

TABLE DES MATIÈRES

1. L'analyse statistique des données textuelles : champs et objets d'étude

1. Le champ de l'analyse statistique des données textuelles (ADT)
2. Les disciplines voisines
3. Les corpus et les enquêtes

2. Les unités d'analyse et les observations

1. La segmentation du texte en unités
2. L'annotation de surface automatique
3. Les unités séquentielles

3. Les unités en contexte

1. La concordance
2. Une typologie des formes de retour au texte
3. La cooccurrence, une synthèse statistique des contextes
4. Le calcul des spécificités, un outil pour la caractérisation contrastive des contextes locaux et globaux

4. Exploration, visualisation, validation et inférence : les principes de base

1. Les approches exploratoires et confirmatoires
2. Les méthodes d'analyse en axes principaux
3. Les méthodes de classification
4. La validation par rééchantillonnage

5. L'analyse en composantes principales (ACP)

1. Les interprétations géométriques
2. Le problème des échelles de mesure et la transformation des données
3. La représentation des mots et des répondants
4. L'analyse du nuage des p variables (colonnes)
5. Observations et variables supplémentaires
6. L'analyse factorielle en facteurs communs et spécifiques
7. La validation par rééchantillonnage (*bootstrap*)
8. Deux exemples d'application

6. L'analyse des correspondances (AC)

1. La démarche d'après un exemple
2. La représentation simultanée des lignes et des colonnes

3. Les éléments supplémentaires
4. Les aides à l'interprétation
5. La validation par rééchantillonnage
6. L'analyse des correspondances multiples (ACM)
7. D'autres méthodes

7. La classification des mots et des textes

1. La classification ascendante hiérarchique (CAH) d'après un exemple
2. Les méthodes de classification hiérarchique, les représentations arborées
3. Les méthodes de partitionnement
4. La classification mixte et autres modèles
5. La sériation
6. La validation des classifications

8. Les stratégies d'analyse et la complémentarité entre analyse en axes principaux et classification

1. Les forces et les faiblesses des méthodes en axes principaux
2. L'utilisation conjointe des axes principaux et de la classification
3. La description statistique des classes ou des catégories : valeurs-test et spécificités
4. Les fragments caractéristiques (ou réponses modales)
5. Les stratégies d'analyse : le cas des corpus de réponses libres (ou des textes courts, nombreux, qualifiés)
6. La fragmentation d'un corpus en unités de contexte

9. L'articulation entre les analyses exploratoires et confirmatoires

1. Explorer, valider, prévoir...
2. La stylométrie et la discrimination globale
3. Les unités statistiques de la stylométrie
4. Un exemple de modèle statistique en stylométrie
5. Les analyses discriminantes globales
6. Discrimination et validation : un exemple
7. La discrimination et les réseaux de neurones
8. Les recherches de thèmes (*Topic Modeling*) : un point de vue

AUTEURS

LUDOVIC LEBART, ex-directeur de recherche au Centre national de la recherche scientifique (CNRS), est statisticien et enseignant-chercheur à Télécom ParisTech. Ses sujets de recherche sont la statistique multidimensionnelle, la qualité des enquêtes socio-économiques, l'inférence statistique en analyse des données et les logiciels d'analyse des données qualitatives et textuelles. Il est l'auteur de nombreux livres sur ces thèmes traduits en plusieurs langues.

BÉNÉDICTE PINCEMIN est chargée de recherche en linguistique au CNRS, au sein de l'Institut d'histoire des représentations et des idées dans les modernités de l'École normale supérieure de Lyon. Elle est membre du projet Textométrie, qui développe le logiciel TXM. Ses travaux portent sur la modélisation de la textualité et de l'activité interprétative pour l'analyse sémantique de corpus.

CÉLINE POUDAT est linguiste et maître de conférences en analyse du discours à l'Université Côte d'Azur à Nice. Elle étudie les typologies textuelles et les genres de la communication médiée par les réseaux, qu'elle explore avec les méthodes de l'analyse de données textuelles et de la linguistique de corpus. Elle codirige le consortium national français Corpus, Langues et Interactions.

Financé par le
gouvernement
du Canada

Funded by the
Government
of Canada

Canada



Conseil des arts
du Canada

Canada Council
for the Arts

SODEC

Québec



Distribution

Canada : Prologue inc.
Belgique : SOFEDIS / SODIS

France : SOFEDIS / SODIS
Suisse : Servidis SA



418 657-4399
puq@puq.ca



Presses
de l'Université
du Québec

50 ans
de savoir