Projet Traitement Automatique de Corpus

26 janvier 2019

1 Consignes

Vous allez faire un projet, par groupe de deux. Vous devez tout écrire. Ne pas utiliser de bibliothèques tièrces (et notament pas de JSON). L'objectif est de réaliser un programme qui réalise un ensemble de tâche automatisé sur un corpus de dialogue.

2 Les fonctionnalités à implémenter

Votre programme doit procéder automatiquement aux traitements suivant :

1. Télécharger l'archive data-dstc.tar.gz du Dialogue State Tracking de camdial. Utilisez la commande wget. Si le fichier existe déjà, alors ne pas la télécharger. (2)
2. Décompresser l'archive avec tar. (1)
3. Créer un dossier "reponses" dans lequel vous mettrez tous les fichiers crées automatiquement par la suite. (1)
4. Lister dans un fichier "dates" toutes les dates de tous les fichiers (une date par ligne). (2)
5. Créer un fichier "statistiques" (pour l'ensemble des logs) dans lequel apparaîtra :
   - Le nombre total de phrases pour la clef "transcript" (dans "ouput").Compter avec grep (2)
   - Le score moyen total. Pour celà, utiliser le programme python mean.py (la commande est la suivante python mean.py chemin/du/fichier). Il fait la moyenne de tous les nombres d'un fichier, avec un nombre par ligne. (2)
6. Dans tous les fichiers de Apr11_S1, transformer les dates dans ce format : JJ-MM-AA. Utiliser des groupes de capture avec sed. (3)
7. Lister dans un fichier les couples (session-id, nombre de tours) pour chaque fichier log.jon. Pour cela, compter le nombre de tours dans "turns". Utiliser la command find avec l'option "-exec" pour executer une opération sur chaque fichier. S'inspirer de cette réponse. (3)

8. Créer un fichier moyenne, dans lequel seront repertoriés les couples (session-id, score moyen). (4)

Si il y suspicion de copier-coller, travail à plus de 2 ou autre entourloupe, c'est zéro pour tout le monde. Que ce soit le copieur, ou le copié. Je ferai passer votre code dans un programme pour vérification automatique + vérification manuelle.

Evaluation J'évaluerai votre projet lors de la dernière scéance. Il faut que le programme fonctionne sur les machines linux de l'université.