INGÉNIERIE DES CONNAISSANCES

Projet

PROJET DU SEMESTRE

- Vous êtes une équipe d'ingénieurs et d'ingénieures linguistes et vous avez été recrutées pour participer à un projet de recherches visant à étudier la variation sociolinguistique diastratique dans deux types de données : des données issues de l'oral et de la communication médiée par les réseaux.

- Variables listées par les chercheurs :
  - temps verbaux
  - négation complète (il *ne mange pas/rien/jamais/point/aucunement…*) vs. négation incomplète (*il mange pas/rien/jamais/point/aucunement…*)
  - absence/présence de /y/ dans *tu* (*tu as pris / t(u) as pris*), du schwa dans les clitiques (*ce*, *de*, *je*, *le*, *me*, *ne*, *que*, *se*, *te*) : c(e) matin, mie d(e) pain, j(e) prends l(e) métro, il m(e) parle, je n(e) viens pas, pas qu(e) toi, il s(e) perd, je t(e) parle

PROJET DU SEMESTRE

- Étape 1 : faire l'inventaire des sources de connaissances disponibles (corpus)
- Étape 2 : acquisition des connaissances (corpus)
- Étape 3 : modélisation formelle des connaissances acquises

- Étape 4 : modélisation formelle des connaissances à ajouter (modèles d'annotation)
- Étape 5 : ajout de nouvelles connaissances (annotation automatique)
- Étape 6 : formalisation des connaissances obtenues (création d'une nouvelle version du corpus)

PROJET DU SEMESTRE

- Format du projet :
  - Dossier
  - Nouvelles versions du corpus ESLO : XML et TXM
  - Nouvelle version du corpus 88MSMS : XML et TXM
- Date limite de remise du projet :
  - Une semaine avant la date limite de remise des notes (à préciser)
- Modalités de remise du projet :
  - Dépôt Moodle pour le dossier
  - Partage (Wetransfer ou autre) des nouvelles versions des corpus

PROJET DU SEMESTRE

- Dossier :
  - Doit être rédigé !
  - Première page : titre du dossier, composition du groupe
- Organisation libre du dossier, mais doivent obligatoirement apparaître :
  - Description des corpus utilisés au départ : thématique, taille, créateurs, annotation présente, métadonnées présentes…
  - Vos raisonnements, le parcours suivi : pourquoi et comment vous avez fait quoi
  - Méthodes utilisées pour annoter, transformer…
  - Les scripts doivent apparaître en annexe, avec une description au moins minimale
  - Description des corpus obtenus : pourquoi ces formats, que permettent-ils de faire etc.
  - Retour d'expérience : on aurait pu faire ça, on aurait pu faire comme ça

PROJET DU SEMESTRE

- Commentaires et retours d'expérience importants !
- Si vous avez partagé les tâches, indiquez qui s'est occupé de quoi
- Le résultat obtenu n'est pas noté seul : pour la notation, une part plus importante est accordée au parcours suivi.
  - Le script fonctionne et fournit un bon résultat, mais rien n'est expliqué ou commenté : négatif
  - Ça ne fonctionne pas, on ne peut rien faire avec le corpus mais tout est expliqué et on propose d'autres façons de faire : positif

PROJET DU SEMESTRE

- Nouvelles versions des corpus à rendre :
  - XML : votre corpus « de base » avant import dans TXM
  - TXM : le fichier obtenu dans TXM après import. Ce fichier s'obtient, dans TXM, en faisant : clic-droit sur le corpus > exporter > corpus en format binaire

- Cf. démo sur TXM