

# Table de matières

<b>Introduction .....</b>	<b>15</b>
<b>Chapitre 1. L'apprentissage artificiel et la classification automatique des documents ayant un contenu <i>monolingue</i> .....</b>	<b>23</b>
1.1 Introduction.....	24
1.2 Définition de la classification automatique des documents : quoi et comment ? .....	25
1.2.1 Comment représenter un document ? .....	27
1.2.2 La réduction de la dimension de l'espace d'apprentissage.....	28
1.2.3 Le choix du classifieur.....	29
1.2.4 L'évaluation des résultats .....	30
1.3 Quelques applications de la classification automatique des documents.....	30
1.3.1 L'indexation automatique.....	31
1.3.2 L'organisation des documents .....	31
1.3.3 Le filtrage et le routage de documents.....	31
1.3.4 La désambiguïté sémantique automatique (DSA) .....	32
1.3.5 La catégorisation des pages et sites web .....	33
1.4 Les difficultés rencontrées lors de la classification automatique des documents.....	34
1.4.1 Les pièges de la langue naturel.....	34
1.4.2 La malédiction de la grande dimension de l'espace d'apprentissage .....	34
1.4.3 Le sur-apprentissage ( <i>Overfitting</i> ).....	35
1.4.4 La subjectivité de la décision.....	35
1.4.5 L'imprécision des fréquences .....	36
1.5 Conclusion .....	36
<b>Chapitre 2. La Classification de Documents : <i>Comment?</i> .....</b>	<b>38</b>
2.1 Introduction.....	39
2.2 La conception du corpus d'apprentissage .....	39
2.3 La représentation du corpus.....	40
2.4 La réduction de la taille du vocabulaire (RTV) .....	42
2.5 La sélection des attributs .....	44
2.5.1 La Fréquence de Document.....	45
2.5.2 Le Gain d'Information.....	46
2.5.3 Le Rapport de Gain .....	47

2.5.4	L'Information Mutuelle.....	47
2.5.5	La Statistique de $\chi^2$ .....	48
2.5.6	Le Rapport des chances .....	48
2.6	L'extraction des attributs .....	49
2.7	Le calcul du poids des attributs.....	50
2.8	Conclusion .....	51

**Chapitre 3. La construction inductive des classifieurs pour une classification supervisée ..... 53**

3.1	Introduction.....	54
3.2	Les approches de construction d'un modèle de prédiction.....	54
3.2.1	L'approche manuelle .....	54
3.2.2	L'approche automatique .....	56
3.3	Les caractéristiques du modèle construit.....	57
3.4	Les arbres de décisions ( Decision Trees ) .....	57
3.4.1	Le pré-élagage ( pre-pruning ).....	59
3.4.2	Le post-élagage ( post-pruning ).....	60
3.4.3	L'algorithme ID3.....	62
3.4.4	Discussion.....	63
3.5	Le classifieur bayésien naïf (BN) (Naive Bayesian Classifier).....	65
3.5.1	Discussion.....	69
3.6	Les machines à vecteurs de support (Support Vector Machines).....	70
3.6.1	Discussion.....	77
3.7	Le dopage ( « Boosting » ) .....	79
3.7.1	Discussion.....	83
3.8	Conclusion .....	86

**Chapitre 4. Le jeu de données : conception et choix de la nature des attributs ; une approche statistique basée sur l'analyseur morphologique de DIINAR.1 ..... 88**

4.1	Introduction.....	89
4.2	La description du jeu de données.....	89
4.3	La conception du jeu de données .....	89
4.4	La richesse morphologique du mot arabe .....	92
4.5	Une description générale de DIINAR.1.....	95
4.6	L'analyseur morphologique automatique de l'arabe : AraMorph .....	98
4.7	Le choix de la nature de l'attribut.....	100

4.8	Le mot, le lemme, la racine, ou le n-gramme ? Expérimentations et résultats .....	105
4.9	Conclusion .....	109
<b>Chapitre 5. Evaluation d'un classifieur, expérimentations, et résultats .....</b>		<b>110</b>
5.1	Introduction.....	111
5.2	Évaluation des résultats du classifieur .....	111
5.3	Expérimentations et résultats .....	115
5.4	Conclusion .....	119
<b>Chapitre 6. La classification automatique des documents arabes ayant un contenu multilingue : motivation, problématique et solutions .....</b>		<b>121</b>
6.1	Introduction.....	122
6.2	La classification automatique des documents arabes ayant un contenu multilingue : motivations. ....	122
6.3	Les mots écrits en caractères latins : problématique d'intégration et de fidélité .....	125
6.5	Solutions proposées .....	129
6.5.1	Coefficient Catégorique Cumulatif (« 3C ») .....	129
6.5.2	Coefficient Catégorique Cumulatif avec Sélection Composée (« 3C-SC ») .....	130
6.5.3	Le Rapport de la Fréquence de Documents avec Conformité (« RFDC ») .....	131
6.5.4	Modification de la stratégie de sélection d'attributs des méthodes connues .....	134
6.6	Expérimentations et résultats .....	135
6.7	Discussion des résultats .....	136
6.7.1	Le nombre de mots écrits en caractères latins conservés.....	137
6.7.2	L'évaluation de la performance.....	139
6.8	Conclusion .....	140
<b>Conclusion .....</b>		<b>143</b>
<b>Perspectives .....</b>		<b>147</b>
<b>Bibliographie .....</b>		<b>149</b>
<b>Table des matières des annexes .....</b>		<b>166</b>

<b>Annexe des figures .....</b>	<b>167</b>
<b>Annexe des tableaux.....</b>	<b>180</b>
<b>Annexe des algorithmes des solutions proposées .....</b>	<b>189</b>
<b>Extrait des mots vides.....</b>	<b>192</b>
<b>Transcription des caractères arabes.....</b>	<b>193</b>
<b>Un exemple d'une analyse morphologique d'un document .....</b>	<b>194</b>