Data for Machine Learning

unix

http://www.shellunix.com/index.html

Récupérez sur le site *ftp* à l'aide d'une ligne de commande le fichier suivant *splice .tar.gz* que l'on peut trouver à l'URL suivante : ftp.cs.toronto.edu dans le répertoire *pub/neuron/delve.* Bien sûr vous pouvez le faire en mode graphique, mais faisons un peu de *'commandes type unix'* pour cela.

```
   #pour se connecter au serveur
   $ftp ftp.cs.toronto.edu #pour file transfer protocol #ou sftp
   # en général ces serveurs utilisent le USER LOGIN : anonymous et le
#PASSWORD : votre_email pour des raisons de sécurité
   >cd pub/neuron/delve
   >ls
   >cd data
   >ls
   >cd tarfiles
   >ls
   >get splice.tar.gz
   #si pb passer en mode passiv pour que la commande get fonctionne.
   >quit
   $ls
```

```
   #pour decompresser un fichier
   $gunzip splice.tar.gz

   #pour dérouler l'arborescence
   $tar -xvf splice.tar

   #pour se déplacer
   $cd splice/Source
   $gunzip splice.data.gz

   #pour lire ce que contient le fichier
   $more splice.data
   $more splice.names

   # Que contient ce fichier ? Comment est-il organisé ?
   #usage du pipeline avec la commande grep ('$man grep' pour comprendre)
   $more splice.names | grep EI
   $more splice.names | grep [EI,IE]
   $more splice.data

   #que fait la commande cut ? $man cut par exemple pour répondre
   $cut -c 1 splice.data
   $cut -f 2-3 -d, splice.data
```

Décrire ce fichier et exprimez le problème de Machine Learning sous-jacent en termes d'entrées-sorties avec l'explication biologique qui va avec.

gawk

http://www.shellunix.com/awk.html

Récupérer le tableau de mesures *2D.txt* décrivant un nuage de points 2D. Le nuage n'est pas bien formaté. Il faut le formater différemment. Par exemple, certaines lignes comportent une troisième coordonnée égale à 0. Il faut l'enlever. On va utiliser la commande ***gawk*** ou ***awk*** d'Unix :

- Créez un programme ***reformate.awk*** à l'aide de ***gedit*** ou ***gvim*** ou ***emacs*** qui contient les deux lignes de code suivante :

```
{print $1,$2}
END{print NR}
```

- Appliquez-le au fichier *2D.txt* en lançant :

```
$awk -f reformate.awk 2D.txt
```

- puis

```
$awk  '{print $1,$2}' 2D.txt >2Df.txt
```

Comment utiliser cet utilitaire pour formater le fichier splice.data afin d'alimenter vos algorithmes d'apprentissages sous sklearn python. Eventuellement créer un fichier au format .csv.