Linux : Découverte du terminal et des entrées/sorties.

Rédacteurs

Gérald Hermant - organic-ip.fr
Raphael Loyet - Campus
Jeremie SUZAN - floki io

Objectifs

- Découverte et utilisation de linux.
- Installation de paquets
- Gestion de fichiers en ligne de commande

Prérequis

- Linux installé.
- Des bases de programmation (boucles, condition,...)

Compétences

- Installer des paquets et comprendre les notions de dépendances
- Créer renommer, éditer des répertoires et des fichiers en utilisant la ligne de commande
- Redirection des entrées et sorties
- Gestion automatisée de fichiers

Démarche Pédagogique

Le module est découpé en trois étapes:

- Une première étape permet tout d'abord de découvrir la ligne de commande sous linux et l'administration d'ubuntu (installation de logiciels).[Phase1]

- Une seconde étape permet de découvrir des commandes un peu plus avancées pour automatiser des traitements basiques sur des fichiers. [Phase1]

- Une troisième et dernière étape regroupe les commandes et notions vues dans les deux premières étapes afin de trier automatiquement 10 000 fichiers en fonction de leurs noms. [Phase2]

Étape 1

Partie 1 : Découverte de la ligne de commande et installation de paquets

Modalités

- Travail en autonomie.
- Production individuelle.
- Temps estimé: 0,4 jours.

Objectif

- Découvrir le déplacement dans l'arborescence en ligne de commande.
- Comprendre ce qu'est une dépendance.
- Installer des logiciels via l'interface graphique.

Consignes

Les bases de la ligne de commande

Faites une partie (15 minutes au moins) avec Terminus :
http://luffah.xyz/bidules/Terminus/.

Installer un paquet sous linux

Pour comprendre ce qu'est un paquet sur linux:
https://fr.wikipedia.org/wiki/Paquet_(logiciel)

A partir du gestionnaire de packages, choisissez les packages suivants :

- feedgnuplot
- tldr (une application d'aide alternative à la commande man)

Un peu de pratique :

Pour préparer la deuxième partie :

- Ouvrez un terminal et aller dans le dossier Documents
- À l'aide de la commande wget, téléchargez une des archives indiquées ci-dessous.
- Créez un nouveau dossier.
- Copiez l'archive dans ce nouveau dossier.
- Décompressez l'archive dans un dossier nommé text_to_be_processed.
  (l'extension tgz correspond à tar.gz et se décompresse en utilisant la commande tar, l'extension .gz se décompresse avec gzip)

- Affichez les fichiers présents dans ce nouveau dossier.
- Depuis la ligne de commande vérifiez que les étapes précédentes se sont bien passées: le répertoire contient un fichier texte d'environ 20Mo.
- Utilisez la commande head pour afficher les 30 premières lignes du fichier et tail pour afficher les 42 dernières lignes du fichier.

Liens vers les archives:

- Texte français
  https://object.pouta.csc.fi/OPUS-UN/v20090831/mono/fr.txt.gz
- Texte anglais
  https://object.pouta.csc.fi/OPUS-UN/v20090831/mono/en.txt.gz

Livrables

❏ Un mémo (càd une fiche résumée) des commandes définies, avec 3 colonnes :
- ❏ A gauche, l'instruction.
- ❏ Au milieu, l'utilisation (ce que ça fait).
- ❏ A droite, un/plusieurs exemples avec des commentaires si besoin (une ligne par option spécifique par exemple).

Instruction	Utilisation	Exemples
cat	affiche le contenu d'un (ou plusieurs) fichier(s)	cat mon_fichier cat -s fichier1 fichier2
cmd_fictive	commande fictive	cmd_fictive -h

❏ Définir la notion de dépendances.

Preuve de travail

❏ Les logiciels demandés sont disponibles sur votre poste.
❏ Le mémo des commandes : cd, ls, cat, pwd, cp, echo, mv, rm, mkdir, diff, head.
❏ La commande history affiche les commandes utilisées pour effectuer les actions demandées.
❏ Le fichier présent dans le dossier est un fichier texte d'environ 20Mo.

Ressources

❏ La commande man <nom_du_programme> affiche le manuel du programme.
- https://doc.ubuntu-fr.org/tutoriel/comment_installer_un_paquet.
- https://openclassrooms.com/fr/courses/7274161-administrez-un-systeme-linux/7529321-adoptez-l-arborescence-des-systemes-linux
- https://linux.goffinet.org/administration/arborescence-de-fichiers/operations-sur-les-fichiers/.

- Vidéo arborescence des fichiers: https://www.youtube.com/watch?v=CSD9USjCX7k
- https://doc.ubuntu-fr.org/tar#tarextraction_de_fichiers.

Pour aller plus loin

- Installez un fichier .deb (par exemple : vscode).
- Lire le PATH sous linux
- Autres gestionnaires de paquet pip, npm, bundle, gem, … installez sl et cowsay, que font ces commandes?

Étape 2 : Les entrées/sorties

Objectifs

L'objectif de ce mini-projet est de créer un index à partir de fichiers textes directement en console linux (sans développer de programme). Ce type d'index est utilisé pour la recherche d'information et/ou la cryptographie. À la fin de ce module vous serez capable de :

- Chaîner les entrée/sortie de commande linux.

Modalités

❏ Temps estimé: 0.6 jours
L'étape deux comprends deux parties

Partie 1 : Compter des éléments dans un fichier

Modalités

❏ Travail en groupe de 2
❏ Production individuelle

Consignes

- Utilisez les redirections d'entrée/sortie linux et les commandes cat et wc afin de créer un fichier param-<langue_choisie> contenant :
  - le nombre de lignes du fichier txt.
  - le nombre de mots du fichier txt.
  - le nombre d'octets du fichier txt.

Livrables

❏ Une démo commentée afin de recréer le fichier param.
❏ Un fichier texte contenant la commande utilisée.

Preuve de travail

❏ Je suis capable d'expliquer la construction de la commande demandée.
❏ La commande cat <nom_du_fichier> affiche les informations demandées.

Ressources

- Texte français https://object.pouta.csc.fi/OPUS-UN/v20090831/mono/fr.txt.gz
- Texte anglais https://object.pouta.csc.fi/OPUS-UN/v20090831/mono/en.txt.gz
- la commande man <commande> permettant l'accès au manuel de la commande

linux

- https://fr.wikibooks.org/wiki/Le_syst%C3%A8me_d%27exploitation_GNU-Linux/Redirection_des_entr%C3%A9es/sorties.
- https://www.tecmint.com/linux-io-input-output-redirection-operators/.

Partie 2 : Compter les occurrences des mots du fichier

Modalités

- Production individuelle
- Travail en groupe de 2

Objectifs

- Nettoyer le fichier texte.
- Compter les occurrences des différents mots des fichiers.

Consignes

Prétraitement

Pour commencer, la ponctuation et les chiffres seront enlevés. La casse de tout le fichier sera changée.

Pour cela la commande `sed` sera utilisée, celle-ci permet d'appliquer différents filtres et différentes transformations sur un flux de texte. Elle sera utilisée avec les arguments suivants afin d'effectuer les prétraitements : `-e "s/[[:punct:]]//g" -e 's/./\L&/g' -e 's/[0-9]*//g'`.

La commande `grep` permet de rechercher du contenu dans un(des) fichier(s), il sera possible de l'utiliser pour filtrer les lignes vides en utilisant les arguments `-v ^$`.

- Utilisez les commandes `sed`, `cat` et `grep` afin de créer un nouveau fichier `preproc_<langue_choisie>.txt`, qui ne contient pas de majuscules, de chiffre, de ponctuation et de lignes vides.

Comptage d'occurrences

- En utilisant les redirections E/S et les commandes `sort`, `uniq`, `head`, `tail` en plus des commandes précédentes créez:
  - un fichier `index_<langue_choisie>` contenant tous les mots triés par nombre d'occurrence croissant.
  - un fichier `top_30_<langue_choisie>` contenant le 30 mots les plus utilisés dans le texte initial.
  - un fichier `last_30_<langue_choisie>` contenant le 30 mots les moins utilisés dans le texte initial.

Livrables

- ❏ Une démo commentée de création de l'un des trois fichiers.
- ❏ Un fichier texte contenant les commandes pour les différentes étapes.

Preuve de travail

- ❏ Les fichiers contenant l'index global des mots et ceux contenant les mots les plus et les moins présents.

Ressources

- la commande `sed -e "s/\ /\n/g"` permet de remplacer les espaces par des sauts de lignes dans un fichier.
- les manuels de `head`, `tail`, `uniq` et `sort`.

Pour aller plus loin

- Utilisez la commande `sed` pour créer un fichier `occurences.csv` avec comme séparateur une virgule ou un point-virgule
- Créez un histogramme avec `feedgnuplots` comprenant les 10 mots les plus utilisés des deux langues.
- Automatisez les différents traitements dans un script pour les appliquer sur l'autre langue.
- Utilisez la gestion de processus pour paralléliser les traitements https://www.tecmint.com/run-linux-command-process-in-background-detach-process/
- Utilisez la commande `awk` pour remplacer le nombre d'occurrences par la fréquence du mot.
- Utilisez la commande `xargs` ou `parallel` afin de paralléliser les traitements.
- https://adamdrake.com/command-line-tools-can-be-235x-faster-than-your-hadoop-cluster.html