Data sciences et langage 2019/2020
Anne-Lyse Minard : anne-lyse.minard@univ-orleans.fr

Projet / dossier

Vous développerez en groupe un programme permettant de faire une première classification des réponses à une question du grand débat. Ce programme pourra reposer sur des scripts python et des graphes unitex. Les graphes peuvent permettre d'effectuer des pré-annotations (par exemple de la négation).

Modules python :
- Spacy : https://spacy.io/
- NLTK : http://www.nltk.org/
- expression régulière : re

Pour le **20 janvier**, vous déposerez sur CELENE un rapport chacun de 10 pages maximum qui contiendra :

1. l'analyse de 5 contributions au choix qui peuvent être difficile à analyser par un système automatique. Vous expliquerez la raison de cette difficulté (présence de coréférences, ambiguïté, information implicite, etc.). Vous veillerez à avoir des exemples de phénomènes différents.

2. la description du système développé, en particulier les étapes du traitement et les méthodes utilisées.

3. l'analyse de quelques résultats obtenus en mettant en évidence les problèmes encore à résoudre.

4. une réflexion sur le traitement de transcription automatique de l'oral.
Nous avons récupéré les sous-titres de youtube (automatique) d'un débat du monde de l'entreprise au CCI d'Orléans (vidéo : https://www.youtube.com/watch?v=OlhOAsCyw2s, sous-titres : sur CELENE).
Vous expliquez 5 problématiques du traitement de données provenant de la langue parlée en illustrant avec des exemples.