DÉTECTER L'INFORMATION CONFIDENTIELLE

avec un moteur de recherche intelligent

En s'appuyant sur de puissantes technologies d'indexation combinées au potentiel des algorithmes de Machine Learning et de Deep Learning, Sinequa permet aux entreprises d'exploiter leurs données non structurées. *Explications d'Adrien Gabeur (08), Directeur des Solutions Cognitives au sein de Sinequa.*

Adrien Gabeur (08)

Présentez-nous Sinequa.
Sinequa est un éditeur de logiciels indépendant de la French Tech. Nous fournissons aux entreprises multinationales et agences gouvernementales une plateforme d'analyse et de recherche intelligente.
La combinaison unique d'un moteur de recherche propriétaire éprouvé (Enterprise Search) avec des algorithmes avancés de NLP (Traitement du Langage Naturel), de Machine Learning et de Deep Learning permet à notre solution d'extraire des informations métiers à partir de données structurées, mais surtout non structurées.
Grâce à un travail d'innovation constant depuis 2017, Sinequa est reconnu leader dans le Magic Quadrant pour les Insight Engines réalisés par le cabinet d'analyste américain Gartner. Il en est de même pour le Forrester Wave conduit par le cabinet Forrester. Ce sont des reconnaissances prestigieuses pour un éditeur de logiciels européen.
En 2015, nous nous sommes implantés aux États-Unis avec des bureaux à Manhattan. Plus de 50 % de notre chiffre d'affaires est réalisé en Amérique du Nord où notre solution est déployée chez des clients emblématiques, comme la NASA qui a récemment choisi notre plateforme pour naviguer à travers son énorme base documentaire scientifique et réutiliser les savoir-faire accumulés au cours des anciennes missions spatiales.

Comment aidez-vous les entreprises à exploiter leurs données non structurées ?
Si les données non structurées connaissent une croissance exponentielle, elles restent difficilement exploitables, car elles sont de formats extrêmement divers (textuel, image, vidéo…) et sont disséminées dans toute l'entreprise.
Leur exploitation nécessite des solutions capables d'interpréter le langage naturel (texte) et ses subtilités dans toutes les langues.
Notre plateforme permet de relever l'ensemble de ces défis :

- elle propose des traitements avancés pour plus 23 langues ;
- elle s'appuie sur une librairie propriétaire de plus de 200 connecteurs qui permettent d'accéder aux différentes sources de données utilisées par les entreprises ;
- elle extrait les contenus à travers plus de 350 formats de fichiers.

Comment cela se traduit-il concrètement ?
Nous commençons par configurer nos connecteurs pour accéder en lecture aux différentes sources de données. Cela peut parfois représenter plusieurs centaines de millions de documents. Les données sont alors indexées dans notre plateforme et enrichies grâce à nos algorithmes de traitement du langage.
À ce stade, le texte est immédiatement disponible à la recherche et nous sommes déjà en mesure de reconnaître toute sorte de patterns, de concepts ou du vocabulaire spécifique au métier, que nous extrayons sous forme d'entités nommées.
Nous utilisons ensuite des algorithmes de Machine Learning pour entraîner, sur les données du client, des modèles capables de faire une analyse plus fine du contenu et dédiés au cas d'usages que nous adressons.

“Face à la croissance exponentielle du volume des données non structurées, les entreprises se retrouvent avec un corpus documentaire qui déborde d’informations, entre autres confidentielles. L’enjeu est d’analyser en temps réel les données pour identifier les éléments à protéger.”

En parallèle, nous créons aussi des applications métier, dites « Search-Based applications », pour permettre aux utilisateurs d’explorer, d’analyser et d’exploiter le corpus documentaire enrichi par nos analyses. À partir de ces applications, nous pouvons récolter le feedback des métiers. Cela nous permet d’améliorer constamment les modèles, mais aussi d’assurer que les prédictions restent précises, aussi bien dans le temps que dans le cadre de l’évolution des corpus.

Qu’en est-il en termes de cyber sécurité ?
Face à la croissance exponentielle du volume des données non structurées, les entreprises se retrouvent avec un corpus documentaire qui déborde d’informations, entre autres confidentielles. L’enjeu est d’analyser en temps réel les données pour identifier les éléments à protéger. L’évolution rapide de ces corpus rend la plupart des méthodes d’identification manuelle totalement inefficaces.
En parallèle, ces informations confidentielles prennent une multitude de formes en fonction des métiers de l’organisation : plan stratégique, informations clients, savoir-faire industriels, partenariat stratégique…
Nous aidons à résoudre ce problème en entraînant des modèles capables d’appréhender, pour chaque client, le contexte et l’essence du contenu, pour prédire avec précision un niveau de confidentialité, en accord avec ses règles internes de confidentialité. Une fois ces modèles déployés à une échelle industrielle sur notre plateforme, nous automatisons le processus d’identification et mettons à disposition des interfaces utilisateurs qui permettent, entre autres, de comprendre où se trouve la donnée confidentielle ou privée et de vérifier qu’elle est bien protégée.

Comment résumeriez-vous la valeur ajoutée de Sinequa ?
Notre plateforme se distingue par sa combinaison unique de technologies :

- l’évolutivité et la performance : la capacité de gérer, dans le cloud ou on-premise, de gros volumes de données ou Big Data ;
- la connectivité : la capacité de se connecter à toutes les sources de données dans les entreprises grâce à plus de 200 connecteurs ;
- le traitement avancé du texte en plus de 23 langues ;
- un moteur de recherche éprouvé qui permet d’interagir avec les données en fonction des problématiques utilisateurs ;
- la gestion des droits d’accès : dans chaque interface utilisateur que nous déployons, nous répliquons les droits d’accès en place dans la source d’origine ;
- le Machine Learning et le Deep Learning : la capacité d’entraîner sur les données de nos clients puis de mettre en production, à l’échelle, l’usage de modèles d’intelligence artificielle.

Nous sommes particulièrement mobilisés sur les technologies de Deep Learning qui évoluent constamment et constituent une véritable révolution. Nous analysons quotidiennement les papiers produits par la recherche fondamentale pour étudier leur potentielle application à nos cas d’usages et leur applicabilité dans les contraintes de notre marché cible (applicabilité au non structuré, hardware nécessaire, évolutivité, taille des ensembles d’apprentissages requis…). Si cela est pertinent, nous les optimisons et les intégrons ensuite à notre plateforme. Aujourd’hui, nous comptons un nombre croissant de clients qui utilisent cette technologie en production et à l’échelle sur notre plateforme.

La data est au cœur de votre expertise. Quelles problématiques adressez-vous dans ce cadre ?
Le principal défi est de fournir la bonne information au bon utilisateur, au bon moment et au bon endroit. C’est l’essence même de notre métier.
Cela peut prendre différentes formes. Plus particulièrement, dans le domaine de la cyber sécurité, nous créons, par exemple, une cartographie navigable qui donne avec précision une vue d’ensemble de toutes les données sensibles, confidentielles et privées de l’entreprise.

En parallèle, quels sont les sujets qui vous mobilisent actuellement ? Qu’en est-il de vos perspectives ?
Le Deep Learning est une technologie encore émergente. À l’heure actuelle, très peu d’entreprises sont capables de l’utiliser à une échelle industrielle pour générer du retour sur investissement. En parallèle, l’arrivée récente des modèles « Deep Language » et le développement des techniques de Transfer Learning laissent entrevoir une multitude de fonctionnalités nouvelles. Notre objectif est de les embarquer sur notre plateforme pour que nos clients puissent en bénéficier. Mon rôle au sein de Sinequa consiste à identifier les nouveaux cas d’usages qui peuvent en découler et développer cette offre en Europe et aux États-Unis.