Cours TAL – Labo 6 : Classification de dépêches d'agence avec NLTK

Objectif

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

Description des expériences

1. **L'objectif général** est d'explorer au moins deux aspects parmi les choix qui se posent lors de la création d'un système probabiliste de classification de textes.
2. **Données** : les dépêches du corpus Reuters, tel qu'il est fourni par NLTK. Vous respecterez notamment la division en données d'entraînement (*train*) et données de test.
3. **Hyper-paramètres** : veuillez étudier au moins deux hyperparamètres. Pour chacun, veuillez comparer au moins deux valeurs et indiquer laquelle fournit le meilleur score. Vous pourrez choisir parmi les hyperparamètres suivants :
    - options de prétraitement des textes : *stopwords*, lemmatisation, tout en minuscules.
    - options de représentation : présence/absence de mots indicateurs, nombre de mots indicateurs ; présence/absence/nombre de bigrammes, trigrammes ; autres traits : longueur de la dépêche, rapport tokens/types.
    - classifieurs et leurs paramètres : divers choix possibles (voir la documentation NLTK).
4. 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*'.
5. Pour chacun des classifieurs**,** optimisez les hyperparamètres sans toucher aux données de *test* NLTK. Divisez les données d'entraînement NLTK en 80% *train* et 20% *dev*, et choisissez les options qui donnent les meilleurs scores sur *dev*.
6. Veuillez donner les scores de rappel, précision et f-mesure de chacun des trois classifieurs, avec les meilleurs hyperparamètres, sur les données de test.
7. Veuillez définir **un quatrième classifieur multi-classe** qui assigne une étiquette parmi quatre : les trois choisies ci-dessus plus la catégorie '*other*'. Vous devrez nettoyer les données, car un petit nombre de dépêches sont annotées avec plusieurs étiquettes : dans ce cas, gardez seulement la première.

8. Veuillez donner les scores de rappel, précision et f-mesure de ce classifieur pour chacune des trois étiquettes choisies. Comment les scores se comparent-ils à ceux des trois classifieurs binaires ?

9. **Documentation** : livre NLTK, chapitre 2 pour accéder au corpus Reuters et le chapitre 6 pour la classification ; puis http://www.nltk.org/howto/classify.html pour les classifieurs dans NLTK ; enfin, *Introduction to Information Retrieval* (https://nlp.stanford.edu/IR-book/information-retrieval-book.html), chapitre 13, pour une discussion générale de méthodes de classification, et des exemples de scores obtenus sur certaines étiquettes.