

Taux d'abstention aux élections municipales

Le but de ce projet informatique est de traiter des données à travers le paradigme de programmation *MapReduce*, la structure de fichiers *HDFS* et à l'aide du package `rnr2` du logiciel R.

1 Récupération des données

Les fichiers disponibles aux adresses suivantes :

```
https://www.data.gouv.fr/fr/datasets/elections-municipales-2008-resultats-572150/  
https://www.insee.fr/fr/statistiques/2044683  
https://www.insee.fr/fr/statistiques/2044748  
https://www.insee.fr/fr/statistiques/2044618  
https://www.insee.fr/fr/statistiques/2128672  
https://www.insee.fr/fr/statistiques/2044713  
https://www.insee.fr/fr/statistiques/2044707
```

contiennent d'une part des données sur les résultats des élections municipales en France en 2008, et d'autre part des indicateurs socio-économiques et démographiques sur les communes françaises. Le but de ce projet informatique est de s'intéresser au taux d'abstention aux élections municipales en France et de le relier éventuellement à des données socio-économiques et démographiques.

2 Traitement des données

Ce projet informatique est très libre, en particulier toutes les initiatives personnelles et les approches innovantes seront fortement appréciées. Néanmoins, l'un des buts étant de vous familiariser avec la problématique de la programmation *MapReduce*, il vous sera demandé d'implémenter chaque question en utilisant le modèle de programmation *MapReduce* à travers le package `rnr2`. Afin de vérifier la cohérence de vos résultats, vous pouvez éventuellement les comparer avec un traitement standard des données en R.

Attention : pour le traitement des données en *MapReduce*, on utilisera directement les données converties au format *HDFS* à l'aide de la commande `to.dfs`.

Il n'est pas autorisé d'utiliser de pré-traitement des données en R avant de les convertir en *HDFS* !

Voici quelques questions qui pourront être traitées avec *RHadoop* :

- Sur l'ensemble de la France, calculer des statistiques descriptives de base du type moyenne du taux d'abstention, variance, minimum, maximum. Faire le même type d'analyse par département. Quel est le département où le taux d'abstention moyen est le plus fort / plus faible ?
- Quelle sont les variables socio-économiques et démographiques les plus corrélées avec le taux d'abstention ?
- Proposer un modèle de prévision du taux d'abstention en fonction de variables socio-économiques et démographiques.

Nous insistons sur le fait que les questions précédentes ne sont que des suggestions et n'ont pas vocation à être exhaustives. Toute prise d'initiative sera appréciée.

3 Travail à effectuer

Il est demandé de nous envoyer un compte-rendu sous la forme d'un fichier Rmarkdown (et le .pdf associé) avec vos codes *R* correctement commentés.