TP Spark

Yves Denneulin, Vincent Leroy

2021

1 Présentation

Le but de ce TP est de vous familiariser avec Spark en utilisant le langage fonctionnel Scala. Vous aurez à manipuler des tweets en utilisant Spark.

Vous avez dans cette archive 3 fichiers : un contenant des tweets[1], un autre une bibliothèque de fonctions pour faciliter le traitement de ces tweets et enfin un fichier permettant d'associer des sentiments à des mots.

Pour faire ce TP, vous utiliserez l'interprète de commandes de Spark que vous lancez par la commande `spark-shell` [2]

1. charger le fichier de tweets fourni dans un RDD en utilisation la fonction `sc.textFile`
2. créer un RDD contenant tous les tweets mentionnant Donald Trump. Affichez le nombre d'élements de ce RDD.
3. en utilisant la librairire `TweeetUtilities`[3] construisez un RDD qui contient des couples (tweet, sentiments). Affichez les 5 premiers éléments de ce RDD en utilisant `take(5).foreach(println)`.
4. en utilisant ce RDD, créer un RDD qui représente le sentiment associé à chaque Hashtag. Affichez les 5 premiers éléments de ce RDD en utilisant `take(5).foreach(println)`.
5. Vous pouvez maintenant afficher le top K des hashtags postifs et négatifs en utilisant `top(k)` et `takeOrdered(k)`.

[1] c'est un fichier brut, qui peut contenir des éléments incorrects.

[2] sur les machines de l'Ensimag, pour utiliser la bonne version de Java, il faut taper `JAVA_HOME=/usr/lib/jvm/java-11-openjdk-amd64 spark-shell`

[3] Pour la charger dans l'interpréteur, tapez la commande :load TweetUtilities.scala