Cours TAL – Labo 7 : Classification de dépêches d'agence avec NLTK

Objectif

L'objectif de ce labo est de réaliser des expériences de **classification de documents** avec la boîte à outils NLTK sur le corpus de dépêches Reuters. Le labo est à effectuer en binôme. Le rendu sera un *notebook* Jupyter présentant vos choix, votre code, vos résultats et les discussions. Le labo sera jugé sur la qualité des expériences et sur la discussion des différentes options explorées.

Documentation

- *Livre NLTK* : chapitre 2 pour le corpus Reuters, chapitre 6 pour la classification, et *howto* pour les classifieurs.
- *Introduction to Information Retrieval*, chapitre 13, pour les scores de certaines étiquettes.

Description des expériences

1. **Données** : les dépêches du corpus Reuters, tel qu'il est fourni par NLTK. Veuillez respecter la division en données d'entraînement et données de test.
2. **Hyperparamètres** : dans ce qui suit, veuillez étudier au moins deux hyperparamètres, avec pour chacun au moins deux valeurs différentes, par exemple :
   - Suppression des *stopwords* → {oui, non}
   - Lemmatisation → {oui, non}
   - Capitalisation → {d'origine, tout en minuscules}
   - Vocabulaire → {unigrammes, unigrammes et bigrammes}
   - Représentation des documents → {Bernoulli, multinomiale}
   - Attributs supplémentaires → {aucun, longueur de la dépêche, rapport tokens/types}
   - Classifieurs → {Naive Bayes, Decision Tree, Maximum Entropy}
3. Veuillez définir et entraîner **trois classifieurs binaires** : chacun prédit si une dépêche est étiquetée ou non avec la catégorie respective. Le premier classifieur binaire sera pour l'étiquette '*money-fx*', le deuxième concernera '*grain*', et le troisième sera pour '*nat-gas*'.
4. Pour chacune des étiquettes, veuillez diviser les données d'entraînement en 80% *train* et 20% *dev*, en respectant les proportions des étiquettes (stratification). Veuillez **trouver les meilleures valeurs** des hyperparamètres en entraînant sur *train* et en testant sur *dev*.
5. Veuillez calculer sur les **données de test, les scores de rappel, précision et f-mesure** de chacun des trois classifieurs binaires, avec les meilleurs hyperparamètres.

6. Veuillez entraîner **un quatrième classifieur multi-classe** qui assigne une seule étiquette parmi '*money-fx*', '*grain'*, '*nat-gas*' et '*other*'. Vous pouvez choisir les valeurs des hyperparamètres selon vos observations du point (4). Note : quelques dépêches sont annotées avec plusieurs étiquettes : dans ce cas, gardez seulement la première.

7. Veuillez donner les scores de **rappel, précision et f-mesure du classifieur multi-classe** pour chacune des étiquettes '*money-fx*', '*grain'*, et '*nat-gas*'. Comment ces scores se comparent-ils à ceux des trois classifieurs binaires du point (5) ? Quelle est donc la meilleure stratégie de classification ?