Université de Kinshasa

Faculté de Sciences

Département de Mathématiques et Informatique

TP de BI : Partie 1

L2 Informatique : Gestion, 2016-2017.

Assistants :
Eddy Kambilo
Hermann Fundatela
Joachim Kalonga
Jordan F. Masakuna
Joël Ngalamulume

Titulaire :
Prof. Dr. Kafunda Katalay

6 décembre 2016

Table des matières

Résumé

Ce document contient deux types d'informations que nous jugeons relevantes aux étudiants en L2 Informatique. Il comprend :

1. des notions importantes à connaître et utiliser entant que scientifique, finaliste informaticien et futur chercheur.
2. des notions préalables du cours de BI dans son nouveau concept de Big data.

1 Shell : Bash

Un *shell* est simplement un macro-processeur qui exécute des commandes. Le terme macro-processeur désigne une fonctionnalité dans laquelle le texte et les symboles sont étendus pour créer des expressions plus larges.

Un shell est à la fois un interpréteur de commandes et un langage de programmation. En tant qu'interpréteur de commandes, le shell fournit l'interface utilisateur riche munis de beaucoup d'utilitaires GNU. En tant qu'un langage de programmation, il permet de combiner d'utilitaires. Les fichiers contenant des commandes peuvent être créés, et deviennent ensuite des commandes (script).

Les shells peuvent être utilisés de manière interactive ou non interactive. En mode interactif, ils acceptent l'entrée à partir du clavier. Lors de l'exécution non interactive, les shells exécutent des commandes lues à partir d'un fichier.

Bash (*B*ourne-*A*gain *SH*ell) est un shell, l'interpréteur de langage de commande des systèmes d'exploitation GNU.

Bash est le shell par défaut de systèmes d'exploitation GNU.

Utilisation : voir TP.

2 Version Control

Le version control (système de contrôle de versions) est un moyen de conserver une sauvegarde de la modification des fichiers, un historique de ces changements et, surtout, de permettre à une équipe de personnes de modifier simultanément les mêmes fichiers. Il existe de nombreux systèmes de contrôle de version.

Git : Github, BitBucket ou GitLab.

Utilisation : voir TP.

3 Programmation parallèle

Dans la programmation parallèle, plusieurs processus sont exécutés simultanément. Les problèmes complexes peuvent souvent être divisés en problèmes moins complexes, qui peuvent ensuite être résolus en même temps. Il est à noter que tout problème n'est pas parallélisable.

3.1 Concepts de base

Il existe plusieurs modèles de programmation parallèle en usage commun :
- Shared Memory (sans threads)

— Threads
— Distributed Memory / Message Passing
— Data Parallel
— Hybrid
— Single Program Multiple Data (SPMD)
— Multiple Program Multiple Data (MPMD
Tous pour un usage commun.

Utilisation : voir TP.

3.2 Cluster Computing

Le cluster computing consiste à créer plusieurs nœuds (ordinateurs) fonctionnant comme une entité unique. les nœuds du cluster sont interconnectés par le biais de réseaux locaux. Il y a principalement deux raisons de déployer un cluster : la performance du système et tolérance aux pannes.

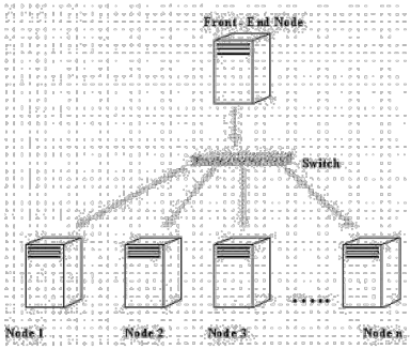

FIGURE 1 – Cluster Computing

Avantages : gestion, image système unique et haute disponibilité.
Désavantages : programmabilité, détection de source d'erreurs.

3.3 Grid Computing

Le Grid Computing est la ségrégation des ressources de plusieurs sites. Il combine l'utilisation de multiples clusters qui sont moins couplés, hétérogènes et géographiquement dispersés.

Avantages : accès aux ressources supplémentaires, équilibrage des ressources et fiabilité.
Désavantages : instable, haute connexion internet requise et différents domaines d'administration.

FIGURE 2 – Grid Computing

3.4 Cloud Computing

Le cloud computing est le nouveau paradigme de l'informatique qui fournit un grand bassin de ressources dynamiques évolutives et virtuelles en tant que service sur demande. Il offre l'informatique, le stockage et le logiciel comme service ou utilitaire. Seule la connexion internet est requise pour l'utilisation du cloud.

FIGURE 3 – Cloud Computing

Avantages : ressources partagées, Pay-As-You-Go et meilleure gestion du matériel.
Désavantages : moins de fiabilité, internet requis et non-interopérabilité.

4 Statistique descriptive

La statistique consiste à collecter les données, organiser (tableaux, diagrammes) et interpréter les résultats.

4.1 Concepts de base

- *Population* : est l'ensemble qu'on observe.
- *Échantillon* : est un sous ensemble de la population considérée.
- *Caractère* : la propriété d'analyse.
- *Caractéristiques de position* : mode, médiane, quantiles, moyenne.
- *Caractéristiques de dispersion* : variance, écart-type

4.2 Types de variables

Une variable est tout facteur qui peut exister des quantités des types différents. Une expérience a généralement deux types de variables : indépendante et dépendante ; et de deux types : variables catégoriques et continues.

Une variable indépendante, parfois appelée variable expérimentale ou prédictive, est une variable qui est manipulée dans une expérience afin d'observer l'effet sur une variable dépendante, parfois appelée variable de résultat.

Les variables catégoriques sont également appelées variables discrètes ou qualitatives, elles peuvent être classées en catégories nominale, ordinaire ou dichotomique. Les variables nominales sont des variables qui ont deux catégories ou plus, mais qui n'ont pas d'ordre intrinsèque. Les variables dichotomiques sont des variables nominales qui n'ont que deux catégories. Les variables ordinales sont des variables qui ont deux ou plusieurs catégories ordonnées.

Les variables continues sont également appelées variables quantitatives. Elles peuvent être catégorisées en variables d'intervalle ou de rapport.

Question. Catégoriser les variables suivantes : Age, sexe, état civil, pays et pourcentage.

4.3 Visualisation des données

La visualisation des données améliore la communication des informations qu'on souhaite partager de manière très claire et compréhensible. La visualisation peut être textuelle, tabulaire ou graphique. Bien qu'il y ait des gens qui l'ont dans les gènes, mais cette habileté (visualisation) peut être apprise ; dans la plupart de cas, dans la pratique.

Dans quel cas, la visualisation graphique peut-elle être plus appropriée que la visualisation textuelle ou tabulaire ?

Question : Choisissez une visualisation textuelle ou tabulaire ou graphique :

- On veut connaître la moyenne des effectifs des étudiants de l'Unikin au cours des dix dernières années. Quel type de visualisation est le plus approprié ? Pourquoi ? Donnez la structure de votre sortie. Et s'il veut distinguer les résultats des différentes facultés ?
- On veut connaître le nombre total d'étudiants de la faculté polytechnique par province à partir de 2014-15. Quel type de résultat est le plus approprié ? Pourquoi ? Donnez sa structure.
- On type veut connaître 5 provinces qui avaient plus d'étudiants de l'Unikin au cours des cinq dernières années. Encore une fois, quelle est la visualisation la plus appropriée ?

Croyez-moi, nous pouvons obtenir deux meilleures réponses mais différentes ... Juste question de savoir comment le faire ... Uhm ... Ai-je raison ?

5 Normalisation des données

Normalisation signifie ajuster les valeurs mesurées sur différentes échelles à une échelle théoriquement commune. C'est à dire quitter l'intervalle $[x_{min}, x_{max}]$ pour l'intervalle $[a, b]$. Une des possibilités est la mise en échelle (feature scaling) qui est donnée par :

$$x_{new} = a + \frac{(x_{old} - x_{min})(b - a)}{x_{max} - x_{min}}, \tag{1}$$

où $x_{old} \in [x_{min}, x_{max}]$ et $x_{new} \in [a, b]$.

6 Features Extraction (Méthodes Factorielles)

En Machine Learning (en reconnaissance de formes ou traitement d'images), l'extraction des caractéristiques (features extraction) consiste à extraire des données (caractéristiques) informatives et non redondantes à partir d'un dataset ayant beaucoup de variables contenant généralement des informations redondantes. Ceci facilite les différentes étapes d'apprentissage et de généralisation conduisant parfois à de meilleures interprétations. L'extraction de caractéristiques est liée à la *réduction de dimensionnalité*, qui est le processus de réduction du nombre de variables considérées (méthodes factorielles), via l'obtention de variables principales.

Quelques techniques : **Complex Moment Invariant, Generic Fourier Descriptor, Exact Legendre Moment et ACP.**

7 Python sur Jupyter Notebook et Spyder

Python est un langage de programmation très utilisé, de haut niveau, généraliste, interprété et dynamique. Sa philosophie de conception met l'accent sur la lisibilité du code et sa syntaxe permet aux programmeurs d'exprimer des concepts dans peu de lignes de code que possible, diffèrent des langages tels que C++ et Java. Il fournit des constructions permettant l'écriture de programmes clairs à la fois sur une petite et grande échelle.

Python prend en charge de multiples paradigmes de programmation, y compris la programmation orientée objet, impérative et procédurale. Il dispose d'un système de type dynamique et la gestion de la mémoire automatique et dispose d'une bibliothèque standard aussi large.

Les interpréteurs Python sont disponibles pour de nombreux systèmes d'exploitation, permettant au code Python de s'exécuter sur une grande variété de systèmes.

Le Jupyter Notebook est un IDE web qui permet de créer et de partager des documents contenant du code interactif, des équations, des visualisations et des textes explicatifs. Spyder est un IDE desktop.

Utilisation : voir TP.

8 MongoDB

MongoDB est une base de données Open Source et NoSQL. Il a été développé en C ++. MongoDB est une base de données multi-plateforme orientée document qui fournit, une haute performance, une haute disponibilité et une évolutivité facile. Il travaille sur le concept de collection et de document.

8.1 Structure de MongoDB

1. **Collection** : est un groupe de documents MongoDB. C'est l'équivalent d'une table SGBDR. Une collection appartient à une seule base de données. Les documents d'une collection peuvent avoir des champs différents. En général, tous les documents d'une collection sont de nature similaire ou connexe.
2. **Document** : est un ensemble de paires clé-valeur. Les documents ont un schéma dynamique. Le schéma dynamique signifie que les documents d'une même collection n'ont pas besoin d'avoir la même structure.

8.2 Différence entre SGBDR et MongoDB

SGBDR	**MongoDB**
Base des données	Base des données
Table	Collection
Enregistrement/ ligne	Document
Colonne	Champs
Jointure	Documents embarqués
Clé Primaire	Clé Primaire

Contrairement aux bases de données relationnelles, MongoDB n'a pas de notion de relation.

1. **Avantages** :
   — Moins de Schéma : MongoDB est une bd orientée document, dans lequel une collection peut contenir plusieurs documents de structures différentes.
   — Clarté de la structure.
   — Pas de jointures complexes.
   — Capacité de requête, dynamisme de requêtes.

— Tuning.
— Facile à mesurer.
— Utilisé pour les bases des données objet.
— Accès rapide des données.

2. **Utilisation de MongoDB** :
— Big Data.
— CMS.
— Télécommunication et Réseaux Sociaux.
3. **Pourquoi MongoDB** :
— Stockage orienté document : JSON.
— Indexation de n'importe quel attribut
— Réplication et disponibilité
— Requêtes riches
— Rapide en mise à jour

Utilisation : voir TP.

9 Latex pour ton mémoire, un bon choix

LaTeX (*La*mport *TeX*) est un système de préparation de documents. L'utilisateur saisit de textes bruts par opposition au texte formaté des logiciels de traitements de texte tel que Microsoft Word.

L'utilisateur se sert des commandes afin de définir la structure de son document (ex. un article, un livre), styliser le texte d'un document (ex. gras ou italique) et ajouter des citations. Une distribution TeX telle que MikTeX, est utilisée pour produire de fichiers de sortie tel que PDF.

LaTeX est largement utilisé pour la publication de documents scientifiques dans plusieurs domaines (ex. mathématiques et informatique).

Utilisation : voir TP.

10 Outils

1. **MikTex et TexMaker** : pour Latex (https://miktex.org/2.9/setup et http://www.xm1math.net/texmaker/download.html#windows).
2. **Anaconda** : pour Python (https://www.continuum.io/why-anaconda).
3. **Git Bash** : pour Bash (https://git-for-windows.github.io/).
4. **Git** : pour Git (déjà installé sur les systèmes Windows).
5. **MongoDB** : pour base des donnees NoSQL (https://www.mongodb.com/download-center#community).

Cours dans le repository : git@github.com:jmf-mas/TPBI-2016-2017.git