HELLO MARCEL !

LE 18 MARS 2010 ***LÉO GOURVEN***

Analyser À la recherche du temps perdu à l'aide des outils de visualisation de données : voilà le projet original entamé par Léo Gourven, étudiant à Hetic. Il explique sa démarche, que vous pourrez suivre sur son blog Data_Proust et fait appel aux bonnes volontés pour l'aider.

Analyser *À la recherche du temps perdu* à l'aide des outils de visualisation de données : voilà le projet original entamé par Léo Gourven, étudiant à Hetic. Il explique sa démarche, que vous pourrez suivre sur son blog Data_Proust et fait appel aux bonnes volontés pour l'aider.

Je travaille depuis un petit mois sur un drôle de projet. Je me remettais doucement de la lecture de *A La Recherche Du Temps Perdu* de Marcel Proust et parallèlement, je travaillais dans le cadre de mes études autour d'un projet lié aux visualisations graphiques. Alors je me suis dit (innocemment) : pourquoi le petit Marcel n'aurait pas droit à sa data visualization ?

Et au fur et à mesure je me suis rendu compte que l'œuvre de Proust justifiait tout particulièrement cette approche scientifique barbare :

Le roman est immense ! 1,5 millions de mots !
C'est un roman fleuve, il va de l'enfance à la mort.
L'écriture de Marcel Proust est quasi scientifique. Il suit une sorte de recette, on avance par étape.
Tout le monde connait Proust ! (et personne ne l'a lu). Et tout le monde se demande depuis ses 4 ans si ses phrases sont si longues que ce l'on raconte ?
Libre de droit et numérisé.
J'ai (re)découvert que dans les années 80 (quand je n'étais même pas né quoi), un certain Brunet Étienne avait déjà travaillé sur le **sujet**, mais en se concentrant sur l'aspect statistique (les occurrences les plus répétées, nombre de mots, de phrases etc). Ce n'est pas énorme (ça l'était pour l'époque), mais c'est déjà extrêmement intéressant.

Dans mon cas, une des premières choses à faire, c'est transformer l'information en donnée structurée (J'avais l'habitude de faire le contraire mais bon). C'est-à-dire mettre la *Recherche* dans une base de données, séparer chaque phrase, l'identifier et – dans un second temps- l'enrichir (De quel tome vient-elle ? Où se déroule l'action de cette phrase ? Quel temps est utilisé ?).

À partir de là, je pourrai opérer quelques traitements statistiques, a priori je débuterai par une étude du nombre de mots par phrase. Mais l'intérêt de cet outil prendra tout son sens une fois que l'on pourra superposer le nombre de mots par phrase avec les lieux, les éléments clés de l'action, etc. (Ce qui permettra de répondre à des questions du type : de quelle manière évolue la longueur des phrases en fonction face à la mort de sa grand mère ?)

Mais pour cela il faut que je trouve un outil d'analyse linguistique qui puisse me séparer mes phrases (Pas si simple qu'il n'y parait). Si quelqu'un maitrisant ce domaine arrive sur ce blog, j'ai besoin d'aide ! Envoyez moi un mail !

En bref, je vous raconterai sur ce blog comment mon projet avance, je causerai technique, je diffuserai mon code, je vous appellerai à l'aide mais je ne vous dirai pas que je suis fatigué !

JEAN
le 22 mars 2010 - 9:17 • SIGNALER UN ABUS - PERMALINK

Je trouve votre démarche très amusante.. Je pense aussi que la data vizualisation serait une évolution logique des éditeurs, dont le métier consiste justement à organiser de la donnée… J'en ai fais une petite note : http://notrelienquotidien.com/2010/03/04/la-data-visualization-nouvel-eldorado-des-editeurs/

VOUS AIMEZ 0 VOUS N'AIMEZ PAS 0 LUI RÉPONDRE

1 ping

uberVU - social comments le 19 mars 2010 - 0:21

Social comments and analytics for this post…

This post was mentioned on Twitter by mathemagie: [#owni] Hello Marcel ! http://goo.gl/fb/g90Y...

Raw Data_proust now ! | Owni.fr le 1 avril 2010 - 17:01

[...] Gourven a un projet fou: analyser À la recherche du temps perdu à l'aide des outils de visualisation de données. Il [...]