Sommaire

1 Codification : du corpus aux tableaux statistiques **1**
1.1 Données textuelles et contextuelles 1
1.1.1 Données textuelles 1
1.1.2 Données contextuelles 2
1.2 Exemple et notation 2
1.3 Choix de l'unité textuelle 5
1.3.1 Forme graphique 5
1.3.2 Lemme 5
1.3.3 Stem 5
1.3.4 Segment répété 6
1.3.5 En pratique 6
1.4 Prétraitement 7
1.4.1 Unicité des graphies et des caractères 7
1.4.2 Automatisation partielle du prétraitement 8
1.4.3 Sélection des mots 8
1.5 Index des mots et segments 9
1.6 Codification sous forme de tableau lexical et de tableau contextuel 9
1.7 Corpus ***Life_Fr*** : résultats préliminaires 10
1.7.1 Thèmes mentionnés dans le corpus 10
1.7.2 Description univariée des données contextuelles 11
1.7.3 Note sur la gamme des fréquences 13
1.7.4 Analyse à partir des lemmes 15
1.8 Mise en œuvre avec **R** 16
1.9 En résumé 16

2 Analyse des correspondances de données textuelles **17**
2.1 Données et objectifs 17
2.1.1 AC, outil pour l'analyse de données linguistiques 17
2.1.2 Données : un exemple de taille réduite 17
2.1.3 Objectifs 19
2.2 Associations entre documents et mots 19
2.2.1 Profils des documents et des mots 19

Analyse textuelle avec R – Mónica Bécue-Bertaud
ISBN 978-2-7535-6533-3 — Presses universitaires de Rennes, 2018, www.pur-editions.fr

2.2.2 Modèle d'indépendance 20
2.2.3 Test du χ^2 22
2.2.4 Taux d'association entre documents et mots 23
2.3 Nuages des lignes et des colonnes 23
2.3.1 Espaces des profils-lignes et des profils-colonnes 23
2.3.2 Équivalence distributionnelle et distance du χ^2 23
2.3.3 Inertie des deux nuages 24
2.4 Ajustement des nuages 25
2.4.1 Axes factoriels 25
2.4.2 Représentation des lignes et des colonnes 27
2.4.3 Représentation des lignes et des colonnes de l'exemple . . . 29
2.4.4 Relations de transition et représentation superposée des lignes et colonnes 31
2.5 Aides à l'interprétation 33
2.5.1 Sélection des facteurs à étudier 33
2.5.2 Contribution d'un point à l'inertie d'un axe 33
2.5.3 Qualité de représentation d'un point 34
2.5.4 Aides à l'interprétation dans la pratique 34
2.6 Lignes et colonnes supplémentaires 35
2.6.1 Tableau principal et tableaux supplémentaires 35
2.6.2 Tableaux de fréquences supplémentaires 36
2.6.3 Tableaux quantitatifs ou qualitatifs supplémentaires 36
2.7 Validation des visualisations 36
2.8 Interprétation des résultats 39
2.8.1 Diagramme des valeurs propres 39
2.8.2 Interprétation d'un axe factoriel 39
2.8.3 Interprétation d'un plan factoriel 40
2.8.4 Éléments supplémentaires 40
2.8.5 Retour aux données 40
2.9 Mise en œuvre avec **R** 41
2.10 Démarche de l'analyse des correspondances 41

3 Applications de l'analyse des correspondances 43
3.1 Niveau de granularité de l'analyse 43
3.2 Analyse de réponses ouvertes agrégées 44
3.2.1 Données et objectifs 44
3.2.2 Sélection des mots 44
3.2.3 Longueurs des documents 44
3.2.4 Inertie et V de Cramer 46
3.2.5 Représentations des nuages sur le premier plan 47
3.2.6 Éléments supplémentaires 52
3.2.7 Mise en œuvre avec **R** 53
3.3 Analyse directe de réponses ouvertes 54
3.3.1 Données et objectifs 54

Analyse textuelle avec R – Mónica Bécue-Bertaud
ISBN 978-2-7535-6533-3 — Presses universitaires de Rennes, 2018, www.pur-editions.fr

3.3.2 Objectifs de l'analyse directe des réponses ouvertes 55
3.3.3 Traits principaux d'une analyse directe de textes courts . . 55
3.3.4 Analyse directe de la question sur la culture 56
3.3.5 Mise en œuvre avec **R** 61
3.4 Analyse discursive du discours de Badinter 63
3.4.1 Méthodologie 63
3.4.2 Résultats 64
3.4.3 Flux des arguments 66
3.4.4 Conclusion sur l'étude du discours de Badinter 69
3.4.5 Mise en œuvre avec **R** 69

4 Classification en analyse textuelle **71**
4.1 Classification de documents 71
4.2 Mesures de dissimilarité entre documents 72
4.3 Mesure de la qualité d'une partition 73
4.3.1 Les classes de documents dans l'espace factoriel 73
4.3.2 Qualité d'une partition 73
4.4 Mesures de dissimilarité entre classes de documents 74
4.4.1 Méthode du saut minimum 74
4.4.2 Méthode du saut maximum 74
4.4.3 Méthode liée à la variance : méthode de Ward 75
4.5 Classification ascendante hiérarchique (CAH) 75
4.5.1 Algorithme de construction d'une hiérarchie ascendante . . 75
4.5.2 Choix d'une partition 76
4.5.3 Description des classes 76
4.6 Partition directe 77
4.7 Stratégie mixte en classification 77
4.7.1 Consolidation d'une partition issue d'une CAH 78
4.7.2 Partition directe suivie de CAH 78
4.8 Démarche d'analyse combinant AC et CAH 78
4.9 Exemple d'utilisation conjointe de l'AC et de la CAH pour l'analyse de réponses agrégées 79
4.9.1 Données et objectifs 79
4.9.2 Prétraitement par une AC 81
4.9.3 Construction de l'arbre hiérarchique 81
4.9.4 Choix d'une partition 82
4.10 CAH sous contrainte de contiguïté 83
4.10.1 Principe et algorithme 83
4.10.2 Mise en œuvre avec **R** 84
4.11 Exemple : classification de réponses ouvertes 85
4.11.1 Données et objectifs 85
4.11.2 Prétraitement 86
4.11.3 Construction de la CAH et choix de la partition 90
4.12 Description des classes 93

Analyse textuelle avec R – Mónica Bécue-Bertaud
ISBN 978-2-7535-6533-3 — Presses universitaires de Rennes, 2018, www.pur-editions.fr

4.12.1 Description lexicale des classes 94
4.12.2 Description des classes par les variables contextuelles 95
4.12.3 Mise en œuvre avec **R** 99
4.13 Résumé de la démarche d'une classification sur coordonnées factorielles issues d'une analyse des correspondances 100

5 Caractérisation lexicale des parties du corpus 101
5.1 Mots caractéristiques 102
5.2 Mots caractéristiques et analyse des correspondances 103
5.3 Mots caractéristiques et classification 103
5.3.1 Classification opérée à partir du contenu verbal 104
5.3.2 Classification à partir des variables contextuelles 104
5.3.3 Mots ou spécificités hiérarchiques 104
5.4 Documents caractéristiques 105
5.5 Exemple : éléments caractéristiques et AC 105
5.5.1 Mots caractéristiques des catégories 106
5.5.2 Mots caractéristiques et premier plan factoriel 108
5.5.3 Documents caractéristiques de catégories 108
5.6 Exemple : mots caractéristiques en complément d'une classification 111
5.7 Mise en œuvre avec **R** 111

6 Analyse factorielle multiple en analyse textuelle 113
6.1 Tableaux multiples en analyse textuelle 113
6.2 Données et objectifs 114
6.2.1 Prétraitement 114
6.2.2 Problématique de la lemmatisation 114
6.2.3 Résumés numériques des corpus 115
6.2.4 Index des mots les plus fréquents 115
6.2.5 Tableau multiple analysé et notation 116
6.2.6 Objectifs 117
6.3 Présentation de l'analyse factorielle multiple pour tableaux de contingence (AFMTC) 118
6.3.1 Limites de l'AC du tableau juxtaposé 118
6.3.2 Principe de l'AFMTC 119
6.3.3 Intégration de variables contextuelles 120
6.4 Analyse des réponses ouvertes de l'enquête « Aspiration Internationale » en trois langues 120
6.4.1 AFMTC : valeurs propres de l'analyse globale 120
6.4.2 Représentation des documents et des mots 120
6.4.3 Représentation superposée des configurations globale et partielles 126
6.4.4 Relations entre les axes de l'analyse globale et ceux des analyses séparées 128
6.4.5 Représentation des groupes 128

Analyse textuelle avec R – Mónica Bécue-Bertaud
ISBN 978-2-7535-6533-3 — Presses universitaires de Rennes, 2018, www.pur-editions.fr

6.4.6 Mise en œuvre avec R 130
6.5 Analyse simultanée de deux questions ouvertes ; impact de la lemmatisation 130
6.5.1 Objectifs 131
6.5.2 Prétraitement 131
6.5.3 AFMTC sur la droite et sur la gauche, lemmatisées et non lemmatisées 132
6.5.4 Mise en œuvre avec **R** 135
6.6 Autres applications de l'AFMTC en analyse textuelle 135
6.7 Résumé de la démarche de l'AFMTC 136

7 Stratégie d'analyse à partir d'applications 137
7.1 Règles générales de l'exposé des résultats 137
7.2 Étude d'une base bibliographique par Annie Morin 139
7.2.1 Présentation des données lupus 139
7.2.2 AC du tableau documents × mots 141
7.2.3 Analyse du tableau agrégé par année 145
7.2.4 Étude chronologique sur les noms de médicaments 147
7.2.5 Mise en œuvre avec R 147
7.2.6 Conclusion de cette étude 149
7.3 Analyse du scénario d'un film par Fionn Murtagh 149
7.3.1 Données et objectifs 149
7.3.2 Quelques mots sur le script du film 150
7.3.3 Résultats préliminaires 151
7.3.4 Méthodologie 152
7.3.5 Résultats 153
7.3.6 En conclusion 155
7.3.7 Mise en œuvre avec **R** 156
7.4 Twitter par Fionn Murtagh 156
7.4.1 Introduction au traitement 157
7.4.2 Données et objectifs 158
7.4.3 L'accès aux données Twitter 159
7.4.4 Analyse de Twitter : un flux de tweets 160
7.4.5 Analyse des correspondances des tweets 160
7.4.6 Utilisation des hashtags 162
7.4.7 Considérations générales sur l'analyse du contenu des flux de Twitter 163
7.4.8 Mise en œuvre avec **R** 165
7.5 Discours politiques par Ramón Álvarez-Esteban et Mónica Bécue-Bertaut 165
7.5.1 Données et objectifs 165
7.5.2 Méthodologie 167
7.5.3 Résultats 167
7.5.4 Mise en œuvre avec R 180

Analyse textuelle avec R – Mónica Bécue-Bertaud
ISBN 978-2-7535-6533-3 — Presses universitaires de Rennes, 2018, www.pur-editions.fr

Annexe : Packages d'analyse textuelle en R par Annie Morin **181**

Bibliographie **183**

Index **187**

Analyse textuelle avec R – Mónica Bécue-Bertaud
ISBN 978-2-7535-6533-3 — Presses universitaires de Rennes, 2018, www.pur-editions.fr