RAW DATA_PROUST NOW !

LE 1 AVRIL 2010 ***LÉO GOURVEN***

Léo Gourven a un projet fou: analyser À la recherche du temps perdu à l'aide des outils de visualisation de données. Il nous tient informé des dernières évolutions, notamment en ce qui concerne la transformation du texte en base de données.

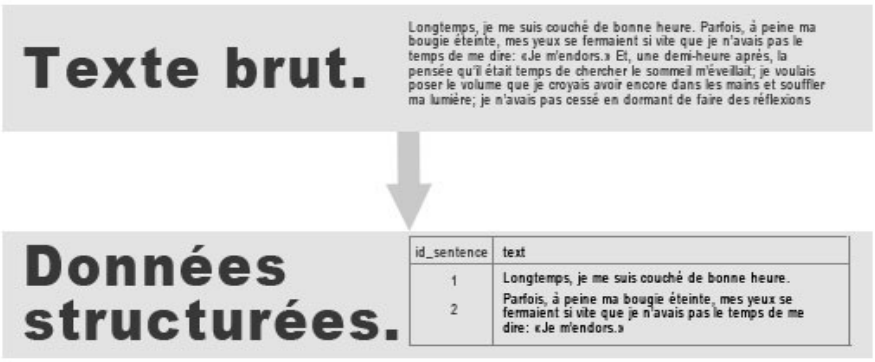
Léo Gourven a **un projet fou:** analyser *À la recherche du temps perdu* à l'aide des outils de visualisation de données. Il nous tient informé des dernières évolutions, notamment en ce qui concerne la transformation du texte en base de données.

—

Les semaines passées, je me suis cassé la tête pour essayer de trouver un outil qui me permette de passer du texte brut à une sorte de base de données. Je viens enfin de trouver chaussure à mon pied !

EXPLICATIONS :

Pour pouvoir faire des statistiques, il faut des variables. Elles décrivent des caractéristiques : des lieux, des personnages, un numéro de Tome ou tout ce que l'on souhaite. Dans mon cas, **je dois déterminer et extraire des caractéristiques intéressantes du texte.**



Ce genre de chose, **ça s'appelle le traitement automatique des langues (TAL)**. En gros ça veut dire, que des chercheurs développent **des algorithmes pour extraire automatiquement des données** d'un corpus. Parfait. Manque de chance, ces gens qui ont oublié d'être cons, ont aussi oublié ce qu'était une interface graphique. Ce qui me compliquait légèrement la tache (faut de la force pour se plonger là dedans). MAIS, j'ai fini par trouver la petite perle.

GATE est un logiciel open source qui regroupe les codes des chercheurs cités plus haut, **mais avec une interface graphique** (y'a même des screencasts, que demande le peuple). C'est un peu complexe, mais accessible. J'ai donc rentré mon petit texte et exécuté mes traitements.

MAIS QUE FUT DONC MA SURPRISE, en découvrant le nombre de données que m'a sorti Gate en permettant de séparer : Les mots, les phrases, les paragraphes, les lieux, les métiers, les personnages, les sommes d'argents et d'autres choses encore.

Je ne pensais pas que l'on pouvait aller aussi loin dans l'analyse et mon projet gagne encore en grandeur (comprenez en travail !). Je vais donc continuer à tripatouiller Gate et essayer d'avoir un beau XML qui décrive des caractéristiques intéressantes à analyser.

Prochaines étapes :

> Diffuser les XML !
> Je rencontre deux amis graphistes pour discuter ergonomie ce week end.
> Ouvrir un serveur de développement et un Github.
> Et par là, proposer aux personnes que ça motive de m'aider dans le développement !

—

> Article initialement publié sur Data Proust