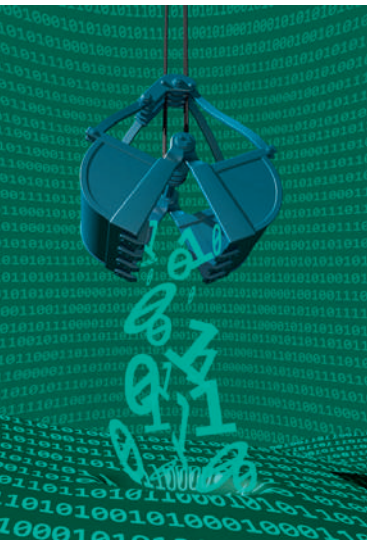
La grande prospection



Dans le langage courant, le terme de «prospection» évoque l'activité des chercheurs d'or ou l'exploration des potentialités du marché. L'informatique parle de data mining (en français, «prospection» ou «exploration de données»), faisant référence au traitement automatique de larges volumes de données liées statistiquement. Un grand nombre de personnes utilisent des algorithmes. En effet, les moteurs de recherche et les correcteurs d'orthographe font partie du quotidien. Il en va de même des programmes de reconnaissance de l'écriture, des profils de clients ou de la cryptologie, qui servent à protéger les rapports médicaux ou à effectuer des enquêtes fiscales. Les consommateurs, c'est-à-dire nous tous, ne s'en inquiètent pas, tant que cette industrie minière digitale leur permet de découvrir des informations utiles et agréables.

Une exploration assez poussée peut mener à des découvertes. Le terril des métadonnées foisonne de nouvelles opportunités de carrière. «Compter plutôt que lire», telle est la nouvelle devise adoptée depuis que *Google Book Search* a numérisé 15 millions de livres. Le déchiffrage du pseudonyme de l'auteure d'Harry Potter à l'aide d'un programme d'analyse de texte a au moins permis de combler le creux médiatique de l'été. La combinaison de la statistique et de l'analyse stylistique ont fait naître une nouvelle discipline: la «culturomique». Il s'agit d'un mélange entre les sciences culturelles, la linguistique et l'économie qui cherche les tendances cachées à l'intérieur des corpus, parmi les innombrables phrases et les mots qui les composent. Les linguistes informaticiens et les informaticiens spécialistes de la littérature, qui y consacrent des thèses, veulent déchiffrer les langues en se servant des méthodes mathématiques de la génomique. C'est de là que vient d'ailleurs le terme «culturomique» qui, par analogie à cette science, fait référence au décodage du patrimoine génétique de toutes les cultures. Qui est capable de lire des millions de livres? La lecture à distance de larges corpus vient compléter la méthode traditionnelle, c'est-à-dire le pénible travail de lecture et d'interprétation d'œuvres individuelles. *Les algorithmes remplacent l'interprétation subjective par des recherches évaluables et quantitatives*. Le diplôme de master en technologies du langage de l'EPFZ correspond aux profils professionnels recherchés actuellement par les bibliothèques et les éditeurs. Ils satisfont également à la demande dans les domaines de la sociolinguistique, de la psycholinguistique, de la neurolinguistique, de la patholinguistique, de l'informatique judiciaire et de la recherche en filtration de données. Dans tous les cas, un gigantesque volume de textes est scanné, puis trié selon des modèles sémantiques, lexicaux et syntaxiques. L'une des premières prises de position concernant ce procédé s'intitule «Comment ne pas lire un million de livres» [1]. Plusieurs professions utilisent depuis longtemps des mots déclencheurs et des profils d'auteurs à des fins d'identification. Si Google offre gratuitement aux consommateurs un jeu géographique, un sous-produit de Streetview [2], c'est sans doute pour redorer son blason après les constantes violations des droits d'auteurs et de la vie privée dont elle s'est rendue coupable. Un autre programme [3] permet d'effectuer l'analyse statistique de plus de 5 millions de livres numérisés, qui ont été publiés entre les années 1800 et 2000 en anglais, en français, en allemand et en espagnol. Pour vous donner une petite idée de ce que signifie l'exploration de données, cela représente un volume de données d'environ 500 milliards de mots.

Les textes sont découpés en lettres, en phonèmes et en mots. Les fragments qui en résultent se nomment des n-grammes; ils permettent entre autres de créer des programmes tels que *Loading Friendship*, propriété de Facebook. Des cercles et des lignes de diverses couleurs établissent une carte du réseau social à partir du carnet d'adresses de l'utilisateur. Twitter propose également des «nuages d'amis» similaires, ou clusters. Il s'agit de graphiques colorés montrant exactement qui est en relation avec qui, et à quelle fréquence. Les ennemis de l'Etat tels que Julian Assange, de WikiLeaks, et Edward Snowden, de l'Agence nationale de la sécurité américaine (la NSA), ont rendu publique l'efficacité des programmes d'espionnage. Le fait que ses filiales du monde entier collectent des données, qu'il s'agisse d'e-mails privés, de conversations téléphoniques, de recherches sur Internet, de communications cryptées ou de données provenant d'achats en ligne, n'est un secret pour personne.

Le plus surprenant, c'est que ces nouvelles sont accueillies avec une certaine indifférence. Si elles n'avaient pas été utilisées à des fins stratégiques lors de batailles électorales, elles seraient tout simplement passées inaperçues durant le creux des vacances d'été. Une saturation médiatique ou un désintérêt pour tout ce qui touche à la politique, combiné à une certaine tendance à oublier les conséquences des dictatures, peuvent expliquer ce phénomène. Peut-être avons-nous trop besoin de nous sentir en sécurité et avons-nous peur de remettre en question notre sentiment naïf de liberté. Peut-être que les défenseurs de la cause nous énervent simplement parce qu'ils nous rappellent que ces technologies si pratiques et faciles à utiliser contribuent inexorablement à la surveillance par l'Etat. «Cela n'arrive qu'aux autres», voilà ce que pense tout un chacun, avant de reprendre sa petite vie bourgeoise.

Erhard Taverna

1 http://people.lis.illinois.edu/
2 www.geoguessr.com
3 www.ngrams.googlelabs.com

erhard.taverna[at]saez.ch