

Tables des matières

Table des matières	1
Introduction générale	4
Les comptes rendus hospitaliers	5
Chapitre I. la catégorisation automatique des textes	6
1. Introduction	7
2. Formalisation du problème	7
3. Définition	8
4. Application de la catégorisation des textes	9
4.1 Indexation automatique des textes.....	9
4.2 Organisation des documents.....	9
4.3 Filtrage des textes.....	9
4.4 Informatique linguistique.....	10
4.5 La catégorisation hiérarchique des pages web.....	10
5. Les méthodes de classification	10
5.1 Regroupement hiérarchique.....	10
5.2 Regroupement base sur une partition.....	11
5.3 Méthodes basé sur la densité ou une grille.....	11
5.4 Méthode basé sur un modèle.....	11
6. Les types de classificaion	11
6.1 Classification supervisé (catégorisation).....	11
6.2 Classification non supervisé (clustering).....	11
7. Les différents contextes de classificaion	12
7.1 La classification bi-classe (Filtrage).....	12
7.2 La classification multi-classe (Routage).....	13
7.1 La classification multi-classe disjointe.....	13
7.1 La classification ordonnée.....	13
8. Lien avec la recherche documentaire	13
9. Comment catégoriser un texte	13
9.1 Le prétraitement	16
9.1.1 La segmentation.....	17
9.1.2 Suppression des mots fréquents et élimination des "Mots vides"	17
9.1.3 Suppression des mots rares.....	19
9.1.4 Le traitement morphologique.....	19
9.1.5 Le traitement syntaxique.....	20
9.1.6 Le traitement sémantique.....	20
9.1.7 Réduction de la dimension.....	21
9.2 Codage des termes	21
9.2.1 Codage Term frequency \times inverse document frequency (TF \times IDF).....	21
9.2.2 Le Codage LNU.....	22

9.3 Sélection des termes	23
9.3.1 L’algorithme de sélection des termes.....	23
10. Méthodes de représentation des documents	24
10.1 La représentation en sac mots	24
10.2 La représentation des textes par des phrases	25
10.3 Représentation des textes avec des racines lexicales et des lemmes.....	26
10.4 Représentation des textes avec la méthode des n-grammes.....	26
11. Difficultés particulières de la catégorisation des textes	28
11.1 Redondance(Synonymie)	28
11.2 Polysémie (Ambiguïté)	29
11.3 L’homographie	29
11.4 Déséquilibre	29
11.5 La graphie	29
11.6 Les variations morphologiques	30
11.7 Les mots composés	30
11.8 Présence-Absence de termes	30
11.9 Complexité de l’algorithme d’apprentissage.....	30
11.10 Sur-apprentissage.....	30
I.12 Les différents corpus utilisés dans la catégorisation des textes	31
Conclusion	31
Chapitre II. L’état de l’art	32
II.1 Introduction	33
II.2 L’état de l’art	33
II.2.1 Les projets effectués dans le domaine de la classification des rapports textuels	33
II.2.2 Choix du classifieurs	34
II.2.2.1 Classificateur bayésien	34
II.2.2.2 Les arbres de décision.....	37
a. Définition	37
b. Algorithme général d’apprentissage par arbres de décision.....	37
II.2.2.3 Algorithme des k-voisins les plus proches	38
a. Définition.....	38
b. Algorithme de classification par k-PPV.....	39
II.2.2.4 Machines à support vectoriel (svm)	39
a. Définition.....	40
II.2.3 Comparaison des algorithmes	40
II. 2.4 Analyse du corpus Reuters	41
II. 2.5 Quelle est la meilleure méthode ?	41
Conclusion	43
Chapitre III. Conception et implémentation	44
III.1 Introduction	45

III.2 Le processus du travail	45
III.3 Le corpus utilisé dans la phase d'apprentissage	47
III.4 La phase d'apprentissage	48
III.4.1 présentation de l'environnement de prétraitement Weka.....	48
III.4.2 Charger le corpus d'apprentissage	48
III.4.3 Eliminer les mots vides.....	49
III.4.4 techniques utilisés dans la sélection des termes.....	49
III.4.5 sauvegarder les données.....	50
III.4.6 Les résultats obtenus dans la phase d'apprentissage	51
III. 5 La phase de test	51
III.5.1 L'algorithme de la classification d'un nouveau document.....	51
III. 6 Configuration	53
Conclusion	54
Conclusion générale et perspectives	55
Bibliographie	56
List des figures	61
Glossaire	62