Masters LITL
Fouille de données, Fouille de textes
Évaluation : *Les titres des publications de HAL*

Le rapport (4 pages maximum tout compris, format PDF) doit être déposé sur la page IRIS d'ici le 8 décembre.

Nous avons extrait du dépôt d'archives ouvertes HAL (hal.archives-ouvertes.fr) environ 50 000 notices bibliographiques de travaux académiques rédigés en français, dont nous avons retenu les informations suivantes (dans l'ordre des colonnes) :

1 Domaine
2 Type de support
3 Année de publication
4 Nombre d'auteurs
5 Titre

Nous nous sommes limités à 3 domaine (disciplines) : *Lettres* (littérature), *Linguistique* et *Informatique*, ainsi qu'à trois types de support : Articles de journaux (*ART*), communications dans un congrés (*COMM*) et chapitres d'ouvrage (*COUV*).

L'archive disponible sur la page du cours contient plusieurs fichiers de données au format CSV (UTF-8), qui ne diffèrent que par la quantité de données disponible. En fonction de la puissance de votre machine et des choix que vous ferez en termes de méthodes, vous devrez peut-être limiter le nombre d'items traités.
Le fichier `litl-exam.csv` contient l'ensemble des données disponibles (plus de 50 000 titres). Les 3 fichiers `litl-exam-`*`NNNN`*`.csv` contiennent respectivement un échantillon aléatoire de ces données de 3 000, 10 000 et 20 000 items. Vous indiquerez pour chaque question quel fichier vous avez utilisé pour y répondre, en essayant bien entendu d'utiliser le fichier le plus volumineux possible.
NOTE : le recouvrement entre ces différents échantillons n'a pas été contrôlé. Il est donc fortement déconseillé d'utiliser un de ces fichiers comme corpus de test pour un modèle entraîné sur un autre fichier.

1/ Analyse des champs non textuels

En utilisant Weka ou tout autre outil d'analyse des données structurées, dégagez quelques régularités ou tendances observables dans les *champs non-textuels* (i.e. tous sauf le titre).

Par exemple vous pouvez vous poser des questions du type : le nombre d'auteurs varie-t-il en fonction du type de publication ou de l'année, les supports sont-ils représentés de la même façon entre les disciplines, etc.

Essayez de dégager 2 faits (ou tendances) de ce type en indiquant la façon dont vous vous êtes pris pour les mettre en évidence.

2/ Analyse du titre en fonction du domaine

Procédez dans un second temps à l'analyse des *titres* en les croisant avec le domaine des publications.

En indiquant quel outil, méthode et paramétrage vous utilisez, vous dégagerez les principaux *traits textuels* permettant de distinguer les titres des publications de chaque discipline par rapport aux autres. Vous indiquerez un maximum de 10 traits par discipline, en indiquant (si vous la connaissez) la valeur prédictive de chacun d'eux et proposerez une interprétation (courte) de l'ensemble.

3/ Classification automatique d'un titre par domaine

1/ Construisez un modèle de classification permettant de déduire du titre d'une publication la discipline dont elle relève. Vous prendrez soin de décrire **précisément** la méthode utilisée (modèle, traits, critères de sélection, paramètres éventuels) et son évaluation (modalité d'évaluation et score global obtenu).

Vous argumenterez vos choix techniques en fonction de vos connaissances et de vos expérimentations (vous pouvez indiquer le cas échéant quels modèles vous avez envisagés, évalués puis rejetés).

Vous donnerez une évaluation **quantitative ET qualitative** du classifieur retenu et de la difficulté globale de la tâche.

Pour ce classifieur :

1/ Identifiez **deux erreurs de classification** commises par votre modèle, que vous commenterez en proposant une explication.

2/ Intégrez un ou plusieurs **traits non-textuels** (type, année, nombre d'auteurs) dans votre classifieur, et observez la variation de ses performances. Commentez les résultats.

3/ Proposez et implémentez (en Python) **un trait (ou famille de traits) linguistique construit sur le titre** en justifiant votre choix, et évaluez l'impact de ce(s) trait(s) sur l'efficacité de votre modèle quand vous l' (les) ajoutez aux autres prédicteurs.