Modélisations mathématiques

4. Travail à rendre et création des *runs*

Solen Quiniou

solen.quiniou@univ-nantes.fr

IUT de Nantes

Année 2017-2018 – Info 2

Plan du travail à rendre

Plan du travail à rendre

Objectifs du travail

L'objectif de ce dernier travail sur les modèles de langage est le suivant :

1. Rédiger un court rapport indiquant les choix que vous avez faits pour détecter les auteurs, dans vos systèmes de reconnaissance.

2. Créer des *runs* avec vos systèmes de reconnaissance d'auteurs, sur le fichier de test qui vous sera mis à disposition pendant quelques jours.

→ **Le travail rendu aura été réalisé en binôme (préférablement) ou seul.**

Plan du travail à rendre

Principe d'une compétition en TAL

Une compétition en TAL se déroule en 3 phases :

1. Des données d'apprentissage vous sont fournies et vous disposez de quelques mois pour mettre au point vos ou votre systèmes de reconnaissance par rapport à la tâche fixée par les organisateurs de la compétition.

→ La tâche choisie ici est la reconnaissance des auteurs des phrases d'un fichier.

2. Des données de test vous sont ensuite fournies et vous disposez de quelques semaines pour soumettre un ou plusieurs *runs* sur ces données de test. Chaque *run* correspond au résultat obtenu avec un de vos systèmes de reconnaissance sur les données de test.

→ Un *run* correspond ici à un fichier de références d'auteurs qui contient l'auteur trouvé pour chacune des phrases du fichier de test.

3. Les organisateurs évaluent vos *runs* et ceux des autres participants pour réaliser un classement des différents systèmes et présentent les résultats à l'issue de la compétition.

Création des *runs* pour notre compétition

1. Vous devez tout d'abord récupérer le fichier de test `sentences.txt` sur madoc, à partir du **10 janvier 2018**, et le placer dans le répertoire `data/author_corpus/test`.

→ Le fichier `sentences.txt` contient des phrases d'auteurs inconnus.

2. Vous devez ensuite choisir vos 2 meilleurs systèmes de reconnaissance, parmi vos classes `AuthorRecognizer1`, `UnknownAuthorRecognizer1`, `UnknownAuthorRecognizer2`...

3. Avec la première classe choisie, vous devez créer le fichier *run* `data/author_corpus/test/authors-hyp1.txt`, en utilisant le fichier de test `data/author_corpus/test/sentences.txt`.
4. Avec la seconde classe choisie, vous devez créer le fichier *run* `data/author_corpus/test/authors-hyp2.txt`, en utilisant le fichier de test `data/author_corpus/test/sentences.txt`.
   → La création de chaque *run* se fait comme pour créer le fichier `authors_100sentences_hyp-1.txt` à partir du fichier `authors_100sentences.txt` (voir la méthode `main(..)` de la classe `AuthorRecognizer1`).

Travail à rendre

- L'archive à créer se nommera **obligatoirement** `nomEtudiant1-nomEtudiant2.zip` et contiendra :
  - Le code source commenté des 2 projets Eclipse `Etudiant-mm_langModel` et `Etudiant-mm_authorReco`.
    - ⋆ Les fichiers `authors-hyp1.txt` et `authors-hyp2.txt`, correspondant aux *runs*[1] de vos systèmes de reconnaissance sur le fichier `sentences.txt`, doivent se trouver dans le répertoire `data/authors_corpus/test`.
  - Un rapport de 5-10 pages, au format `pdf`, à placer dans un répertoire `report`, contenant :
    - ⋆ l'état d'avancement de l'implémentation des 2 projets ainsi que les problèmes rencontrés et les solutions proposées ;
    - ⋆ la description des algorithmes de vos systèmes de reconnaissance ;
    - ⋆ les caractéristiques des modèles *n*-grammes que vous avez utilisés (langue, ordre des modèles, type de modèles, type de corpus d'apprentissage. . .) ;
    - ⋆ les performances obtenues avec vos systèmes de reconnaissance sur le fichier de validation `data/author_corpus/validation/sentences.txt`.

→ L'archive est à déposer sur `http://filex.univ-nantes.fr` et le lien FileX est à envoyer par mail, à votre enseignant de TD, avant le **jeudi 25 janvier 2018, 23h55** (en modifiant l'expiration du fichier à 30 jours).

1. Si vous ne soumettez qu'un seul *run*, le fichier se nommera `authors-hyp1.txt`.