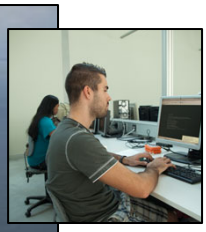
Laboratoire de Technologies de e-commerce (Mobiles & Java)
Partie "Technologies Java et Data mining" : projet " ScienceAirport"- suite: 1ère partie

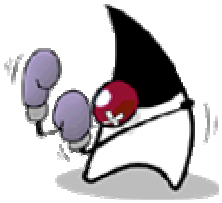
3ème Informatique de gestion
2021-2022





Projet *ScienceAirport* - la suite

Claude Vilvens et Christophe Charlet




1. Préambule

L'Unité d'Enseignement "**Programmation réseaux, web et mobiles**" (10 ECTS - 135h) se structure en Informatique de gestion en trois Activités d'apprentissage de la manière suivante :

- AA: Réseaux et technologies Internet (60h - 45%)
- AA: Programmation.Net (30h - 22%)
- AA: Technologie de l'e-commerce et mobiles (45h - 33%)

Le contexte de ce laboratoire de "Technologie de l'e-commerce et mobiles" est le même que celui du laboratoire de Réseaux et technologies Internet, à savoir celui de "ScienceAirport" qui vise à la gestion d'un aéroport.

2. Règles d'évaluation

Comme on sait, la note finale pour l'UE considérée se calcule par une moyenne des notes des AA constitutives, sachant que le seul cas de réussite automatique d'une UE est une note de 10/20 minimum dans chacune des AAs.

Pour ce qui concerne l'évaluation de l'AA "Technologie de l'e-commerce et mobiles", voici les règles de cotation utilisées par les enseignants de l'équipe responsable de cette AA.

1) L'évaluation établissant la note de l'AA "Technologie de l'E-commerce et mobiles" est réalisée de la manière suivante :

- examen de théorie: un examen écrit en janvier 2022 (sur base d'une liste de points de théorie à développer fournis au fur et à mesure de l'évolution du cours théorique) et coté sur 20;
- laboratoire en évaluation continue: une évaluation ("évaluation 1" ci-dessous) cotée sur 20 qui constitue la note d'évaluation continue;
- examen de laboratoire: un examen oral en janvier 2022 consistant en la présentation de la 2ème partie du laboratoire ("évaluation 2" ci-dessous) et coté sur 20;
- note finale : **moyenne géométrique de la note de l'examen de théorie (poids de 50%), de la note d'évaluation continue (poids de 20%) et de la note de l'examen de laboratoire (poids de 30%)**.

Dans ces conditions, ***il est clair qu'une note beaucoup trop basse parmi les trois ne peut que conduire à l'échec de l'AA considérée***.

Cette procédure est d'application tant en 1ère qu'en 2ème session.

2) *Dans le cas où les travaux sont présentés par une équipe de deux étudiants*, chacun d'entre eux doit être capable d'expliquer et de justifier l'intégralité du travail sans de longues recherches dans le code de l'application proposée (pas seulement les parties du travail sur lesquelles il aurait plus particulièrement travaillé).

3) *Dans tous les cas,* tout étudiant doit être capable d'expliquer de manière générale (donc, sans entrer dans les détails) les notions et concepts théoriques qu'il manipule dans ses travaux (par exemple: keystore SSL, régression multiple et tests, etc).

4) En 2ème session, un **report de note** est possible pour **des notes supérieures ou égales à 10/20** en ce qui concerne :

- la note de théorie;
- les notes de laboratoire des évaluations 1 et 2 (évaluation à l'examen).

Les évaluations de théorie et du laboratoire ayant des **notes inférieures à 10/20** sont donc **à représenter dans leur intégralité** (le refus de représenter une évaluation complète de laboratoire entraîne automatiquement la cote de 0).

Le laboratoire de "Technologie de l'E-commerce et mobiles" comportera donc deux évaluations. La première (data mining avancé et Big data, Android) sera **évaluée** par l'un des professeurs du laboratoire **à partir du 8 novembre 2021** (avec rentrée d'un dossier papier tel que décrit dans l'énoncé). La deuxième (SSL, messagerie électronique, Big data et exploration des données, …) sera évaluée lors de l'examen de laboratoire en **janvier 2022** (un dossier papier ne sera plus nécessaire).

Les travaux de l'évaluation 1: Data mining avancé et Android

Compétences développées :

- Maîtriser les techniques avancées de data mining : REG-CORR multiple, ANOVAs, CHI-CARRE;
- Développer un applicatif Java middleware avec RServe;
- Maîtriser les bases de la programmation Java Android intégrée au développement Java classique.

Dossier attendu :

1. code et explications des résolutions des problèmes 1.2 (REG-CORR multiple) et 2.2 (ANOVA 2) proposés avec R et RServe
2. code du client **Applic_DataMining**;
3. schéma général de l'application Android en termes d'activités, d'intents, d'events (formalisme libre).

1. La dépendance de plusieurs variables

Il s'agit ici de résoudre trois exercices portant sur la dépendance (ou non) de variables quantitatives et/ou qualitatives.

1.1 Les déménagements

Une entreprise de déménagement s'intéresse au nombre d'heures demandées par un déménagement (temps de transport non compris) en fonction du volume estimé de biens à déplacer et du nombre de très grands meubles (lit breton indémontable, bahut mérovingien d'époque, etc). Etudier et tester l'adéquation d'un modèle linéaire (donner la valeur des divers symboles de ce modèle) sur bases des données échantillonnées suivantes (à transformer au préalable en fichier csv) :

temps	volume	nombre de grandes pièces
24.00	545	3
13.50	400	2
26.25	562	2
25.00	540	2

9.00	220	1
20.00	344	3
22.00	569	2
11.25	340	1
50.00	900	6
12.00	285	1
38.75	865	4
40.00	831	4
19.50	344	3
18.00	360	2
28.00	750	3
27.00	650	2
21.00	415	2
15.00	275	2
25.00	557	2
45.00	1028	5
29.00	793	4
21.00	523	3
22.00	564	3
16.50	312	2
37.00	757	3
32.00	600	3
34.00	796	3
25.00	577	3
31.00	500	4
24.00	695	3
40.00	1054	4
27.00	486	3
18.00	442	2
62.50	1249	5
53.75	995	6
79.50	1397	7

En particulier, prédire sur base de ce modèle le temps nécessaire pour un volume de 1500 et 15 grandes pièces.

1.2 Le fromage boursoulavien

Le Ministère de l'Agriculture de Boursoulavie (MAB), et spécialement son ministre Carlos Mondo de Riches, est très attentif à la qualité du fromage national (le FatCheese). Il a donc mandaté une enquête sur le rendement fromager (RFESC) défini comme le poids en kg de fromage obtenu à partir de 100 litres de lait (provenant des vaches de race LimousineHerveExquis et SalersAurillacPiquouse nécessaires à l'AOC). En particulier, il voudrait savoir si ce rendement est influencé par certaines caractéristiques du lait. Les résultats de différentes mesures se trouvent dans le fichier "**RdtFromage.txt**".

Les caractéristiques en question sont mesurées de plusieurs manières:

- des mesures de base, applicables à toute fabrication alimentaire (les 7 premières colonnes) et à la portée des laboratoires du MAB;
- des mesures qui réclament un laboratoire spécialisé;

- des mesures de propriétés physiques et en particulier mécaniques (comme les déformations ou les écoulements).

Dans un premier temps, le MAB ne prendra donc en compte que les mesures qui lui sont directement accessibles pour vérification, donc les mesures de base. Celles-ci sont :

- MAT : matière azotée totale (g/l)
- CNE : concentration en caséine (g/l)
- NPN : azote non protéique (g/l)
- CAT : concentration en calcium total (g/l)
- CAS : concentration en calcium soluble (g/l)
- CAI : concentration en calcium ionique (g/l)
- ES : extrait sec du lait en %, standardisé à 25 mg/l
- RFESC : rendement fromager

Le MAB recherche le modèle mathématique qui serait le mieux adapté pour réaliser des calculs prédictifs de RFESC.

A priori, on commencera par étudier la régression/corrélation multiple de cette variable expliquée en fonction de toutes les autres. Ensuite, on peut éliminer les variables explicatives (les régresseurs) qui sont manifestement peu utiles parce que

- leur test de régression amène à accepter H_0;
- elles sont en corrélation serrée avec RFESC.

La fonction plot() peut être utile pour détecter ces derniers cas.

1.3 Les tâches ménagères

Le Ministère de la Famille d'Exuvie a commandité à des statisticiens britanniques une étude sur la répartition des tâches ménagères entre les deux membres d'un couple traditionnel (le fait qu'ils soient parents n'a pas été pris en compte). Le résultat se trouve dans le fichier taches_menageres.txt. Quelles conclusions peuvent être tirées ?

2. ANOVA

Il s'agit ici à nouveau de résoudre deux exercices portant sur la dépendance (ou non) de variables quantitatives et qualitatives.

2.1 Les civilisations précolombiennes

Le Ministère de la Culture de Batracie a étudié le nombre de récipients contenant de la bière fermentée sur divers sites archéologiques correspondants chacun à l'un des 4 types de civilisations précolombiennes suivantes: Cuacuacomeqiqi, Oxomatl, Tlaloc et Tenochtitlan. Les résultats de ces comptages (par are) sont repris ci-dessous (à transformer au préalable en fichier csv):

Cuacuacomeqiqi	Oxomatl	Tlaloc	Tenochtitlan
93	85	100	96
120	45	75	58
65	80	65	95
105	28	40	90
115	75	73	65
82	70	65	80
99	65	50	85

87	55	30	95
100	50	45	82
90	40	50	
78		45	
95		55	
93			
88			
110			

Observe-ton des différences significatives entre les quatre traitements et quels sont ceux qui sont, si ils existent, à résultats similaires ?

2.2 Les médicaments contre la GCE

Une entreprise pharmaceutique s'intéresse à une maladie tropicale (la Gengivite Cephalopodique Endiablée - GCE) et a mis au point trois molécules susceptibles de soigner cette maladie : AlphaVictoire, BetaTriomphe et GammaSucces. Les tests cliniques ont été pratiqués pour mesurer un coefficient relatif d'amélioration de l'état de patients gravement atteints (plus ce coefficient d'immunité est élevé et plus l'action sera considérée comme efficace). Mais, de plus, on souhaite également tenir compte du mode d'administration des différentes molécules (par voie orale ou par injection intraveineuse). Les résultats (à transformer au préalable en fichier csv) sont :

	AlphaVictoire	**BetaTriomphe**	**GammaSucces**
voie orale	10	7	12
	12	14	9
	8	10	11
	10	11	27
	6	9	7
	13	10	8
	9	11	13
	10	7	14
	9	9	10
	8	9	11
injection	11	8	7
	18	9	6
	12	10	10
	15	9	7
	13	11	7
	8	13	5
	15	7	6
	16	14	7
	9	15	9
	13	12	6

Observe-t-on une différence significative d'efficacité soit selon la molécule, soit selon le mode d'administration ou encore selon une combinaison des deux facteurs ?

3. Data mining avec le serveur RServe

Il s'agit de tester le serveur RServe en réalisant une application Java qui permet de réaliser quelques questions statistiques évoquées ci-dessus :

- les déménagements
- les civilisations précolombiennes

Cette application Java avec interface graphique est cliente de RServe qui fournit les résultats statistiques en retour, l'application Java les rendant présentables et interprétables dans son interface graphique. Pour rappel, il est possible de sauver en R les graphiques générés en utilisant les commandes jpeg(file="...jpg", width=800, height=700), plot(…), dev.off(), dev.new().

2. Data mining et informations statistiques

2.1 Présentation

ScienceAirport a décidé de se doter d'un serveur **Serveur_DataMining** dans le but d'améliorer ses offres et de mieux répondre aux souhaits des voyageurs ("un voyageur satisfait est un voyageur qui revient" - comme disait on ne sait plus qui). Il utilise en back-end un serveur **RServe** pour effectuer toutes les études statistiques.

Ce serveur, multithread Java/Windows-Unix (en modèle pool de threads), attend sur le PORT_DM des requêtes formulées par les analystes de ScienceAirport. Ceux-ci manipulent une application **Application_DataMining**, dont le rôle est de demander le traitement (notamment statistique) des informations disponibles dans BD_AIRPORT. Ceci sous-entend que cette base de données contient les informations nécessaires aux questions qui seront posées. De plus, le serveur écrit aussi dans une autre base BD_DECISIONS qui est la base décisionnelle : elle mémorise les résultats stratégiques importants obtenus lors des requêtes.

Le ***protocole applicatif*** (basé TCP) est **LUGANAP** (**LUG**gage **ANA**lysis **P**rotocol).

Sur base de ces spécifications, il vous appartient de définir les commandes de ce protocole LUGANAP, **donc de leur donner un nom et de choisir la manière de les implémenter (objets, chaînes de caractères, etc).**

Les outils utilisés ici seront, outre le JDK 1.8 classique, encore une fois :

- la librairie graphique **JFreechart** ;
- la librairie d'accès au serveur **RServe**.

2.3 Usages de l'application Application_DataMining

L'utilisation du premier prototype de l'application s'effectue de la manière suivante.

1) Un analyste de ScienceAirport se connecte au serveur.
2) Il choisit une année, éventuellement un mois dans cette année, éventuellement une compagnie aérienne.
3) Il peut alors formuler l'une des requêtes suivantes, qui se restreint au mois précisé (à toute l'année sinon) et à la compagnie choisie (toutes si aucune n'a été précisée). Toutes ces requêtes portent sur des questions de bagages, car une gestion optimale de ceux-ci permet de substantielles économies (ou des dépenses inconsidérées !)

3.1) REG_CORR_LUG : on demande d'étudier la relation entre le poids des bagages des voyageurs (donc poids moyen par personne) et la distance parcourue par le vol qu'ils ont pris.

L'application client génère ensuite (avec JFreeChart), sur base des résultats retournés par le serveur, un graphique "nuage de points" et un histogramme sur les distances parcourues : histogramme simple pour une année ou comparé avec le mois suivant si un mois a été précisé.

3.2) REG_CORR_LUG_PLUS : idem (mais sans les graphiques), en tenant compte en plus du nombre d'accompagnants et/ou de l''âge des voyageurs.

3.3) ANOVA_1_LUG : on demande d'étudier la relation entre le poids des bagages des voyageurs (donc poids moyen par personne) et la zone géographique de destination (zones possibles: EUR, AM-N, AM-S, AS, AFR-N, AFR-SUBSAH, AUST, OCEA). L'application client génère ensuite (avec JFreeChart), sur base des résultats retournés par le serveur, un "diagramme à moustaches" et diagramme sectoriel du poids total par région.

3.4) ANOVA_2_LUG_HF : idem (mais sans les graphiques) en tenant compte en plus du sexe du voyageur (Homme ou Femme).

3. Les applications mobile Android

3.1 L'application mobile Android pour le personnel sur pistes

L'application Android **Application_Pistes** est destinée aux employés de l'aéroport qui travaillent sur la piste, notamment les conducteurs des véhicules transportant les bagages, les bagagistes travaillant dans les soutes des avions et les responsables de portes d'embarquement. Il s'agit en fait d'une version très allégée de l'application Application_Bagages (voir évaluation 3 du projet ScienceAirport du laboratoire de "Réseaux et technologies Internet"). En effet, il s'agit simplement pour un employé "sur piste" de réaliser les opérations suivantes :

- se connecter au serveur
- choisir sa langue d'utilisation
- obtenir la listes des bagages qu'il doit trouver sur les charriots parvenus à l'entrée de la soute
- cocher tous les bagages au fur et à mesure de leur entrée
- retourner la liste complétée une fois les charriots vides.

Sur base de ces spécifications, il vous appartient de définir les commandes du protocole LUGAPM (LUGgage hAndling Protocol for Mobile), donc de leur donner un nom et de choisir la manière de les implémenter (objets, chaînes de caractères, etc).

3.2 L'application mobile Android pour décideurs

Il s'agit ici de construire une première version pour mobile de l'application **Application_DataMining** décrite ci-dessus (**Application_DataMiningMob)**. Il s'agit d'une version

- simplifiée car on n'implémentera que les requêtes REG_CORR_LUG (avec "nuage de points" et histogramme simple) et ANOVA_1_LUG (avec seulement diagramme sectoriel);
- adaptée : les résultats principaux sont sauvegardés dans une "base décisionnelle" SQLite locale au téléphone; on pourra consulter cette "base" à la demande;
- améliorée : l'utilisateur peut choisir sa langue d'utilisation pour REG_CORR_LUG;

- évoluée : on utilise la technologie des fragments.

Le ***protocole applicatif*** (basé TCP) **LUGANAP** est évidemment adapté en (**LUGANAP** for **M**obile).