TP3: Manipulation des RDD

Exercice 1:

- Ecrire une application Spark en Python pour avoir les **3 mots** qui se répètent le plus dans un fichier texte.

  Indication: Utiliser les transformations **flatmap**, **map**, **reduceByKey**, **sortBy** et l'action **take**

- Lancer votre application via le **shell pyspark** sur le cluster déjà créé
- Accéder à la page du monitoring du cluster pour voir les ressources allouées
- Accéder à la page du monitoring de l'application **pyspark**
- Déployer votre application sur le cluster avec **spark-submit**

TP3: Manipulation des RDD

Exercice 2:

- Soit le fichier **foot1.csv** de match de foot contenant les colonnes suivantes séparées par ";":

date;home_team;away_team;home_score;away_score;tournament;city;country

- Ecrire les programmes Python permettant d'avoir:

1. Le nombre de matchs par tournoi
2. Le nombre de match joués dans chaque ville.
3. Le nombre maximal de buts marqués dans un seul match par pays.
4. Le nombre de matchs gagnés par pays (équipe)
5. Le nombre de buts marqués et celui des buts encaissés par un pays saisie au claver.

   N.B: Utiliser deux accumulateurs pour calculer ces deux nombres.
6. Les noms des pays ayant marqué le nombre maximal de buts.