hackaTAL : *extraction de citations*

groupe : **Alèd** (ça n'allait vraiment pas très bien)

Natacha, Florian, et Léna (M1 pluriTAL)

Notre approche : les regex

On voulait dans un premier temps extraire les citations (discours direct) à l'aide d'une regex. (travail avec script python, module re)

Regex finalement choisie : `([^«»]*)(«.+?»)([^«]*)`

Avantage : attribut "group" des objets *match* du module re (cela nous a permis au début de faire un petit concordancier)

Ce que nous sommes parvenus à faire

- **concordancier** : repérer les schémas qui se répétaient autour des citations ! (inversion sujet verbe à droite, usage assez consistent des verbes de parole...)
- **extraire une partie des citations** avec guillemets “français” (avec position du début, et fin de citation)
- repérer les entités nommées autour des citations (filtre pour récupérer les “organisations” et “personnes”, car plus probable que le discours de ces entités soit cité)

Difficultés rencontrées

- différents guillemets utilisés : “ ”, « » ;
- beaucoup de guillemets fermants mais pas ouvrants ! (en lisant les contenus textuels des articles, il est donc difficile de repérer les citations) ;
- beaucoup plus difficile de gérer les relations entre les entités nommées et les citations :(
- ce début de travail avec les regex ne permet de repérer que les citations en discours direct.

En conclusion

Un défi un peu difficile à notre niveau, mais nous sommes contents d’avoir pu participer ! :)