

# Table des matières

<b>Chapitre 1. R, généralités et installation</b> . . . . .	9
1.1. Histoire . . . . .	9
1.2. Installation de R . . . . .	10
1.3. Mise en route de R. . . . .	12
1.4. Evaluation du temps de calcul. . . . .	14
1.5. Installation d'un package. . . . .	15
<b>Chapitre 2. Charger des données : entrées/sorties sous R</b> . . . . .	19
2.1. Création d'un objet corpus élémentaire . . . . .	19
2.2. Sauvegarde d'un objet . . . . .	22
2.3. Comptage d'une chaîne. . . . .	22
<b>Chapitre 3. Ressources statiques : dictionnaires et collections</b> . . . . .	25
3.1. WordNet . . . . .	25
3.2. Corpus d'actualités . . . . .	26
3.3. Corpus technique . . . . .	28
3.4. Corpus étiqueté. . . . .	28
3.5. Corpus génériques. . . . .	29
<b>Chapitre 4. Ressources dynamiques : récupération de données</b> . . . . .	33
4.1. Données de la base Twitter. . . . .	33
4.2. Authentification et connexion . . . . .	34

---

4.3. Téléchargement d'information « réseau » . . . . .	37
4.4. Téléchargement de tweets . . . . .	38
4.5. Fabrication d'un objet corpus « tm » . . . . .	39
4.6. Données de la base Factiva . . . . .	39
<b>Chapitre 5. Analyse des distributions de mots . . . . .</b>	<b>41</b>
5.1. Quelques tests statistiques . . . . .	41
5.2. Modèles de distribution. . . . .	42
5.3. Partie théorique de ce chapitre. . . . .	48
<b>Chapitre 6. Analyse de questions ouvertes . . . . .</b>	<b>51</b>
6.1. Acquisition de données . . . . .	51
6.2. Normalisation de données . . . . .	53
6.3. Exploration lexicale . . . . .	54
6.4. Notion de polarité . . . . .	58
6.5. Système de codage qualitatif. . . . .	58
6.6. Visualisation . . . . .	62
6.7. Interface pour des extensions d'exploration. . . . .	66
6.8. Partie théorique de ce chapitre. . . . .	67
<b>Chapitre 7. Analyse généraliste : matrice et associations . . . . .</b>	<b>69</b>
7.1. Acquisition du corpus. . . . .	69
7.2. Stockage du corpus . . . . .	70
7.3. Prétraitements . . . . .	71
7.4. Recherche d'expressions . . . . .	74
7.5. Calcul de la matrice de contingence « termes-documents » . . . . .	75
7.6. Extraction de listes de termes . . . . .	77
7.7. Extraction de groupes de termes . . . . .	79
7.8. Partie théorique de ce chapitre. . . . .	79
<b>Chapitre 8. Analyse généraliste : termes et analyse des correspondances . . . . .</b>	<b>83</b>
8.1. Exécution de l'interface graphique . . . . .	83
8.2. Import du corpus. . . . .	84
8.3. Synopsis du corpus . . . . .	86
8.4. Analyse descriptive du lexique . . . . .	88

---

8.5. Analyse des correspondances . . . . .	91
8.6. Classification hiérarchique ascendante . . . . .	92
8.7. Partie théorique de ce chapitre . . . . .	95
<b>Chapitre 9. Analyse de la lisibilité . . . . .</b>	<b>103</b>
9.1. Acquisition du corpus et segmentation lexicale . . . . .	103
9.2. Utilisation d'un étiqueteur syntaxique (POS Tagger) . . . . .	106
9.3. Notion d'indice de lisibilité . . . . .	107
9.4. Formalité . . . . .	113
9.5. Mesure de diversité lexicale . . . . .	114
9.6. Détection d'auteur . . . . .	117
9.7. Partie théorique de ce chapitre . . . . .	118
<b>Chapitre 10. Détection d'entités nommées . . . . .</b>	<b>129</b>
10.1. Acquisition du corpus et paramétrage . . . . .	129
10.2. Extraction des entités nommées . . . . .	137
10.3. Evaluation . . . . .	144
10.4. Partie théorique de ce chapitre . . . . .	145
<b>Chapitre 11. Extraction de relations entre entités . . . . .</b>	<b>153</b>
11.1. Acquisition du corpus et paramétrage . . . . .	153
11.2. Extraction des entités nommées . . . . .	154
11.3. Extraction des relations entre entités nommées . . . . .	154
11.4. Evaluation . . . . .	158
11.5. Partie théorique de ce chapitre . . . . .	159
<b>Chapitre 12. Modèles de partitionnement de données . . . . .</b>	<b>167</b>
12.1. Sélection des mots . . . . .	167
12.2. Pondération d'un terme . . . . .	172
12.3. Liste des mots les plus fréquents . . . . .	173
12.4. Matrice de similarité . . . . .	174
12.5. Nuage de mots . . . . .	175
12.6. Classification hiérarchique ascendante . . . . .	176
12.7. K-moyennes . . . . .	179
12.8. K-médoïdes . . . . .	180

---

12.9. K plus proches voisins . . . . .	183
12.10. Partie théorique de ce chapitre . . . . .	183
<b>Chapitre 13. Analyse d’opinion . . . . .</b>	<b>191</b>
13.1. Estimation d’un sentiment . . . . .	191
13.2. Fonction de score. . . . .	198
13.3. Modèle de régression inverse . . . . .	199
13.4. Partie théorique de ce chapitre . . . . .	206
<b>Chapitre 14. Recherche de documents . . . . .</b>	<b>211</b>
14.1. Interroger un index existant . . . . .	211
14.2. Interroger son index plein texte . . . . .	212
14.3. Partie théorique de ce chapitre . . . . .	222
<b>Chapitre 15. Modèles pour la catégorisation . . . . .</b>	<b>227</b>
15.1. Modèle par extraction des n-grammes . . . . .	227
15.2. Comparaison de chaînes . . . . .	229
15.3. Analyse latente . . . . .	234
15.4. Analyse latente probabiliste . . . . .	235
15.5. Visualisation des thèmes . . . . .	249
15.6. Partie théorique de ce chapitre . . . . .	256
<b>Chapitre 16. Visualisation sémantique . . . . .</b>	<b>263</b>
16.1. Définitions . . . . .	264
16.2. Création d’un réseau d’association . . . . .	264
16.3. Visualisation interactive . . . . .	278
16.4. Visualisation locale des proches voisins . . . . .	282
16.5. Visualisation à hautes dimensions. . . . .	286
16.6. Partie théorique de ce chapitre . . . . .	288
<b>Bibliographie . . . . .</b>	<b>299</b>
<b>Index . . . . .</b>	<b>309</b>