**Classification des textes avec le modèle naïf de Bayes**

Le but de ce TP est d’implémenter un classificateur de texte en s’appuyant sur la modélisation naïve de Bayes. On dispose d’un corpus de 1000 textes positifs et de 1000 textes négatifs :

* Diviser ce corpus en **corpus** **d’entrainement** et **corpus de test** (en premier temps opter pour une division aléatoire de 80% de textes pour l’entrainement et 20% pour la phase des tests). Les textes sont disponibles sur le réseau.
* Développer le modèle d’entrainement avec le modèle naïf de Bayes et sauvegarder le sur le disque (on utilise uniquement les textes du corpus d’entrainement dans cette phase).
* Évaluer votre modèle de classification en l’appliquant sur les textes du corpus de test. Utiliser le calcul de

une fois pour les textes positifs et ensuite pour les textes négatifs. En déduire un tableau comme :

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Expérience** | **Type corpus** | **Taille réseau** | **% positifs** | **% négatifs** | **% Moyenne** |
| Avec mots bruts | division |  |  |  |  |
| Avec mots bruts | Cross-validation |  |  |  |  |
| Avec mots en formes canoniques | … |  |  |  |  |
| … |  |  |  |  |  |

* Améliorer votre traitement en (1) créant des formes canoniques à vos mots à fin de réduire la taille du réseau Bayésien (mettre les noms et adjectifs en singulier, les verbes en infinitifs, etc.) (2) enlevant les mots non signifiants à la classification (à, et, être, avoir, etc.) (3) ne gardant que les mots plus grands que certaine taille, etc.
* Implémenter une méthode pour plutôt organiser dynamiquement votre corpus avec l’approche de cross validation (diviser le corpus en 10 dossiers, prendre un dossier pour le test et le reste pour l’entrainement. Répétons ceci 10 fois sur tous les dossiers et à chaque fois calculer les précisions et en faire la moyenne enfin).
* Conclusion : quels sont les meilleurs paramètres pour votre classificateurs ?

Durée de travail 3 semaines