Travail pratique # 2

CLASSIFICATION DE TEXTES

IFT-7022
Luc Lamontagne
Automne 2016

Instructions :

- Travail individuel.
- Rapport et logiciel à remettre, via pixel, au plus tard le 21 novembre.
- Ce travail est noté sur 100 et vaut 20% de la note de session.

1. Objectif

L'objectif de ce travail est de mener quelques expérimentations afin de vous familiariser avec les techniques de classification de texte. Votre travail consistera à évaluer un ou quelques algorithmes de classification (au choix) sur un corpus de textes.

Les sections suivantes présentent les deux options que je vous propose. Cependant, vous êtes autorisé à utiliser un autre corpus si vous le souhaitez (merci de me consulter).

Vous pouvez, pour ce travail, utiliser toutes librairies logicielles qui peuvent vous aider à faire le prétraitement des textes, à construire des modèles de classification et à évaluer ces modèles.

2. Option 1 : Classification de questions

À partir du corpus utilisé dans le TP1 sur la classification de question (*questions.txt*), comparez la performance de deux algorithmes différents d'apprentissage (au choix) pour classifier le type de chacune des questions.

Comme il s'agit d'un problème multiclasses, vous pouvez soit construire un seul classificateur global qui discrimine parmi les différentes classes ou des classificateurs individuels pour chacune des classes (voir chapitre 7 de Jurafski 3^{e} édition pour plus d'informations).

Évaluez la performance de l'algorithme en terme de précision, de rappel et d'exactitude. Discuter des principales sources d'erreur de vos algorithmes.

Pour plus d'informations sur la classification de questions, vous pouvez vous référer aux articles suivants:

- Xin Li et Dan Roth (2002) *Learning Question Classifiers*, COLING'02. http://ucrel.lancs.ac.uk/acl/C/C02/C02-1150.pdf
- P. Blunsom, K. Kocik, J. Curran (2006) *Question Classification with Log-Linear Models*, SIGIR'06. http://www.it.usyd.edu.au/~james/pubs/pdf/sigir06qc.pdf

3. Option 2 : Analyse de sentiments

La 2^{e} option consiste à apprendre à classifier des textes selon leur polarité (positive ou négative). Tout comme pour la première option, vous devez comparer au moins deux algorithmes différents d'apprentissage (au choix), à évaluer la performance de l'algorithme en terme de précision/rappel/exactitude et à discuter des principales sources d'erreur de vos algorithmes.

Voici quelques corpus que vous pouvez utiliser pour mener vos expérimentations (au choix) :

- Movie reviews dataset : http://www.cs.cornell.edu/People/pabo/movie-review-data/
- Mutli-domain sentiment dataset : http://www.cs.jhu.edu/~mdredze/datasets/sentiment/

- LARA review dataset : http://sifaka.cs.uiuc.edu/~wang296/Data/index.html
- Restaurant review dataset (évaluation de 1 à 5) : http://www.cs.cmu.edu/~mehrbod/RR/
- Twitter Sentiment Corpus : http://www.sananalytics.com/lab/twitter-sentiment/
- UMICH SI650 – Sentiment Classification : https://inclass.kaggle.com/c/si650winter11

4. À remettre

- Votre projet et vos fichiers d'expérimentations (si différents de ceux proposés).
- Un rapport qui décrit :
    - Les outils que vous avez utilisés et/ou le code que vous avez développé;
    - Les expérimentations que vous avez menées;
    - Les résultats que vous avez obtenus;
    - Les conclusions que vous tirez de vos expérimentations;
    - Des instructions pour installer et exécuter votre projet.

5. Évaluation du travail

▪ Choix de logiciels ou implémentation de code	10%
▪ Démarche générale	10%
▪ Prétraitement des textes	10%
▪ Expérimentations	30%
▪ Évaluation et analyse des résultats	30 %
▪ Qualité du rapport	10 %