Description Sujet 1 :

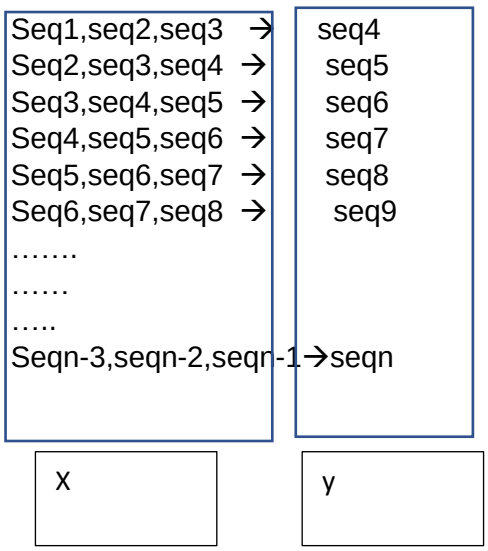
Pour pouvoir réaliser la prédiction du sujet d'intérêt d'un utilisateur web, il serait plus facile de faire les tâches suivantes :

1- Réduire chaque tweet ou commentaire à des mots clés pouvant le résumer « My heart goes out to the Malaysian people. This is such a tragedy. Words can't express how sad it is. I wish we could just have peace. #MH17 »-→Les mots clés : tragedy, Malaysian people, sad (Vous pouvez appliquer TF-IDF et choisir ceux qui présentent un poids important, 2-gram)

Tweet1	Mot_11	Mot_12	Mot_13
..	..	..	..
..	..	..	..
..	..	..	..
..	..	..	..
..	..	..	..
..	..	..	..
Tweet n	Mot_n1	Mot_n2	Mot_n3

2- Fixer un nombre précis pour les mots clés pour chaque tweet (exemple 3). Vous pouvez utiliser (from keras.preprocessing.sequence import pad_sequences// from keras.preprocessing.text import Tokenizer) pour pouvoir ajuster chaque séquence de mots.

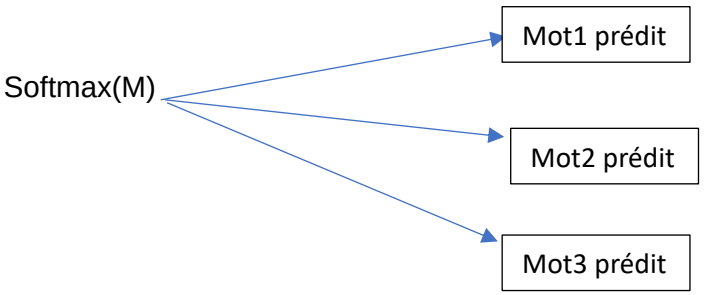
3- Transformer chaque séquence de mots en une représentation numérique : vous pouvez utiliser « Embedding layer » de keras. Vous pouvez choisir 3 séquences successives pour prédire une séquence



4- Le réseau utilisé sera un réseau pour apprentissage supervisé : en effet, la prédiction sera à chaque fois des trois mots successifs du prochain tweet. Le réseau de neurones aura comme sortie pour chaque frame une matrice M de taille(n,p) tel que :

- n est le nombre de mots pour chaque séquence prédite
- p est le nombre de mots dans le vocabulaire+1 ; l'ajout du 1 est pour le mot « unknown » dans le cas où le mot retrouvé est inconnu.

5- Le réseau cherchera à labelliser chaque vecteur ligne par le mot le plus proche en utilisant la fonction softmax.



-