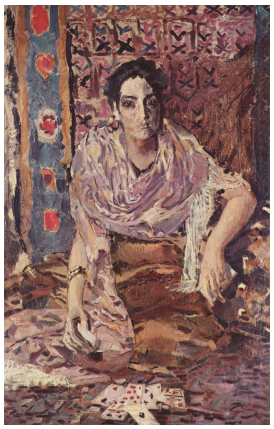
Soutien informatique : UNIX. Fouille de données. TD 2

Sergey Kirgizov



Diseuse de bonne aventure (1895)
Mikhaïl Aleksandrovitch Vroubel

1 Retour chariot et saut de ligne

De nombreux caractères différents sont utilisés pour marquer la fin des lignes. Les deux caractères les plus couramment utilisés sont décrits dans le tableau suivant :

Nom français	**Nom anglais**	Code ASCII	Abréviation	Échappement antislash
Retour chariot	Carriage Return	13	CR	\r
Saut de ligne	Line feed	10	LF	\n

Les systèmes de la famille Unix utilisent le caractère \n pour marquer le début d'une nouvelle ligne. D'autres systèmes peuvent utiliser d'autres caractères ou même des séquences de caractères. (Voir `https://en.wikipedia.org/wiki/Newline` et `https://fr.wikipedia.org/wiki/Fin_de_ligne` pour plus de détails).

☞ **EXERCICE 1.1.** Lire `'man 1 echo'` .

☞ **EXERCICE 1.2.** Comparer et comprendre les résultats d'exécution des lignes suivantes :

```
echo -e 'Aa\nB'
echo -e 'Aa\rB'
```

2 Fouille de données

☞ **EXERCICE 2.1.** Avec `curl` (ou `wget`) télécharger le fichier .zip qui contient les prénoms attribués aux enfants nés en France hors Mayotte entre 1900 et 2021 :
`wget https://www.insee.fr/fr/statistiques/fichier/2540004/nat2021_csv.zip`

☞ **EXERCICE 2.2.** Dézipper le fichier .zip avec la commande `'unzip'`. Vous devriez obtenir un fichier `'nat2021.csv'` après la décompression

☞ **EXERCICE 2.3.** Afficher les lignes liées aux naissances d'enfants nommés ZIDANE, ou ayant un autre prénoms de votre choix.

☞ **EXERCICE 2.4.** Le fichier `'nat2021.csv'` contient les caractères '\r' qui peuvent nous empêcher de travailler avec le fichier. Utiliser `'tr -d "\r"'` pour supprimer les caractères '\r' du fichier de prénoms.

Soit `'numbers.txt'` un fichier qui contient des nombres sur chaque ligne, par exemple

```
0
1
2
3
5
8
13
21
```

La commande `'cat numbers.txt | paste -s'` présentera le fichier sous forme d'une chaîne en utilisant les symboles de tabulation pour séparer les lignes du fichier d'origine :

```
0 1 2 3 5 8 13 21
```

La commande `'cat numbers.txt | paste -s -d+'` utilisera le symbole '+' comme séparateur les lignes du fichier d'origine, afin d'afficher une expression arithmétique suivante :

```
0+1+2+3+5+8+13+21
```

Et finalement la commande `'cat numbers.txt | paste -sd+ | bc'` calculera la somme des nombres contenus dans le fichier `'numbers.txt'`.

```
53
```

☞ **EXERCICE 2.5.** Compter le nombre d'enfants nommés ZIDANE (ou ayant un autre prénoms de votre choix)

☞ **EXERCICE 2.6.** Que fait les trois commandes suivantes ?

```
cat nat2021.csv |  tr -d "\r" > nat.csv
cut -f 2 -d ";" nat.csv | uniq > prenoms
cat prenoms | tr '\n' '\0' | xargs -0 -I {} bash -c 'echo -n "{} "; grep ";{};" nat.csv | cut -f 4 -d ";" | paste -sd+ | bc'
```

☞ **EXERCICE 2.7.** Trouver 10 prénoms les plus rares

☞ **EXERCICE 2.8.** Trouver quelque chose d'intéressant sur la plateforme ouverte des données publiques françaises `https://www.data.gouv.fr/`, par exemple le fichier contenant de l'information sur émissions de CO_2 et de polluants des véhicules : `https://www.data.gouv.fr/fr/datasets/emissions-de-co2-et-de-polluants-des-vehicules-commercialises-en-france/` ou bien d'autres données publiques qui vous intéressent. Utiliser la puissance d'Unix pour étudier ces données.