Spécialité DAC **Module LODAS** **TD/TME du 3 février 2021**

Ce TME porte sur le traitement de corpus avec Orange.

1. **Corpus**

   On va charger deux corpus, le premier est téléchargeable à l'URL suivant :

   https://nuage.lip6.fr/s/BFqRSnKqkwn3ni5 Pour cela, on utilise le widget « import document »

   Le second s'obtient pas le widget « wikipedia », avec des mots clefs au choix. Je propose de prendre deux ou trois mots clefs bien distincts, par exemple « covid-19 » et « théâtre » ou « dorique » de façon à extraire une cinquantaine ou une centaine de textes.

   Regarder comment on passer des corpus à des datas et des datas à des corpus avec les data tables.

2. **Visualisation du corpus**

   Avec le widget « corpus viewer », regarder l'ensemble des textes des corpus. Vous pouvez en sélectionner quelques-uns ou tous

3. **Utiliser le sentiment analysis**

   Avec le l'option multilingual analysis pour le français, regardez la tonalité affective des textes.

4. **Topic Modeling et word cloud**

   Utilisez le « Topic Modeling » et « Word Cloud » sur quelque textes ou tous.

   Que remarquez-vous ?

5. **Préprocessing**

   Utiliser le préprocesseur qui opère un séquence d'opérations sur les textes.

   Cela permet de mettre en œuvre différentes transformations, par exemple, mettre tout en minuscule.

   La seconde étape est la tokénisation. L'idéal, dans notre contexte est d'utiliser regexp avec \w+

   Enfin, on opère les transformation, en particulier pour filtrer les « mots vides ». À cet effet, on téléchargera une liste de mots vides du français accessible à l'URL https://nuage.lip6.fr/s/dGGNfYAGyXao7tQ

6. **Recommencer l'utilisation du Topic Modeling et word cloud**

   Recommencez le « Topic Modeling » et le « Word Cloud » sur quelque textes après le préprocessing.

7. **Transformer un texte en un « sac de mots »**

   Avec le widget « bag of words », transformer les différents textes des corpus en sacs de mots. Regardez le résultat sur la Data Table