

Classification de données de produits

Rakuten France Multimodal Product Data Classification

E. Hamza F. Quentin

hamza.errahj@student-cs.fr

fournier.quentin69@gmail.com

May 27, 2024

Sommaire

Contexte

Objectif du projet

Prétraitement du texte

Obtention de la matrice TF-IDF

Modélisation

Comparatif des modèles

Conclusion

Images

Bibliographie

Contexte

- ▶ Challenge de classification des produits de Rakuten France
- ▶ Objectif: Prédire le code type (prdtypecode) des produits
- ▶ Utilisation de données textuelles (désignation et description) et images
- ▶ Simplification: Pas d'utilisation des images

Objectif du projet

- ▶ Prétraiter le texte
- ▶ Convertir les données en matrice TF-IDF
- ▶ Appliquer différentes approches de classification
- ▶ Comparer les performances des modèles
- ▶ Conclure sur l'approche la plus appropriée

Prétraitement du texte

- ▶ Jeu de données: Environ 99,000 listes de produits en CSV
- ▶ Features: Désignation, Description, Images, Codes de produits
- ▶ Étapes:
 - ▶ Concaténation des champs textuels
 - ▶ Nettoyage des textes (minuscules, suppression des caractères spéciaux, stop words, lemmatisation)

Obtention de la matrice TF-IDF

- ▶ Utilisation de `TfidfVectorizer` de `scikit-learn`
- ▶ Conversion des textes en une matrice TF-IDF
- ▶ Exemple de code:

```
tfidfvectorizer = TfidfVectorizer(maxfeatures = 5000)Xtfidf = tfidfvectorizer.fit_transform(texts['text'])
```

Modélisation

- ▶ Algorithmes utilisés:
 - ▶ Support Vector Machines (SVM)
 - ▶ Régression Logistique
 - ▶ k-Nearest Neighbors (KNN)
 - ▶ Random Forest
 - ▶ Voting Classifier

Comparatif des modèles

Modèle	Précision
SVM	0.75
Régression Logistique	0.78
KNN	0.72
Random Forest	0.79
Voting Classifier	0.80

Table: Comparatif des performances des modèles

Conclusion

- ▶ Voting Classifier combine les forces de plusieurs algorithmes
- ▶ Meilleure généralisation sur les données non vues

Images

path/to/your/image1.png

Bibliographie

- ▶ TF-IDF Vectorizer
- ▶ Logistic Regression
- ▶ Support Vector Machines
- ▶ Classification Multiple
- ▶ Saving and Loading Models with Pickle