**Rapport de résultats : classificateur Python par la méthode de Bayes**

Ce rapport parle des différents résultats obtenus par le groupe composé de MM. Danick Fort et Dany Jupille. Le script Python utilisé lors de la réalisation de ces tests est le fichier nommé « classifier.py ». Les différents résultats présentés dans ce document sont également visibles dans le fichier Excel « comparative-results.xlsx ».

Deux types de tests ont été effectués sur le classificateur Python.

1. Des tests de « cross-validation » : on divise en 10 dossiers ordrés les différents dossiers de sources de textes mis à disposition, et on prend 9 dossiers pour l’entraînement puis 1 dossier pour les tests. On effectue cette opération pour les 10 dossiers, une fois les sources « brutes » puis une fois les sources « taguées ».
2. Des tests de « division aléatoire » : il s’agit du même principe que les tests de « cross-validation », mais on mélange les fichiers dans les dossiers avant la division.

**Résultats des tests de « cross-validation »**

Les résultats donnés par ces tests sont assez inattendus. Tantôt les résultats sont parfaits, tantôt le taux d’échec est élevé.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Corpus de test** | **% positifs** | **% négatifs** | **% succès** | **% échec** |
| 1ère tranche de 10% de données brutes | 100.00% | 10.00% | 55.0% | 45.00% |
| 2ème tranche de 10% de données brutes | 98.00% | 63.00% | 80.5% | 19.50% |
| 3ème tranche de 10% de données brutes | 98.00% | 100.00% | 99.0% | 1.00% |
| 4ème tranche de 10% de données brutes | 100.00% | 100.00% | 100.0% | 0.00% |
| 5ème tranche de 10% de données brutes | 99.00% | 99.00% | 99.0% | 1.00% |
| 6ème tranche de 10% de données brutes | 100.00% | 100.00% | 100.0% | 0.00% |
| 7ème tranche de 10% de données brutes | 98.00% | 100.00% | 99.0% | 1.00% |
| 8ème tranche de 10% de données brutes | 100.00% | 99.00% | 99.5% | 0.50% |
| 9ème tranche de 10% de données brutes | 100.00% | 96.00% | 98.0% | 2.00% |
| 10ème tranche de 10% de données brutes | 43.00% | 100.00% | 71.5% | 28.50% |
| 1ère tranche de 10% de données canoniques | 100.00% | 20.00% | 60.0% | 40.00% |
| 2ème tranche de 10% de données canoniques | 96.00% | 73.00% | 84.5% | 15.50% |
| 3ème tranche de 10% de données canoniques | 98.00% | 100.00% | 99.0% | 1.00% |
| 4ème tranche de 10% de données canoniques | 100.00% | 99.00% | 99.5% | 0.50% |
| 5ème tranche de 10% de données canoniques | 99.00% | 99.00% | 99.0% | 1.00% |
| 6ème tranche de 10% de données canoniques | 100.00% | 100.00% | 100.0% | 0.00% |
| 7ème tranche de 10% de données canoniques | 98.00% | 99.00% | 98.5% | 1.50% |
| 8ème tranche de 10% de données canoniques | 99.00% | 95.00% | 97.0% | 3.00% |
| 9ème tranche de 10% de données canoniques | 99.00% | 97.00% | 98.0% | 2.00% |
| 10ème tranche de 10% de données canoniques | 44.00% | 98.00% | 71.0% | 29.00% |

|  |  |
| --- | --- |
| **Types de données** | **Efficacité du script (% succès total)** |
| Données brutes | 90.2% |
| Données canoniques | 90.7% |

**Résultats des tests de « division aléatoire »**

Les résultats donnés par ces tests sont déjà plus réalistes que ceux des tests précédents. On peut en déduire que les mots présents dans les fichiers « adjacents » dans la liste sont approximativement les mêmes et en même quantité.

De ce fait, mélanger les fichiers avant de faire l’entraînement accroît nettement la régularité des résultats.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Corpus de test** | **% positifs** | **% négatifs** | **% succès** | **% échec** |
| 1ère itération de 10% de données brutes | 88.00% | 92.00% | 90.0% | 10.00% |
| 2ème itération de 10% de données brutes | 94.00% | 87.00% | 90.5% | 9.50% |
| 3ème itération de 10% de données brutes | 90.00% | 88.00% | 89.0% | 11.00% |
| 4ème itération de 10% de données brutes | 92.00% | 91.00% | 91.5% | 8.50% |
| 5ème itération de 10% de données brutes | 94.00% | 83.00% | 88.5% | 11.50% |
| 6ème itération de 10% de données brutes | 94.00% | 91.00% | 92.5% | 7.50% |
| 7ème itération de 10% de données brutes | 94.00% | 83.00% | 88.5% | 11.50% |
| 8ème itération de 10% de données brutes | 93.00% | 89.00% | 91.0% | 9.00% |
| 9ème itération de 10% de données brutes | 95.00% | 87.00% | 91.0% | 9.00% |
| 10ème itération de 10% de données brutes | 96.00% | 92.00% | 94.0% | 6.00% |
| 1ère itération de 10% de données canoniques | 97.00% | 90.00% | 93.5% | 6.50% |
| 2ème itération de 10% de données canoniques | 94.00% | 89.00% | 91.5% | 8.50% |
| 3ème itération de 10% de données canoniques | 97.00% | 89.00% | 93.0% | 7.00% |
| 4ème itération de 10% de données canoniques | 97.00% | 93.00% | 95.0% | 5.00% |
| 5ème itération de 10% de données canoniques | 97.00% | 89.00% | 93.0% | 7.00% |
| 6ème itération de 10% de données canoniques | 93.00% | 91.00% | 92.0% | 8.00% |
| 7ème itération de 10% de données canoniques | 92.00% | 91.00% | 91.5% | 8.50% |
| 8ème itération de 10% de données canoniques | 92.00% | 89.00% | 90.5% | 9.50% |
| 9ème itération de 10% de données canoniques | 94.00% | 92.00% | 93.0% | 7.00% |
| 10ème itération de 10% de données canoniques | 96.00% | 92.00% | 94.0% | 6.00% |

|  |  |
| --- | --- |
| **Types de données** | **Efficacité du script (% succès total)** |
| Données brutes | 90.7% |
| Données canoniques | 92.7% |

D’une manière générale, les résultats sont bons (> 90%).

**Journal de bord**

|  |  |  |
| --- | --- | --- |
| Date | Etudiant | Travail |
| 04.04.2014 | Danick Fort | Création de la fonction d’apprentissage : créer un dictionnaire avec chaque mot rencontré et sa fréquence. |
| 04.04.2014 | Dany Jupille | Création du squelette du code général. Chargement de la liste des mots à ignorer. Filtre de la ponctuation dans les textes. |
| 11.04.2014 | Danick Fort | Finition de la fonction d’apprentissage. |
| 11.04.2014 | Dany Jupille | Version classe du script, création de la classe gérant Bayes. |
| 25.04.2014 | Danick Fort | Finition de la version procédurale du script (fonctionnelle et complète). |
| 25.04.2014 | Dany Jupille | Version classe du script, ajout d’un générateur pour parcourir les fichiers et gestion de la liste des mots courants. |
| 02.05.2014 | Danick Fort | Implémentation d’un système de *cross-validation* pour les tests. |
| 02.05.2014 | Dany Jupille | Version classe de l’application complète. Remplissage d’un tableau Excel des pourcentages de réussite. |