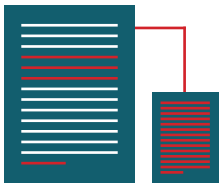


Le Text Mining

Le Text Mining, qu'est-ce que c'est ?

Le Text Mining est une branche du Data Mining qui se spécialise dans le traitement de corpus de textes pour en analyser le contenu puis en extraire des connaissances. Les principales tâches à accomplir consistent en la reconnaissance de l'information présente dans le document et son interprétation. Tout cela est possible grâce à une recherche sémantique reposant sur l'analyse du langage naturel et la gestion de bases de connaissances spécialisées. On peut par exemple distinguer une plainte d'un client à une demande d'information, ou même un spam d'un message publicitaire, en inspectant la tournure des phrases.

Pourquoi le Text Mining ?

La croissance des réseaux sociaux

« Sur les 3,025 milliards d'internautes à travers le monde, 2,060 milliards sont actifs sur les réseaux sociaux, soit 68% des internautes et 28% de la population mondiale. Le temps passé sur les réseaux sociaux quant à lui est de 2h par jour dans le monde et d'1h30 en France. »
@ blogdumoderateur

Depuis l'avènement des réseaux sociaux, les sources et les contenus de données n'ont cessé de croître exponentiellement. Face à cette tendance, les entreprises maintiennent une volonté de conserver un maximum d'informations exploitables, suffisamment pour remplir des péta-octets de données brutes. La majorité de ces informations reste encore aujourd'hui inexploitée.

Des données non structurées à valoriser

« De nos jours, de plus en plus d'informations sont stockées dans des formats non structurés et partiellement structurés (messages électroniques de clients, notes de centre d'appel, réponses ouvertes à des enquêtes, actualités, formulaires Web, etc.). Ce flot d'informations pose problème à de nombreuses organisations qui souhaitent trouver la méthode leur permettant de collecter, d'étudier et d'exploiter ces informations. »
@ IBM

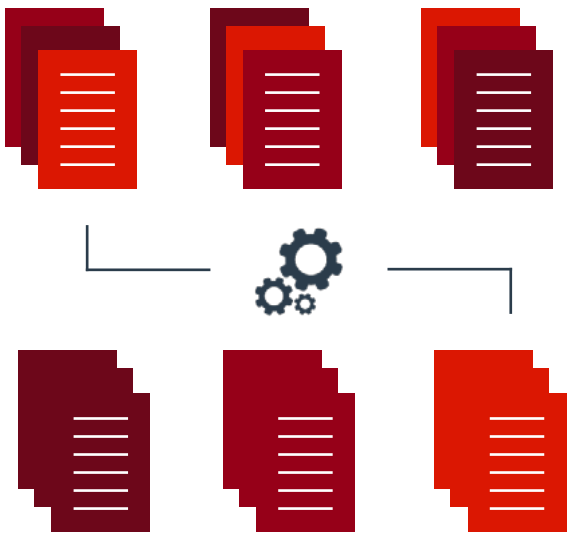
En moyenne, 80% des données des SI des entreprises sont non structurées. Jusqu'à encore quelques années, ces données ne faisaient qu'être stockées, car difficilement exploitables de par leur format. Cependant, de nouvelles technologies nous permettent d'exploiter toujours plus ces données.

Le Text Mining, les concepts

Le Text Mining permet de rechercher une information précise expertisée pour un contexte spécifique, de comparer plusieurs documents et en déduire une opération adaptée. Il permet ensuite la réalisation de tâches telles qu'architecturer l'information extraite de façon à ce qu'elle soit réutilisable rapidement.

On fait appel au Text Mining aussi bien pour une base de textes volumineuse que pour un choix réduit d'articles. Dans les deux cas, les tâches réalisées consistent en l'analyse linguistique, statistique et sémantique de documents pour en extraire les informations pertinentes.

Il peut enfin servir à enrichir l'index d'un moteur de recherche pour améliorer la consultation des documents grâce à une représentation sémantique des données et aider ainsi à leur classification automatique.



De nouvelles possibilités de services

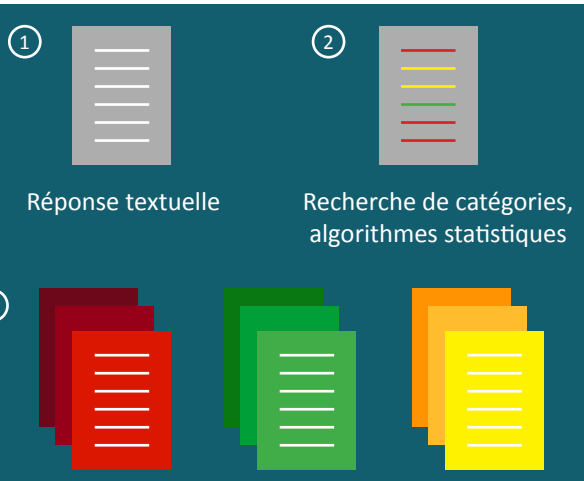
Fonctionnalités du Text Mining

Le Text Mining apporte de nouvelles fonctionnalités au Data Mining, présentées ci-dessous.

Clustering

Un cas usuel d'utilisation Text Mining est la catégorisation de documents de manière non supervisée suite à la collecte de données texte non-structurées. Prenons l'exemple d'un sondage d'opinion public, réalisé par l'entreprise Eaagle aux Etats Unis et qui a posé la question suivante : What should be the priorities for the next US president?

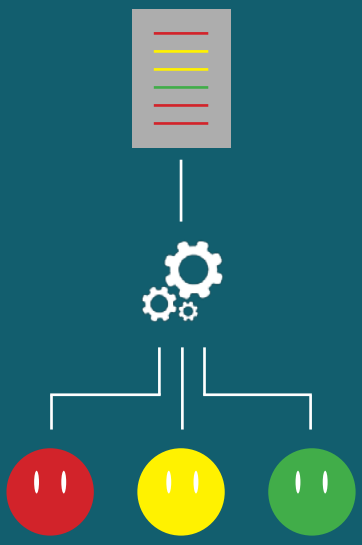
Comme il s'agit d'une question ouverte, il est essentiel d'avoir un outil capable de traiter les réponses textuelles. La première étape consiste alors en la création de catégories, pour trier les réponses, en analysant la fréquence d'apparition des termes ou de groupes de termes. La deuxième étape consiste en la classification les réponses à l'aide d'algorithme de Clustering selon les catégories.



Réponse textuelle

Recherche de catégories, algorithmes statistiques

Catégorisation des réponses par thème à l'aide d'algorithmes de Clustering



Analyse de sentiments

Qu'ils soient sur les réseaux sociaux ou sur les sites de ventes en ligne, les utilisateurs sont incités à donner leur avis. Si dans certains cas il est possible de noter un produit ou un film en cochant des étoiles, il s'agit la plupart du temps d'un commentaire écrit par l'internaute, ce qui représente une donnée non-structurée.

À partir de modèles pré-entrainés, le Text Mining permet d'évaluer si un commentaire est plutôt positif ou négatif. Les méthodes les plus élémentaires, basées sur des analyses statistiques comme l'attribution de poids aux mots clés à valeur sentimentale, permettent de générer des pourcentages de positivité.

Quant aux méthodes actuelles les plus poussées, elles passent par des outils d'analyse du langage naturel afin de détecter le sens de la phrase et d'en estimer le sentiment correspondant, ce qui rend possible la détection de plainte des clients ou encore de demande d'informations.

Extraction de relations

L'extraction de relations est une étape majeure issue du traitement du langage naturel. En effet, elle permet de détecter une relation sémantique entre un ou plusieurs groupes de mots. Par exemple:
La petite souris grignote le morceau de fromage dans la cuisine.

Cet outil est notamment utilisé pour déceler des relations précises dans un texte ou pour découvrir de nouvelles données alimentant des bases de connaissances.

Des cas d'utilisations très variés

Le texte non structuré est très commun et peut représenter la majorité des informations disponibles pour un projet de recherche ou d'exploration de données particulier. De ce fait, les connaissances que l'on pourrait en tirer sont la raison de l'expansion des techniques de fouille de texte et leurs applications dans plusieurs disciplines et secteurs d'activité variés.

Marketing

Connaissance client

Les données issues du contact avec le client (enquêtes, mails, lettre de réclamation, retranscription de messages téléphoniques..) constituent une source pertinente d'informations sur le client, venant renseigner notamment sur les besoins et leur adéquation avec les offres de service, ainsi que sur leur satisfaction et leur intention de fidélité. Le volume de ces données est important et ne cesse de croître. Le Text Mining va permettre de faire ressortir l'information pertinente de ces gros corpus de textes, pour une meilleure connaissance client.
Par exemple, la suite logicielle TEMIS Insight Discovered est utilisée depuis quelques années par EDF pour parfaire sa connaissance client.

Publicité

Les réseaux sociaux ont créé un nouvel univers qui permet de faire communiquer des personnes du monde entier sur des plateformes internet communes. Dans ce contexte, les agences Web, qui cherchent à proposer des annonces publicitaires pour mieux cibler les clients, se retrouvent face à plein de données à exploiter mais sous format non structuré. C'est alors qu'intervient le Text Mining, spécialisé dans ce type de données.

En effet, à l'aide du Text Mining, on peut s'inspirer des habitudes de l'internaute ainsi que de l'analyse sémantique du contenu texte qu'il lit et qu'il écrit, afin que s'affiche sur sa page une publicité suffisamment pertinente pour qu'elle attire son attention et qu'il clique dessus.

Fort de cette tendance, Critéo, une entreprise française fondée en 2005, utilise cette technologie aussi bien sur le web que dans les applications mobiles, ce qui a entrainé des ventes records chez ses clients.

Banques et assurances

Les banques et les assurances font de plus en plus appel à des systèmes automatisés pour optimiser leurs tarifs et pour mieux connaître leurs clients. Depuis des années l'usage du Data Mining est devenu coutumier, et, pour compléter cet outil les chercheurs ont mis en place des outils de Text Mining analysant directement les plaintes, constats et factures pour éviter toute analyse manuelle.

Des logiciels comme SAS® Enterprise Miner et Clementine Text Mining sont capables de traiter des données non-structurées en utilisant d'abord les technologies du Text Mining, pour l'analyse des textes, et par la suite des outils de Data Mining plus classiques.

L'enjeu est grand, le but ici est à la fois de tenter de prévoir le nombre d'accidentés pour l'année suivante, d'élaborer des profils précis des assurés, de connaître la satisfaction globale des clients. D'autres outils plus poussés permettent de détecter des constats et factures frauduleuses.

Biomédical

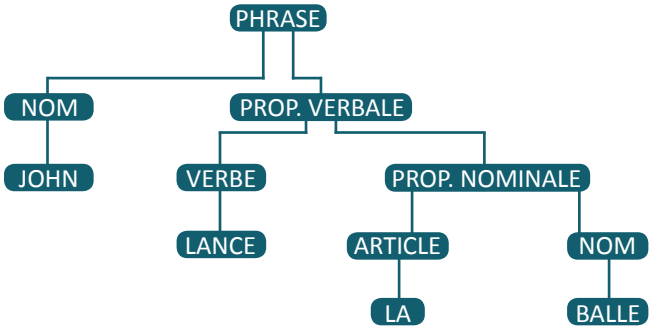
Il existe un intérêt croissant pour la fouille textuelle et les stratégies d'extraction de l'information appliquées à la littérature sur la biologie moléculaire et biomédicale, en raison du nombre croissant de publications électroniques disponibles stockées dans des bases de données telles que PubMed.

« Avec le développement des systèmes biologiques, les chercheurs ont tendance à comprendre les systèmes biomédicaux complexes d'un point de vue systèmes biologiques. Ainsi, la pleine utilisation du Text Mining pour faciliter les systèmes biologiques de recherche sur le cancer est en train de devenir une préoccupation majeure. »
@PubMed

Le Text Mining et ses technologies

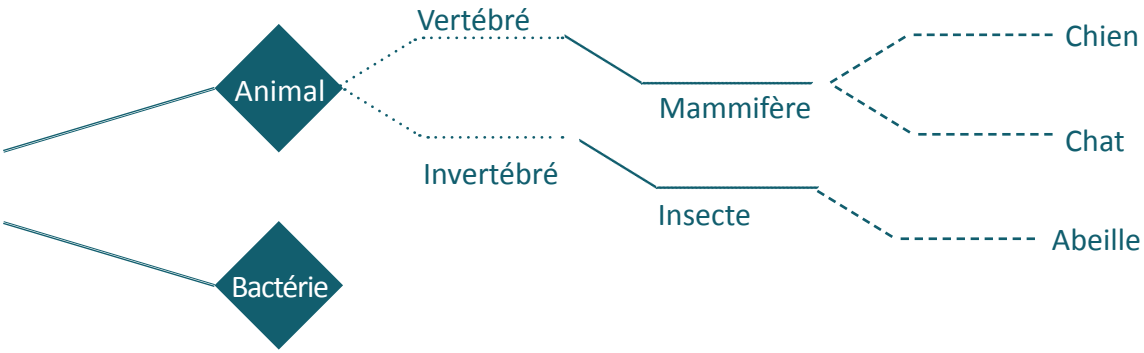
Le traitement du langage naturel

Les outils de traitement du langage naturel constituent la base du Text Mining. Leur rôle est d'apporter au texte un premier niveau de compréhension à partir d'une analyse syntaxique et grammaticale. Ces outils s'enchaînent un à un pour augmenter au fur et à mesure la compréhension des phrases. Ainsi, une détection des mots est suivie par une détection des phrases. Un outil va chercher les lemmes des mots, c'est-à-dire leur forme de base sans pluriel ni conjugaison, tandis qu'un autre va chercher leur nature grammaticale. Enfin, le dernier outil, le parseur, aura pour rôle de comprendre la structure globale de la phrase en analysant sujet, verbe, complément etc., en s'adaptant à la tournure des phrases.
Le résultat de ces outils de traitement du langage naturel servira ensuite d'entrée pour des fonctionnalités plus poussées comme l'analyse sémantique et la découverte d'informations.



Les bases de connaissances

Pour compléter la compréhension syntaxique des phrases, il faut saisir le sens des mots. L'utilisation de bases de connaissances permet de répertorier soit le sens des mots, au niveau linguistique, à la manière d'un dictionnaire, soit leurs fonctionnalités métiers liées à un contexte professionnel. Ce sont les liens entre ces sens qui vont permettre le raisonnement sur les phrases et donc une compréhension plus complète des notions formulées. Les connaissances sont stockées soit au moyen de graphes, soit selon les technologies du web sémantique comme les ontologies . Ces formats de données optimisent les parcours de graphes pour favoriser la connectivité entre les concepts. L'utilisation de bases de connaissances permet de limiter les reconnaissances lexicales de mots à base de statistiques pour identifier des sens au sein d'un document. Une des applications de ces technologies est la mise en œuvre de moteurs de recherche sémantique capable de comprendre une question et d'y répondre rigoureusement.



Le Machine Learning

Enfin, le Text Mining repose aussi sur des algorithmes non déterministes pour donner un sens au texte analysé. Pour cela, il s'appuie sur les principes du machine learning, qui consiste en l'entrainement de modèle via des données prétraitées, dit base d'entrainement, afin de faciliter les traitements futurs. De nombreux algorithmes peuvent être appliqués pour réaliser cette tâche. Il peut s'agir : d'algorithmes basés sur les statistiques (réseauBayesian, ...), mais aussi sur des bases géométriques (KNN, Bag of Feature, ...), arbre de décision (Deep Forest, ...), Support Vector Machine, réseau de neurones.

Vous voulez approfondir le sujet avec nous ?

Dans le cadre des recherches de la cellule Innovation, Aubay poursuit actuellement plusieurs projets de R&D afin d'approfondir son expertise technique autour du Data Mining, et plus particulièrement du Text Mining, et de mettre en œuvre de nouvelles solutions dans des approches techniques novatrices.
Si vous êtes intéressé soit pour être tenu au courant de l'avancée de ces travaux et des résultats obtenus, soit pour participer avec nous au déroulement de ces travaux, n'hésitez pas à nous contacter : innov-dt@aubay.com.