Cours TAL – Mini-projet individuel comptant comme travail écrit

Classification de dépêches d'agence avec NLTK

Andrei Popescu-Belis, le 12 mai 2020

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 individuel : vous êtes responsable des différentes options choisies, et en principe les résultats de chaque projet seront différents. Le projet sera jugé sur la qualité des expériences (correction méthodologique) mais aussi sur la discussion des options explorées.

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 en vue d'une vérification.

Vous devrez en outre faire une courte présentation orale (5-7 min.) et répondre aux questions sur votre projet (5-7 min.) lors d'une séance sur Teams (15 min.) 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 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 explorerez deux dimensions. Pour chaque dimension, vous comparerez au moins deux options pour trouver laquelle fournit le meilleur score, et vous tenterez d'expliquer pourquoi. Vous pourrez choisir parmi les options suivantes :
    a. options de prétraitement des textes : *stopwords*, lemmatisation, tout en minuscules.
    b. 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.
    c. classifieurs et leurs paramètres : divers choix possibles (voir la documentation).

4. **Objectif de classification** : vous devrez construire quatre classifieurs. Vous choisirez les meilleurs hyper-paramètres pour chaque classifieur sans regarder les résultats sur les données de test NLTK, mais en divisant les données d'entraînement NLTK en 80% *train* et 20% *dev*. Vous ferez ensuite l'entraînement final sur l'intégralité des données d'entraînement.

    a. Veuillez d'abord définir et entraîner trois classifieurs binaires, correspondant à trois catégories de votre choix. Chaque classifieur prédit si une dépêche appartient ou non à la catégorie, i.e. si elle doit recevoir ou non l'étiquette respective. Veuillez construire un premier classifieur binaire pour une étiquette que vous choisirez librement parmi les trois suivantes : '*money-fx*', '*interest*', ou '*money-supply*'. Le deuxième classifieur binaire concernera une étiquette de votre choix 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 que vous avez conçus et entraînés.

    b. On vous demande également de définir un quatrième classifieur qui assigne l'une des trois étiquettes que vous avez choisies ci-dessus plus la catégorie '*other*' (il assigne donc une seule étiquette parmi quatre). Vous devrez adapter légèrement les données, car un très petit nombre de dépêches (combien ?) sont en réalité annotées avec plusieurs de ces étiquettes, et vous n'en retiendrez que la première.

        – Veuillez évaluer ce classifieur en termes de rappel, précision et f-mesure pour chacune des trois étiquettes choisies ci-dessus, et comparer ces trois scores à ceux des trois classifieurs binaires précédents.

5. **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.