Cours TAL – Mini-projet individuel comptant comme travail écrit

Classification de dépêches d'agence avec NLTK

Andrei Popescu-Belis, le 28 mai 2021

Modalités du projet

L'objectif de ce projet est de réaliser des expériences de classification de documents sous NLTK avec le corpus de dépêches Reuters. Le projet est à effectuer en binôme, et les résultats de chaque binôme seront différents. Le projet sera jugé sur la qualité des expériences (correction méthodologique) mais aussi sur la discussion des différentes options explorées dans le projet.

Vous devez remettre un *notebook* Jupyter présentant vos choix, votre code, vos résultats et les discussions. Le *notebook* devra déjà contenir les résultats des exécutions, mais pourra être ré-exécuté par le professeur ou l'assistant en vue d'une vérification.

Vous ferez en outre faire une courte présentation orale (10 min.) et répondre aux questions sur votre projet (10 min.) lors d'une séance sur Teams avec le professeur et l'assistant.

Description des expériences

1. **L'objectif général** est d'explorer au moins deux aspects parmi les multiples 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** : la définition d'un classifieur comporte un grand nombre de choix de conception, dans plusieurs dimensions. Dans ce projet, et pour chaque objectif de classification (voir ci-dessous) vous devez étudier au moins deux dimensions. Pour chacune, vous devez comparer au moins deux options et indiquer laquelle fournit le meilleur score, en essayant d'expliquer pourquoi. Vous pourrez choisir parmi les options suivantes :
    - 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).

4. **Pour chacun des classifieurs demandés ci-dessous,** vous choisirez les meilleurs hyperparamètres sans regarder les scores sur les données de *test* NLTK. Vous diviserez donc les données d'entraînement NLTK en 80% *train* et 20% *dev*, et vous choisirez les options qui donnent les meilleurs scores sur *dev*. Au final, vous donnerez les scores sur les données de test.

5. Veuillez d'abord définir et entraîner **trois classifieurs binaires**, correspondant chacun à une catégorie de votre choix. Chaque classifieur prédit si une dépêche appartient ou non à la catégorie respective. Le premier classifieur binaire sera pour une étiquette que vous choisirez parmi les trois suivantes : '*money-fx*', '*interest*', ou '*money-supply*'. Le deuxième concernera une étiquette parmi : '*grain*', '*wheat*', '*corn*'. Enfin, le troisième sera choisi parmi : '*crude*', '*nat-gas*', '*gold*'.

    - Veuillez donner les scores de rappel, précision et f-mesure de chacun des trois classifieurs, avec les meilleurs hyperparamètres, sur les données de test.

6. Veuillez définir **un quatrième classifieur multi-classe** qui assigne une étiquette parmi quatre : les trois choisies ci-dessus plus la catégorie '*other*'. Vous devrez adapter légèrement les données, car un petit nombre de dépêches sont annotées avec plusieurs étiquettes (gardez seulement la première).

    - Veuillez donner les scores de rappel, précision et f-mesure de ce classifieur séparément pour chacune des trois étiquettes choisies, et comparer ces scores à ceux des trois classifieurs binaires précédents.
    - Veuillez également comparer ces scores à ceux de la littérature ci-dessous.

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