Projet 3 : Classification bayésienne et analyse factorielle discriminante

***Résumé** : Le jeu de données contient des informations sur les thèses de doctorat françaises, en mettant l'accent sur la similarité sémantique. Cela représente un défi unique pour la classification en raison de la nature textuelle et sémantique des données.*

Objectif principal : Mettre en place une classification bayésienne avancée avec analyse discriminante sur un jeu de données de résumés de thèses de doctorat françaises afin de les catégoriser en domaines d'étude.

Source des données : Recherche de similarité sémantique de thèse de doctorat française à partir de Kaggle.

- **Lien** : https://www.kaggle.com/code/antoinebourgois2/french-doctoral-thesis-semantic-similarity-search

Aperçu des tâches :

1. **Prétraitement des données :**
   - Nettoyage du texte : Supprimez les mots d'arrêt (stopwords), la ponctuation et effectuez un stemming ou une lemmatisation.
   - Vectorisation : Convertir les données textuelles sous forme numérique à l'aide de TF-IDF ou le plongement lexical « plongement de mots ou word embeddings » tels que Word2Vec.
2. **Extraction de caractéristiques (Feature Extraction):**
   - Utilisez des techniques de traitement du langage naturel (NLP) pour extraire des caractéristiques significatives des résumés de thèse.
   - Explorez la modélisation des rubriques (par exemple, LDA pour l'extraction de rubriques) pour comprendre les rubriques principales et les utiliser comme fonctionnalités.
3. **Réduction de la dimensionnalité :**
   - Appliquez l'analyse discriminante linéaire (LDA, à ne pas confondre avec l'allocation de Dirichlet latente (Latent Dirichlet Allocation), également abrégée en LDA) pour réduire la dimensionnalité de l'espace d'entités tout en préservant la séparabilité des classes.
   - Vous pouvez également utiliser des techniques non linéaires telles que l'analyse discriminante du noyau (Kernel Discriminant Analysis) si les méthodes linéaires sont insuffisantes en raison de la complexité des données textuelles.
4. **Classification bayésienne :**
   - Construire un classificateur bayésien pour catégoriser les thèses dans différents domaines d'étude en fonction des caractéristiques extraites.
   - Utilisez des méthodes bayésiennes avancées qui peuvent traiter des données de grande dimension et qui conviennent à la classification de texte.

5. **Optimisation et validation du modèle :**

   - Optimisez les hyperparamètres du modèle à l'aide de techniques telles que la recherche de grille (Grid Search) ou l'optimisation bayésienne.

   - Validez le modèle à l'aide de stratégies de validation croisée appropriées pour les données textuelles.

6. **Performance Evaluation:**

   - Utilisez des métriques adaptées à la classification, telles que l'exactitude, la précision, le rappel, le score F1 (Accuracy, Precision, Recall, and F1-score), et envisagez également la méthode ROC-AUC si le problème est formulé comme une classification binaire pour chaque domaine d'étude.

   - En outre, utilisez des matrices de confusion et des rapports de classification pour évaluer les performances dans différents domaines d'étude.

7. **Interprétabilité :**

   - Analysez les facteurs discriminants pour interpréter les caractéristiques (mots, phrases, sujets) qui ont le plus d'influence sur la distinction entre les domaines d'études.

Livrables:

- Code : script R ou document RMarkdown contenant tout le code pour le prétraitement, l'analyse et la classification.

- Rapport : rapport détaillé (RMarkdown) expliquant la méthodologie, les résultats et les performances du modèle de classification. Incluez des visualisations de sujets (topics) et de facteurs discriminants.

Considérations:

- Assurez-vous de gérer la nature multi-classes du problème de classification.

- Relevez le défi du déséquilibre des classes si certains domaines d'études ont beaucoup plus de thèses que d'autres.

- Tenez compte de l'interprétabilité du modèle bayésien, en particulier de la façon dont vous communiquez le rôle des facteurs discriminants dérivés de données textuelles.

Annexes:

Code in R

Pour le code R, vous devez suivre ces étapes générales :

1. **Chargement et pré-traitement des données :**

   - Chargez les données textuelles.

   - Nettoyez les données textuelles : supprimez les mots d'arrêt (stopwords), la ponctuation et effectuez un stemming ou une lemmatisation.

   - Vectorisez le texte à l'aide de TF-IDF ou de plongements de mots (word embeddings).

2. **Extraction de caractéristiques (Feature Extraction) et réduction de la dimensionnalité :**
   - Appliquez la modélisation thématique (topic modeling) si nécessaire.
   - Utilisez LDA pour la réduction de la dimensionnalité.
3. **Classification bayésienne :**
   - Entraînez un classifieur bayésien naïf sur les entités extraites.
   - Implémentez la validation croisée pour évaluer le modèle.
4. **Optimisation et évaluation du modèle :**
   - Ajustez les hyperparamètres.
   - Évaluez le modèle à l'aide de matrices de confusion et d'autres métriques de performance.

Contenu du rapport

Le rapport comprendrait les sections suivantes :

1. **Introduction**
   - Vue d'ensemble de l'ensemble de données et des objectifs du projet.
2. **Méthodes**
   - Détails sur le prétraitement des données, l'extraction des caractéristiques (feature extraction) et les méthodes de classification utilisées.
3. **Résultats**
   - Résultats de la modélisation thématique (le cas échéant).
   - Performances du modèle de classification.
   - Discussion sur les facteurs discriminants.
4. **Conclusion**
   - Résumé des résultats et des implications ou applications potentielles.
5. **Ressources supplémentaires**
   - Toute information supplémentaire, telle que le code R complet.