# Data Science Avancées

2023-03-28

## 1 Importation du jeu de données

```
bbc <- read.table(
  file="../datasets/bbc.csv",
  header=TRUE,
  sep=",",
  fileEncoding= "latin1")
```

`header=TRUE` : précise que le nom des variables est présent

`sep=";"` : précise que le séparateur de colonnes est le point-virgule (fréquent dans les fichiers csv, pour une tabulation il faudrait écrire `sep="\t"`)

`dec="."` : le séparateur de décimale est le point (parfois dans Excel on trouve la virgule)

`row.names=1` : précise que le nom des individus est dans la premiére colonne du tableau

`check.names=FALSE` : impose que le nom des colonnes soit pris tel que dans le fichier (sinon les espaces sont remplacés par des points et des X sont mis avant les nombres)

Il est important de s'assurer que l'importation a bien été effectuée, et notamment que les variables quantitatives sont bien considérées comme quantitatives et les variables qualitatives bien considérées comme qualitatives.

Voiçi un apperçu du jeu de donnée.

```
# on considère indexation de base
df.bbc = dplyr::tibble(bbc[-1])
head(df.bbc)
```

```
## # A tibble: 6 x 2
##   text                                                                label
##   <chr>                                                               <chr>
## 1 "England coach faces rap after row\n\nEngland coach Andy Robinson is fa~ sport
## 2 "Moody joins up with England\n\nLewis Moody has flown to Dublin to join~ sport
## 3 "Ferguson fears Milan cutting edge\n\nManchester United manager Sir Ale~ sport
## 4 "Henry tipped for Fifa award\n\nFifa president Sepp Blatter hopes Arsen~ sport
## 5 "Arnesen denies rift with Santini\n\nTottenham sporting director Frank ~ sport
## 6 "Charvis set to lose fitness bid\n\nFlanker Colin Charvis is unlikely t~ sport
```