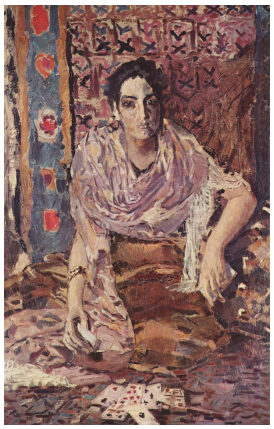
Systèmes UNIX. TD 3 : la fouille de données



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

La plateforme ouverte des données publiques françaises contient beaucoup d'informations intéressantes, par exemple la liste de prénoms attribués aux enfants nés en France depuis 1900 : https://www.data.gouv.fr/fr/datasets/fichier-des-prenoms-depuis-1900/.

EXERCICE 2.1. Avec `wget` télécharger le fichier .zip qui contient les prénoms attribués aux enfants nés en France hors Mayotte entre 1900 et 2020.

```
wget https://www.insee.fr/fr/statistiques/fichier/7633685/nat2022_csv.zip
```

EXERCICE 2.2. Dézipper le fichier .zip avec la commande 'unzip'. Vous devriez obtenir un fichier 'nat2022.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.

Le fichier 'nat2022.csv' contient les caractères '\r' qui peuvent nous empêcher de travailler avec le fichier. On peut voir ces caractères invisibles avec la commande suivante :

```
od -c nat2022.csv
```

EXERCICE 2.4. Supprimer les caractères '\r' du fichier de prénoms :

```
cat nat2022.csv | tr -d '\r' > nat2022-clean.csv
mv nat2022-clean.csv nat2022.csv
```

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 commandes suivantes ?

```
cut -f 2 -d ";" nat2022.csv | uniq > prenoms
cat prenoms | tr '\n' '\0' | xargs -0 -I {} bash -c 'echo -n "{} "; grep ";{};" nat2022.csv | cut -f 4 -d ";" | paste -sd+ | bc'
```

EXERCICE 2.7. Trouver 10 prénoms les plus rares.

Les exercices de complexité élevée sont marquées d'un astérisque★. Ils ne sont pas obligatoires. Ils sont conçus spécialement pour les étudiant(e)s motivé(e)s qui peuvent facilement résoudre les exercices sans astérisque.

EXERCICE★ 2.8. Trouver quelque chose d'intéressant sur la plateforme ouverte des données publiques françaises https://www.data.gouv.fr/. Utiliser la puissance d'Unix pour étudier ces données.