# td2

January 16, 2023

## 1 TD2 Extraction de données

### 1.1 La dernière fois

- Vous avez effectué la lecture d'un fichier et un premier pas de nettoyage.
- Vous avez obtenu un fichier `ding2-1-texte.txt`.

On souhaite maintenant récupérer des informations sur le contenu textuel de la partie. Vous pouvez faire cela avec le fichier `ding2-1-texte.txt`. Dans ce fichier, nous avons un tour de parole (TDP) par ligne avec (1) le numéro de TDP, (2) le marker des locuteurs ('B', 'R', 'O', 'W', 'Y'), et (3) le texte.

```
0001 R   ressource
0002 R   euh où il y a le 5 bah je vais pouvoir collecter en fait prélever euh récolter la ress
0003 O   [discussion]
```

Si vous n'avez pas réussi à obtenir ce fichier, vous pouvez également utiliser ce qu'on vous fournit sur Arche.

### 1.2 Exo 2.1

Maintenant, lire le fichier `ding-2-1-texte.txt` et identiifer :

- (1) le nombre de tours de parole
- (2) le nombre total de mot
- (3) le nombre moyen de mot par tour de parole
- (4) le nombre moyen de lettre par mot
- (5) le nombre de questions
- (6) le nombre de locuteurs et leur nom

Enregistrer ces information dans un nouveau fichier : `res-ding2-1.txt`

[ ]:
```python
import os

# initiate variables
nb_tdp = 0
nb_mot = 0
nb_lettre = 0
nb_q = 0
```

```
ave_mot_tdp = 0
ave_lettre_mot = 0
vocab = set()
set_spk = set()
textf = 'ding2-1-texte.txt'
resf = 'res-ding2-1.txt'
result = ''


# calculate nb of speech turns
# if not os.path.exists(textf):
#     raise FileNotFoundError(f"File {textf} not found. Check the path.")
with open(textf, 'r') as inf:
    # TODO: mettre vos codes ici
    pass
```

### 1.3 Exo 2.2

Maintenant nous allons extraire les informations plus fines : extraire les informations (1) - (5) dans exo 2.1 de chaque locuteur, et afficher vos résultats.

Hint : vous pouvez utiliser un **dictionnaire** pour stocker les informations par locuteur.

Ressource : https://realpython.com/python-dicts/

[ ]:
```
# TODO
```