Résumé sur le package dplyr

Jingwen SU

12/20/2020

Introduction

À la fin du cours, tous les étudiants ont téléchargé les documents pertinents sur Github et partagé les résultats avec nous. Afin d'utiliser ces packages d'installation de R de manière plus complète, je vais lire, analyser et évaluer les articles de mes collègues pour améliorer mes lacunes.

Ceci est une introduction à l'article lu dans cet article. Si vous souhaitez en savoir plus, vous pouvez rechercher des références.

- **Title:** DPLYR
- **Auteurs:** YIN Xueting
- **Lien sur Github:** YIN Xueting

Synthèse du travail en question

L'auteur a d'abord introduit le concept de dplyr et les données qui seront utilisées plus tard. Puis combiné avec la base de données pour le traitement des données, y compris l'application de plusieurs fonctions, telles que: 'filter', 'slice', 'distinct', 'sample', 'arrange', 'join', 'mutate', 'group by', 'rename'.

Contenu principal et explication

Les fonctions données par l'auteur sont très couramment utilisées et font partie des nombreuses fonctions que nous utilisons souvent lors du traitement des données. Grâce aux remarques de l'auteur, nous pouvons voir intuitivement les usages multiples de chaque fonction. Ce sont toutes des choses que nous devons maîtriser.

- **Select**: sélectionner des colonnes d'un tableau de données.

```
select(iris, 1:2) #selectionner la 1ère et la 2ème colonne
select(iris,c(1,2))
select(iris,c("Sepal.Length","Sepal.Width"))
#les résultats sont idem les codes ci-dessus

select(iris, -Species) # ne pas montrer la colonne Species
grep("Sepal", colnames(iris)) # vérifier si "Sepal" existe dans les colonnes de iris et retourner à sa p
select(iris, grep("Sepal", colnames(iris))) # selectionner les colonnes qui contiennent le mot "Sepal"
```

- **filter**: sélectionne des lignes d'une table selon une condition.

```r
filter(iris, Sepal.Length >5) #selectionner les lignes où Sepal.Length est supérieur à 5.
```

- **slice**: Sélectionne des lignes du tableau selon leur position.

```r
slice(iris, 1:5) # selectionner de la ligne 1 à la ligne 5
slice(iris, 2) # selectionner la ligne 2
```

- **distinct**: Filtre les lignes du tableau pour ne conserver que les lignes distinctes, en supprimant toutes les lignes en double.

```r
distinct(iris) # supprimer les lignes redondantes
```

- **sample**: Tirage.

```r
sample_frac(iris, 0.1) # tirer 10% de lignes
sample_n(iris, 10) # tirer 10 lignes
```

- **arrange**: Réordonne les lignes d'un tableau selon une ou plusieurs colonnes.

```r
arrange(iris, Sepal.Length) # selon sepal length, trier en ordre croissant
arrange(iris, desc(Sepal.Length)) #selon sepal length, trier en ordre décroissant
```

- **join**: Fusionner par le même nom de colonne.

```r
a <- 1:10
b <- 20:7
df1 = cbind(a,a^2)
df2 = cbind(b,b^3)

colnames(df1) = c("x1","x2") # renommer les colonnes de df1
df1 = tibble::as_tibble(df1)
colnames(df2) = c("x1","x3") # renommer les colonnes de df2
df2 = tibble::as_tibble(df2)

left_join(df1,df2) # Correspond à toutes les valeurs de df1, les valeurs qui n'existent pas dans df2 af
#nous pouvons également utiliser by = pour correspondre quel ligne
right_join(df1,df2) # Correspond à toutes les valeurs de df2, les valeurs qui n'existent pas dans df1 a
inner_join(df1,df2) # Correspond à toutes les valeurs communes de df1 et df2
full_join(df1,df2) # Correspond à toutes les valeurs de df1 et df2
```

- **mutate**: Mutate permet de créer de nouvelles colonnes dans le tableau de données, en général à partir de variables existantes.

```r
mutate(iris, Petal.Size = Petal.Width + Petal.Length)
```

- **group by**: Permet de définir des groupes de lignes à partir des valeurs d'une ou plusieurs colonnes.

```r
summarise(group_by(iris, Species), mean(Sepal.Length)) # Regrouper par Species et trouver la moyenne
```

- **rename**: Permet de renommer des colonnes.

```r
rename(iris, S.W = Sepal.Width)
```

Ces fonctions sont relativement basiques et faciles à comprendre. Vous pouvez les maîtriser en quelques séances d'entraînement seulement. La pratique vous permet de maîtriser plus facilement les fonctions que la lecture.

Evaluation et résumer

Selon mes normes, je considère cet article comme un article au-dessus de la moyenne.

L'ensemble de l'article a un bon format, explique les concepts associés et donne des démonstrations des fonctions associées. Les détails sont très détaillés et chaque étape est expliquée en détail.

Mais l'inconvénient est qu'il n'y a pas des annexes pertinentes sont fournies, et la fonction est relativement simple, et pas a été effectué des recherches approfondies sur.