

TP COMPIL : Traitement du fichier journal d'un site web

Objectif

On souhaite étudier la fréquentation d'un site web dont on dispose du fichier journal (logs).

Dans le cadre de ce TP, on se limitera à étudier les paramètres suivants :

1. L'estimation du nombre d'internautes différents ayant visité des pages web du site (autrement dit le nombre de sessions),
2. Et la durée moyenne d'une session,

Format des logs

On définira la syntaxe des logs en examinant un exemple de fichier-log fourni dont voici un extrait :

```
147.99.255.218 - - [18/Jun/2013:09:24:38 +0200] "GET /front/res/css/ui-lightness/images/ui-  
bg_highlight-soft_100_eeeeee_1x100.png HTTP/1.0" 200 400  
147.99.255.218 - - [18/Jun/2013:09:24:38 +0200] "GET /front/res/images/readMoreMask.png  
HTTP/1.0" 200 506  
147.99.255.218 - - [18/Jun/2013:09:24:38 +0200] "GET /front/res/ckeditor/config.js?t=B49E5BQ  
HTTP/1.0" 200 889  
147.99.255.218 - - [18/Jun/2013:09:24:38 +0200] "GET  
/front/res/ckeditor/skins/kama/editor.css?t=B49E5BQ HTTP/1.0" 200 5055  
147.99.255.218 - - [18/Jun/2013:09:24:38 +0200] "GET /front/res/ckeditor/lang/fr.js?t=B49E5BQ  
HTTP/1.0" 200 7796  
147.99.255.218 - - [18/Jun/2013:09:24:38 +0200] "GET /front/res/ckeditor/contents.css  
HTTP/1.0" 200 692  
147.99.255.218 - - [18/Jun/2013:09:24:38 +0200] "GET  
/front/res/ckeditor/plugins/styles/styles/default.js?t=B49E5BQ HTTP/1.0" 200 964  
147.99.255.218 - - [18/Jun/2013:09:24:43 +0200] "GET /front/res/images/lgo-up.png HTTP/1.0"  
200 941  
157.55.33.83 - - [18/Jun/2013:09:24:44 +0200] "GET /robots.txt HTTP/1.1" 403 525  
66.249.73.24 - - [18/Jun/2013:09:24:46 +0200] "GET  
/front/document/b78ba2be/a61d/4265/b78ba2be-a61d-4265-833e-  
52bdc42f869a/Cours_XML_UOH/skin/img/txt/pucel.gif HTTP/1.1" 200 591  
66.249.73.24 - - [18/Jun/2013:09:24:52 +0200] "GET  
/front/document/b78ba2be/a61d/4265/b78ba2be-a61d-4265-833e-  
52bdc42f869a/Cours_XML_UOH/skin/img/quiz/fondEvalPrint.png HTTP/1.1" 200 708  
66.249.73.24 - - [18/Jun/2013:09:24:52 +0200] "GET  
/front/document/b78ba2be/a61d/4265/b78ba2be-a61d-4265-833e-  
52bdc42f869a/Cours_XML_UOH/skin/img/btn/printSubWin.png HTTP/1.1" 200 1576  
66.249.73.24 - - [18/Jun/2013:09:24:58 +0200] "GET  
/front/document/b78ba2be/a61d/4265/b78ba2be-a61d-4265-833e-  
52bdc42f869a/Cours_XML_UOH/skin/css/struct_ref.css HTTP/1.1" 200 695  
66.249.73.24 - - [18/Jun/2013:09:24:59 +0200] "GET  
/front/document/b78ba2be/a61d/4265/b78ba2be-a61d-4265-833e-  
52bdc42f869a/Cours_XML_UOH/skin/img/blocks/example.png HTTP/1.1" 200 5119  
66.249.73.24 - - [18/Jun/2013:09:25:10 +0200] "GET  
/front/document/964dccff/e9fc/413a/964dccff-e9fc-413a-987b-ffff0284676c/co/Cours3.html  
HTTP/1.1" 304 212  
173.230.129.121 - - [18/Jun/2013:09:25:32 +0200] "GET /front/notice?id=51a2944d-2e8d-4c4d-  
8579-ad4c61c5d070 HTTP/1.1" 200 32088  
173.230.129.121 - - [18/Jun/2013:09:25:33 +0200] "GET /front/notice?id=51a2944d-2e8d-4c4d-  
8579-ad4c61c5d070 HTTP/1.1" 200 32088  
157.55.33.83 - - [18/Jun/2013:09:25:38 +0200] "GET  
/front/document/cfed958a/0671/4298/cfed958a-0671-4298-a073-  
38950631eb5a/UOHEDUprod/mediassite.html HTTP/1.1" 404 556  
80.14.188.85 - - [18/Jun/2013:09:25:51 +0200] "GET /front/service/sujet?id=3759a400-f2e7-499d-  
ad88-ff2c6b9af8c7 HTTP/1.1" 200 8913
```

Notion de session

On conviendra qu'une session correspond à la séquence d'entrée dans le fichier-log associée à un même numéro IP sachant que deux actions consécutives d'une même session sont en décalage de moins de 10 minutes.

Consignes

1. On commencera par définir la grammaire d'un fichier log.
2. On implémentera ensuite cette grammaire en CUP/JLEX.
3. Et on rajoutera les actions en Java permettant de construire une structure de données adaptée à l'étude visée.

Ce TP est à réaliser par binôme. Cependant, **chaque membre** du binôme devra *uploader* une distribution source du projet dans l'Atelier prévu à cet effet.

NB : Le fichier-log fourni n'est pas public. En conséquence, il est formellement interdit de le redistribuer

Dates à retenir

- Dépôt de votre travail sur le site : le 17 février 2017 à 23 :55
- Fin des évaluations : le 27 février 2017 à 23 :55