

## SOMMAIRE

---

Introduction Générale	1
-----------------------	---

---

### *Partie I: Fouille de Données Textuelles, Recherche d'Information et Corpus Prophétiques*

---

<b>Chapitre I : Corpus des Traditions Prophétiques (Hadith)</b>	<b>7</b>
1. Introduction.....	7
2. Les corpus proprement dits.....	7
2.1. Définition.....	7
2.2. Le corpus dans la science .....	7
2.3. Méthodologie d'utilisation d'un corpus .....	7
3. Les Corpus Arabes.....	8
3.1. Le Corpus LDC.....	8
3.2. Le Corpus Textuel du Journal An-Nahar.....	8
3.3. Le Corpus Al-Hayat.....	8
3.4. Le corpus «Gigaword Arabe».....	9
3.5. Le corpus des textes parallèles arabe-anglais de l'O.N.U .....	9
3.6. Autres Corpus arabes.....	9
4. Les Corpus de traditions prophétiques «Hadith».....	9
4.1. Bref historique.....	9
4.2. Problème de l'authenticité.....	11
4.3. Finalités des Corpus prophétique .....	15
5. Le Hadith et le traitement de l'information.....	15
5.1. 60 ans de travail sur le lexique des Orientalistes.....	15
5.2. Etat de l'art des recherches à caractère commercial .....	16
5.3. Etat de l'art des recherches à caractère scientifique .....	22
5.4. Vers un lexique général et paramétrable, vers une bibliothèque électronique des traditions prophétiques.....	24
6. Conclusion.....	25
 <b>Chapitre II : Fouille de données textuelles et Recherche d'information</b>	 <b>27</b>
1. Introduction.....	27
2. Fouille de données textuelles.....	27
2.1. Définition de la fouille de données textuelles.....	27
2.2. Processus de la fouille de données textuelles.....	28
2.3. Objectifs de la fouille de données textuelles.....	28
2.4. Méthodes de la fouille de données textuelles.....	29
2.5. Domaines d'applications de la fouille de données textuelles .....	29
3. Recherche d'information.....	30
3.1. Le processus de RI.....	30
3.2. Les modèles de RI.....	34

3.3.	La reformulation de requêtes .....	37
3.4.	Evaluation des systèmes de Recherche d'Information .....	38
4.	Conclusion .....	40
<b>Chapitre III : La Fouille de textes appliquée aux corpus des traditions prophétiques "Hadith"</b>		<b>42</b>
1.	Introduction.....	42
2.	Le Corpus Prophétique «Al-Jâmi'us-Sahih» de L'imam Al-Boukhârî .....	42
2.1.	Description textuelle du «Sahîh Al-Boukhârî» .....	42
3.	Extraction d'information dans le corpus «Sahîh Al-Boukhârî».....	43
3.1.	Corpus et tâche.....	44
3.2.	Modèle d'Extraction d'information.....	44
4.	La fouille de textes appliquée au corpus prophétique.....	46
4.1.	Le modèle de l'espace vectoriel .....	46
4.2.	Un processus de fouille des textes.....	47
4.3.	Prétraitement des textes prophétiques .....	47
4.4.	Indexation des textes prophétiques.....	50
4.5.	Traitement des requêtes.....	51
4.6.	Calcul des similarités .....	52
5.	Expériences et résultats.....	53
5.1.	Etude statistique du corpus «Sahîh Al-Boukhârî» .....	53
5.2.	Requêtes et corpus de test.....	56
5.3.	Calcul de poids pour les termes des requêtes .....	57
5.4.	Calcul des similarités .....	57
5.5.	Evaluation du système de recherche .....	58
6.	Conclusion .....	61

---

***Partie II: Classification Thématique des Textes  
Prophétiques***

---

<b>Chapitre IV : Classification de texte et apprentissage automatique: État de L'art</b>		<b>64</b>
1.	Introduction.....	64
2.	La catégorisation des textes.....	64
2.1.	Définition de la catégorisation des textes.....	64
2.2.	Description de la catégorisation de textes.....	64
3.	Représentation des documents.....	66
3.1.	Prétraitement des documents .....	66
3.2.	Réduction de la taille du vocabulaire.....	67
3.3.	Pondération d'attributs «calcul des poids» .....	68
4.	Algorithmes de classification automatique des textes .....	69
4.1.	Classificateur Rocchio.....	69
4.2.	Classificateur bayésien naïf.....	70
4.3.	Classificateur K plus proches voisins.....	71
4.4.	Machines à support vectoriel.....	71

5.	Evaluation des performances des classificateurs.....	73
5.1.	Critères d'évaluation des classificateurs.....	73
6.	Conclusion.....	74

## **Chapitre V : Classification des textes prophétiques basée sur l'algorithme des arbres de décisions** **76**

1.	Introduction.....	76
2.	Classification par les arbres de décisions.....	76
2.1.	Algorithme arbre de Décisions.....	76
2.2.	Calcul de L'Entropie.....	77
2.3.	Le Choix de l'Attribut Gagnant.....	77
2.4.	Un Exemple pour démonstration.....	77
3.	Préparation des corpus .....	79
3.1.	Corpus Hadith.....	79
3.2.	Corpus scientifique .....	79
4.	Méthodologie de recherche.....	80
4.1.	Expériences et résultats.....	80
4.2.	Comparaison des résultats d'évaluation des deux Corpus.....	92
4.3.	Comparaison avec d'autres systèmes de classification.....	93
5.	Conclusion.....	94

## **Chapitre VI : Classification des textes prophétiques basée sur les réseaux de neurones** **96**

1.	Introduction.....	96
2.	Classification par réseaux de neurones.....	96
2.1.	Le Perceptron Multi-couches (MLP).....	97
2.2.	Les Fonctions à bases radiales.....	98
2.3.	Comparaison des réseaux de neurones MLP / RBF.....	99
3.	Modèle de classification des textes arabes.....	100
3.1.	Remplissage de la Matrice Document – Terme.....	100
3.2.	Pré-traitement.....	100
3.3.	Apprentissage du réseau neurones .....	102
4.	Expériences.....	102
4.1.	Pré-traitement et données d'entrée.....	102
4.2.	Résultats et analyses.....	105
4.3.	Discussion .....	111
5.	Conclusion.....	112

---

### ***Partie III: Segmentation Thématique des Textes Prophétiques***

---

## **Chapitre VII : Segmentation thématique : Etat de l'art** **115**

1.	Introduction.....	115
----	-------------------	-----

2.	Définition de la segmentation thématique.....	115
3.	Objectif de la segmentation thématique .....	116
4.	Les différentes segmentations possibles .....	116
4.1.	Supervisées ou non supervisées.....	117
4.2.	Linéaire ou hiérarchique .....	117
5.	Les méthodes de segmentation thématique.....	117
5.1.	Les méthodes basées sur une mesure de similarité.....	117
5.2.	Les méthodes basées sur la répétition de termes.....	118
5.3.	Les méthodes basées sur les chaînes lexicales.....	121
5.4.	Autre méthodes de segmentation thématique.....	124
6.	Conclusion .....	128

## **Chapitre VIII : Étude Comparative des Algorithmes de Segmentation Thématique pour les textes prophétiques 130**

1.	Introduction.....	130
2.	Travaux antérieurs de segmentation thématique pour la langue arabe.....	130
3.	Approches implémentés.....	131
3.1.	Pré-traitement des textes.....	131
3.2.	L'Algorithme TextTiling.....	132
3.3.	L'algorithme C99 .....	135
4.	Méthodologie d'évaluation.....	139
4.1.	Critères d'évaluation.....	139
4.2.	Corpus d'évaluation.....	140
5.	Résultats d'évaluation.....	140
5.1.	Méthode de Jugements des Lecteurs.....	141
5.2.	Méthode de Rappel / Précision.....	142
6.	Discussion.....	144
7.	Conclusion .....	145

## **Chapitre IX : organisation et accès à l'information pertinente 147**

1.	Introduction.....	147
2.	Retour de pertinence.....	147
2.1.	Probabilités de pertinence.....	148
2.2.	Calcul des poids de pertinence.....	148
2.3.	Calcul de similarité pour la requête après expansion .....	152
3.	La segmentation thématique pour le ré-ordonnancement des résultats .....	152
3.1.	Evaluation de la segmentation thématique au sein d'un SRI.....	152
3.2.	Résultats et discussion.....	158
4.	La classification thématique pour l'organisation de l'information pertinente.....	159
4.1.	Evaluation de la classification thématique au sein d'un SRI.....	160
4.2.	Résultats et discussion.....	165
5.	Conclusion .....	165

## **Conclusion Générale 167**

<b>Annexes</b>	<b>170</b>
<b>Bibliographie Personnelle</b>	<b>187</b>
<b>Bibliographie</b>	<b>190</b>
<b>Résumé</b>	<b>202</b>