MAKING-OF (la suite)

En terme d'analyse de texte nous avons réussi, à l'aide de quatre différents scripts, à réaliser un classement dans dix fichiers « .csv » différents :

1) Un fichier **media.csv** regroupant tous les tweets « nettoyés » des impuretés (mots de liaison, Emoji, adresse url, @, etc.)

2) Un fichier **mediaMotFrance.csv** qui contient les Uni-grams (mots seuls) les plus utilisés lorsqu'on tweet aux différents médias français sélectionnés. Ils sont classés en ordre décroissant.

3) Un fichier **mediaMotQuebec.csv** qui contient les Uni-grams (mots seuls) les plus utilisés lorsqu'on tweet aux différents médias québécois sélectionnés. Ils sont classés en ordre décroissant.

4) Un fichier **journalisteMotFrance.csv** qui contient les Uni-grams (donc mots seuls) les plus utilisés lorsqu'on tweet aux différents journalistes français sélectionnés. Ils sont classés en ordre décroissant.

5) Un fichier **journalisteMotQuebec.csv** qui contient les Uni-grams (donc mots seuls) les plus utilisés lorsqu'on tweet aux différents journalistes québécois sélectionnés. Ils sont classés en ordre décroissant.

6) Un fichier **hashtagsMot.csv** contenant tous les #Hashtags sélectionnés (#journalismedemerde, #mainstream, #journaux, #presse, #lapresse, #journaliste, #presselibre, #journalisme, #fakenews, #journalope, #merdias) dans lequel nous avons sélectionnés les trois mots qui précèdent puis qui suivent le hastags dans chaque tweet.

7) Un fichier **twitter_unigrams.csv** qui contient les Uni-grams (donc tous les mots) pour pouvoir déterminer lesquels reviennent le plus souvent. La différence avec l'autre fichier Uni-gram c'est qu'ils correspondent à tous les mots récoltés dans l'ensemble des données (donc pas séparés en fonction de pays, médias, et journalistes)

8) Un fichier **twitter_bigrams.csv** qui contient les Bi-grams (donc les expressions de deux mots) pour pouvoir déterminer lesquels reviennent le plus souvent dans les tweets.

9) Un fichier **twitter_trigrams.csv** qui contient les Tri-grams (donc les expressions de trois mots) pour pouvoir déterminer lesquels reviennent le plus souvent dans les tweets.

10) Un fichier **twitter_hashtags.csv** qui contient tous les #hashtags utilisés dans l'ensemble des données.

On a utilisé la fonction « value_count » dans Jupyter pour essayer de dégager des informations et répondre à nos questions de départ. Néanmoins, après analyse, on se rend compte que les données ne permettent pas d'y répondre ni de tirer de conclusions probantes. Nous avons donc décidé de faire des « Top 20 » dans chaque catégorie pour illustrer cela. Les résultats sont sous forme de tableaux dans le reportage final.

***Les scripts et l'analyse des données sont le travail de Jean-Batiste et moi-même.
***L'écriture du reportage est le travail de Théo et Mélissa.