==========

Date de rédaction : 22/01/2016
Nom du rédacteur du document : Sébastien Iksal
Spécialités : Informatique
==========

Cas d'étude Hubble: MOOCAZ
Scénario hubble : Scénarios 2 et 4
Personnes impliquées pour la collecte et l'analyse : Matthieu Cisel - Pierre Laforcade (élaboration d'un scénario d'analyse) - Christophe Choquet (élaboration d'un scénario d'analyse) - Sébastien Iksal (pré-traitement + Intégration + mise en oeuvre UTL) - Serge Garlatti (Analyse)
Période de la collecte : Printemps 2015
Periode de l'analyse : Décembre 2015 / Février 2016

Dispositif d'apprentissage (Etude de cas de Hubble)

Type de dispositif : MOOC
Finalité de l'apprentissage : Apprendre à monter un MOOC
Utilisation du dispositif et fonctionnalités : Consommation de vidéo, réalisation de quizz, de devoirs, forums de discussion

Contexte de production de données : MOOC organisé sur FUN

Au besoin indiquer les différentes moments de la production (savoir si des données ont été produites sur plusieurs années)

Décrire en quelques mots la problématique posée :
Scénario 2 V1 : On veut suivre les évolutions de types d'apprenants définis à l'avance. (8 comportements définis sur une semaine donnée)

Scénario 4 : On veut essayer de suivre des parcours d'apprenants et les changements d'état d'un format d'engagement à l'autre au fil des semaines

Objectifs de l'analyse : Reproduire des analyses réalisées dans le cadre d'un autre MOOC (Coursera sur une autre plateforme)

Description du stockage des données:

Plateformes/outils utilisés: Les données sont stockées sur les serveurs de STEF en premier lieu, puis après nettoyage intégrée dans UTL.

Points forts de ces plateformes	Points faibles
Une fois nettoyée les données sont faciles à intégrer. Accessibilité à la demande et langage de requête disponible pour préparer les données. Supporte les gros fichiers (MOOCAZV2 => 800Mo)	Ne sait pas gérer les données mal structurées et corrompues.

Production des données avant le traitement :

Décrire le processus de production des données brutes : Les serveurs du CINES stockent en théorie tous les logs de la plate-forme. Une extraction est requise, les logs extraits par l'équipe du CINES sont envoyés à FUN qui nous les retransmet.

Liste des variables initiales : L'ensemble des variables du jeu de données sont disponibles sur la documentation d'edX
http://edx.readthedocs.org/en/latest/internal_data_formats/tracking_logs.html
Parmi les nombreuses variables qui nous intéressent, nous allons pour le moment nous intéresser aux événements de déclenchement de vidéo, de téléchargement de vidéo, et de réalisation de quizz

Plateformes/outils utilisés:

Points forts	Points faibles
Documentation disponible au niveau de FUN	FUN avait jusqu'à peu des problèmes de stockage des logs (des pertes importantes ont été constatées -30- 80%), Données corrompues, format décrit par FUN non respecté

Description des pré-traitements: Transformer un fichier de logs qui se présente comme un mélange de JSON pas très propre en un fichier JSON épuré des lignes corrompues puis conversion en XML pour un import rapide dans la base de données XML Native d'UTL.

Objectifs des pré-traitements : Rendre un fichier de départ collectables par UTL et R.

Décrire le processus de pré-traitement: Suppression dans le fichier original de l'entête des lignes inutile, ce qui a permis d'obtenir un fichier composé de lignes JSON. Suppression des lignes corrompues (tronquées par FUN). Conversion du JSON en XML.

Outils utilisés: Programmes JAVA développés par S. Iksal en utilisant une librairie JAVA pour la manipulation de JSON

Points forts	Points faibles
Rapide, efficace	

Description des analyses : (Faire une description de chacune des analyses conduites)

Description Analyse Scénario 2 V1

Objectif de la création de ces nouvelles données : UTL nécessite de convertir les traces en données brutes, utilisées par des données intermédiaires avant d'obtenir des indicateurs

Description du résultat attendu : Nous souhaitons étudier le comportement des individus tout le long du MOOC. Nous définissons, à partir des trois types de ressources pédagogiques, quatre actions possibles : avoir vu une vidéo, avoir téléchargé une vidéo, avoir répondu à un quiz, avoir effectué un devoir. Nous allons donc étudier la séquence d'actions des individus par semaine. Nous considérons que la même activité au sein d'une même journée (voir deux fois la même vidéo par exemple) est un doublon et ne considérons donc que les actions uniques de chaque jour. En revanche, durant une semaine, un individu peut visionner plusieurs fois la même vidéo. A partir de ces séquences, nous avons déterminé des états par semaine pour chaque individu :

Nom	Description
Viewer	Un apprenant ne regardant que les vidéos
Collector	L'apprenant a uniquement téléchargé les vidéos
Quizzer	L'apprenant se sera contenté d'effectuer les quiz
Active viewer	Un apprenant qui mêle actions de vidéos et de quiz
Completer	Un apprenant qui fait toutes les activités possible au sein de la semaine
Low completer	individus effectuant les devoirs couplés avec un autre type d'actions
Solver	Un individu ayant effectué uniquement les devoirs
Inactive	n'a effectué aucune action cette semaine

Cette classification nous permet d'observer l'évolution des individus durant les semaines de cours, nous excluons les activités précédant la première semaine et ayant lieu après la dernière semaine.

Liste des méthodes mise en œuvre : Filtrage des données brutes afin d'identifier les différentes éléments représentatifs des actions suivies. Construction de données

intermédiaires permettant de classifier les apprenants selon les semaines/activités. Création d'un ou plusieurs indicateurs permettant de compiler ces classifications.

Mode opératoire technique, logiciels utilisés : UTL + description des calculs en DCL4UTL

Résultats : Certains événements ont pu être identifiés (play_video, problem_check, load_video), nous avons néanmoins de gros soucis liés au contenu du champ event_type qui ne correspond pas toujours au contenu décrit par la documentation de FUN. Il n'y a pas d'événement indiquant le téléchargement d'une vidéo. Nous nous interrogeons sur les événements permettant d'identifier les devoirs.

Nous devrions trouver des événements comme :

- edx.course.enrollment.activated
- edx.course.enrollment.deactivated
- edx.forum.searched
- hide_transcript
- load_video
- openassessmentblock.create_submission
- openassessmentblock.get_peer_submission
- openassessmentblock.peer_assess

Mais nous trouvons dans le champ event_type des informations comme :

- /courses/ENSCachan/20002S02/Trimestre_3_2014/xblock/i4x:;_;_ENSCachan;_20002S02;_sequential;_fa03204a375b4569bea35a51eb8d2dcf/handler/xmodule_handler/goto_position
- /courses/ENSCachan/20002S02/Trimestre_3_2014/xblock/i4x:;_;_ENSCachan;_20002S02;_video;_1f2c24f3460a4b3482dba4aaa29baa34/handler/transcript/translation/en
- /create_account
- /export/ENSCachan/20002S02/Trimestre_3_2014
- /get-grades/ENSCachan/20002S02/Trimestre_3_2014/ENSCachan_20002S02_Trimestre_3_2014_grade_report_2014-12-16-1533.csv

Ce qui nous donne presque 3000 event_type différents que nous ne comprenons pas. Le travail de compréhension est en cours.

Points forts des analyses	**Points faibles des analyses**

Analyse Scénario 4 : Parcours de participants au sein de la plate-forme

Liste des méthodes mise en œuvre :

Mode opératoire technique, logiciels utilisés :

Résultats :

Points forts des analyses	**Points faibles des analyses**

Description des Itérations
Pourquoi le processus d'analyse a été reproduit ?

Points forts des itérations	Points faibles des itérations