information voir http://silat.univ-lyon2.fr [Date de dernière visite : Mai 2010].
- **Les arbres de décision** : un algorithme d'apprentissage qui réparti l'espace d'apprentissage sous forme d'un arbre. Les attributs jugés fidèles représentent les nœuds tandis que les branches représentent leurs différentes valeurs et les feuilles représentent la décision à prendre.
- **La fidélité** : dans le contexte de la classification automatique ce terme indique la précision des décisions prises par le classifieur.
- **Le dopage ou Boosting** : dans le contexte de la classification automatique cet algorithme d'apprentissage cherche à améliorer la performance d'un autre

classifieur en l'incitant itérativement à ce concentrer dans chaque itération sur les documents mal-classés dans l'itération précédente.

- **Lemme** : une entrée du dictionnaire qui regroupe plusieurs mots ayant le même sens.
- **Les mots vides** : des mots ne portant pas de sens et récurrents souvent dans les documents. Exemple : les prépositions, les articles, etc.
- **Les machines à vecteurs de support :** dans le contexte d'une classification binaire cet algorithme d'apprentissage cherche à trouver l'hyperplan ayant la marge de séparation maximale entre les deux catégories à utiliser pour la classification.
- **N-gramme** : dans le contexte de la classification automatique des documents un n-gramme est une séquence de *n* lettres formées à partir d'une séquence de mots donnée.
- **Le principe de parcimonie** : ce principe cherche toujours à choisir l'hypothèse la plus simple qui permet d'expliquer un tel phénomène.
- **La racine** : c'est la plus petite unité lexicale qui permet de former des mots et dont les caractères existent, dans la plupart du temps, dans toutes ses différentes dérivations morphologiques.
- **La réduction de la taille du vocabulaire** : une approche qui chercher à réduire la taille d'un corpus d'apprentissage en éliminant tous les attributs qui nuisent à la décision prise par le classifieur pour améliorer sa fidélité.
- **Les réseaux bayésiens naïfs** : dans le contexte de la classification automatique cet algorithme d'apprentissage cherche à trouver la catégorie d'un document en se basant sur la probabilité que ses mots appartiennent à cette catégorie.
- **Le sur-apprentissage** : un phénomène rencontré lorsqu'un algorithme d'apprentissage est très adapté à son corpus d'apprentissage qu'il n'arrive plus à bien raisonner sur des nouvelles données.
- **Un système expert** : d'une manière générale, un système expert est un outil capable de reproduire les mécanismes cognitifs d'un expert, dans un domaine particulier. Il s'agit de l'une des voies tentant d'aboutir à l'intelligence artificielle.