Projet 2 : Analyse factorielle discriminante

Résumé : *Pour le projet 2, axé sur un niveau intermédiaire d'analyse factorielle discriminante (AFD), vous travaillerez avec un ensemble de données Twitter Entity Sentiment Analysis. Cet ensemble de données se compose de tweets associés à diverses entités et du sentiment exprimé à l'égard de ces entités. L'AFD peut être appliquée pour réduire la dimensionnalité des données et pour visualiser la manière dont les tweets se regroupent autour des sentiments.*

Objectif principal : Utiliser l'analyse factorielle discriminante (AFD) pour réduire les dimensions des données des tweets et visualiser le regroupement des sentiments.

Source des données : jeu de données Twitter Entity Sentiment Analysis (Kaggle).

- **Lien** : https://www.kaggle.com/datasets/jp797498e/twitter-entity-sentiment-analysis

Aperçu des tâches :

1. **Chargement et pré-traitement des données :**
   - Chargez le jeu de données dans R ou R Studio ou VS Code.
   - Nettoyez les données textuelles du tweet en supprimant les caractères spéciaux, les URL et les mots vides.
   - Convertissez les étiquettes de sentiment dans un format numérique adapté à DFA.
2. **Ingénierie des caractéristiques (feature engineering):**
   - Extraire les caractéristiques à partir des données textuelles, éventuellement à l'aide de TF-IDF, d'intégrations de mots ou de scores de sentiment.
3. **Analyse factorielle discriminante :**
   - Utilisez les fonctions du package MASS dans R pour effectuer une analyse discriminante linéaire (LDA) afin de projeter les données dans un espace de dimension inférieure qui maximise la séparation des classes de sentiment.
4. **Visualisation:**
   - Visualiser les résultats de l'AFD pour déterminer comment les tweets sont regroupés par sentiment dans l'espace réduit.
5. **Évaluation du modèle :**
   - Si DFA est utilisé pour la classification, évaluez la précision (accuracy) du modèle.
   - S'il est utilisé uniquement à des fins de visualisation, évaluez la qualité de l'agrégation à l'aide de scores de silhouette (silhouette scores) ou d'une mesure similaire.
6. **Rapport et présentation :**
   - Documentez la méthodologie et les résultats dans un rapport détaillé.
   - Préparez une présentation résumant l'approche et les résultats de la visualisation.

Livrables:

- **Code :** script R ou document RMarkdown contenant tout le code pour le pré-traitement, l'analyse et la classification.
- **Rapport :** rapport détaillé (HTML) expliquant la méthodologie, les résultats et les performances du modèle de classification. Incluez des visualisations de sujets et de facteurs discriminants.
- **Attribution :** 2 étudiants (max.).

Annexes:

Schéma du contenu du rapport (RMarkdown ou HTML)

1. **Introduction :**

- Présentation de l'ensemble des données et de l'objectif du projet.

1. **Méthodologie :**

- Les étapes de pré-traitement détaillées, les méthodes d'extraction des caractéristiques et la justification de l'utilisation de l'AFD en général.

2. **Implémentation de l'AFD :**

- Explication de l'AFD, y compris les fondements mathématiques et son application à l'ensemble de données de tweets.

3. **Résultats :**

- Résultats de la visualisation de l'AFD et toute métrique de classification, le cas échéant.

4. **Discussion & Conclusion :**

- Analyse de la répartition des sentiments des tweets dans l'espace AFD et réflexions potentielles.

- Résumé des résultats et de l'utilité potentielle de la AFD dans l'analyse des sentiments.

5. **Travaux futurs :**

- Suggestions pour d'autres analyses ou d'autres modèles d'apprentissage automatique qui pourraient être appliqués.