SAS® TEXT MINER

Instrumenter la recherche et l'analyse de grandes collections de documents

Les entreprises recueillent chaque jour d'énormes quantités d'informations textuelles, dans des langues et dialectes divers. Retours et messages clients, emails, documents web, blogs, flux Twitter, réclamations, sondages, études de marché, brevets, veille concurrentielle... La liste est longue. Personne n'a le temps de tout lire et encore moins d'organiser, de classer ou d'interpréter les informations essentielles.

Pour observer des tendances, déceler de nouveaux thèmes, émettre des alertes sur des problèmes potentiels et identifier de nouveaux indicateurs d'activité, vous devez analyser la totalité de vos données avant d'agir. Or, le langage conversationnel est ambigu, et les messages clés enfouis dans les textes sont difficiles à identifier et à traiter. La plupart des entreprises n'ont pas les ressources nécessaires pour combiner des informations textuelles et des données structurées dans une optique décisionnelle.

Avec SAS Text Miner, vous analysez les données internes stockées dans votre système informatique et collectez dynamiquement des contenus externes pertinents récemment publiés sur le web. Vous explorez de manière interactive et identifiez automatiquement des thèmes grâce à des catégories élaborées à partir des données, et révélez des relations et des associations entre les termes. Au travers d'une interface graphique unique, vous explorez à la fois les informations textuelles et les données structurées, puis intégrez directement les modèles obtenus aux systèmes de scoring en place. Utilisez cette solution pour faire correspondre les CVs aux postes vacants, prévoir les résultats d'un traitement médical selon les différentes thérapeutiques suivies, identifier les facteurs qui poussent l'acheteur à passer à l'acte, etc.

Que fait SAS Text Miner ?

SAS Text Miner met en évidence les informations dissimulées dans des collections de textes. Grâce à la lecture automatisée de données textuelles, et aux algorithmes permettant de réaliser des analyses avancées et rigoureuses, SAS Text Miner permet d'appréhender les tendances à venir et de gérer plus efficacement les nouvelles opportunités en réduisant les risques. Les fonctions linguistiques avancées de SAS Text Miner sont cœur de la solution de data mining SAS® Enterprise Miner™, et offrent ainsi la possibilité d'enrichir les analyses structurées de data mining et d'analyse prédictive avec la connaissance issue des données textuelles.

Pourquoi SAS Text Miner ?

Avec SAS Text Miner, vous gagnez du temps et économisez des ressources en automatisant les tâches de lecture et de compréhension de textes électroniques. En consolidant les sources de données structurées (quantitatives) et les informations textuelles (non structurées) dans un environnement commun, vous obtenez une vue plus précise et plus complète de vos données. L'analyse effectuée sur ces deux types de données produit des modèles descriptifs et prédictifs permettant de repérer précisément les opportunités et les tendances et de prioriser les décisions.

À qui s'adresse SAS Text Miner ?

SAS Text Miner s'adresse principalement aux analystes, statisticiens et data scientists qui doivent passer au crible de gros volumes de texte pour en extraire des informations, des idées et des thèmes communs. La solution s'adresse à tous les secteurs d'activité et contient des procédures d'analyse textuelle haute-performance pour les entreprises ou organisations gérant de très larges collections de documents qui accélèrent l'obtention des résultats.

Principaux atouts

- **Des décisions plus rapides grâce aux processus automatisés.** L'utilisation combinée d'algorithmes intelligents et du traitement du langage naturel limite les tâches manuelles fastidieuses, désormais automatisées telles que l'identification des classes ou la construction des thèmes.
- **Des processus d'exploration et de découverte enrichis par l'expertise métier.** L'identification des concepts clés se faisant par une méthodologie unique, l'interface graphique de SAS Text Miner peut être utilisée pour réévaluer les scores de pertinence et guider les résultats

provenant de l'apprentissage automatique grâce à l'expertise métier. L'utilisation d'entités personnalisées et de méthodes d'apprentissage actives donne la possibilité d'utiliser SAS Text Miner au-delà du mode automatique et de la simple prise en compte de start lists et stop lists.

- **Une vision globale des données avec fonctions d'analyse détaillée.** SAS Text Miner présente l'intégralité du processus de data mining et permet de visualiser et d'explorer le détail des connexions ainsi que les liens entre les éléments d'une série de documents ; une interface interactive facilite l'examen des thèmes dérivés et l'optimisation du modèle.
- **Un repérage des tendances et des opportunités.** SAS Text Miner transforme les données textuelles en une représentation numérique qui résume la collection des documents analysés. Cette représentation apporte une contribution riche à de nombreuses méthodes analytiques. Intégrer ces informations dans des modèles prédictifs permettra de mieux cerner les besoins clients, la demande de produits et services – et d'anticiper les opportunités au moment opportun.

Fig1 : Des règles booléennes peuvent être générées automatiquement et enrichir le modèle directement en supplantant les résultats automatiques.

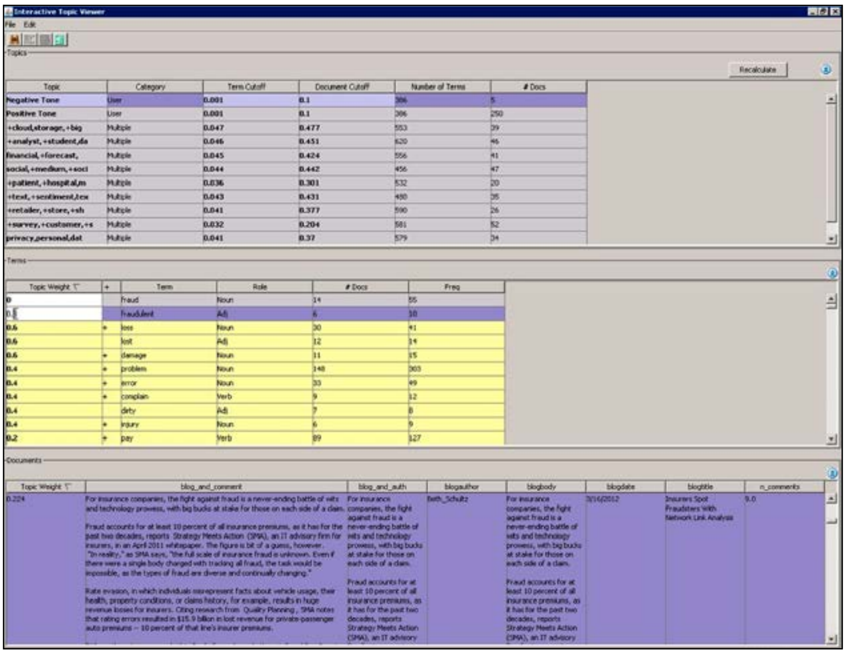

Fig2 : Examen des principaux termes regroupés sous un sujet au travers de l'interface graphique.Si les termes ou les sujets sont similaires, on peut les regrouper et les réassigner pour se rapprocher des résultats escomptés.

Présentation de la solution

SAS Text Miner offre une série complète d'outils de modélisation linguistique et analytique développés spécifiquement pour découvrir et extraire des connaissances à partir de larges collections de documents textuels. A partir d'extraits de textes électroniques, de séries de documents et de téléchargements web, des schémas sont automatiquement identifiés en tant que sujets ou thématiques, définissant des associations explicites entre les termes et les phrases. La solution offre aussi bien des méthodes supervisées que non-supervisées ou semi-supervisées. Une fois les données textuelles transformées en une représentation numérique, elles peuvent être exploitées dans des analyses avancées que ce soit de la modélisation prédictive, du data mining ou de la prévision.

Les procédures haute-performance inclues tirent profit des serveurs multi-processeurs et accélerent grandement les tâches consommatrices de traitement texte.

Des rapports décrivent les résultats issus du nœud « générateur de règle » clarifiant les résultats issus des étapes d'entraînement et de validation des modèles.

Création automatique de règles booléennes et entrainement interactif des modèles

Le nœud « création de règles textuelles» génère automatiquement un ensemble de règles booléennes ordonnées. Ainsi, la solution permet de classifier les documents en s'appuyant sur une table de termes détaillée et permet ensuite de générer la logique booléenne qui a servi à cette classification. Les règles générées peuvent être utilisées pour catégoriser les documents, ou être exportées (y.c. les opérateurs AND, OR, NOT) pour être déployées dans SAS® Enterprise Content Categorization.

Ce nœud permet de mettre en œuvre un apprentissage actif semi-supervisé.

Les utilisateurs peuvent ainsi interagir avec l'algorithme d'apprentissage automatique. La solution découvre automatiquement les catégories et les sujets issus de la collection de documents. Les utilisateurs peuvent aussi prendre la main pour améliorer les résultats du modèle. Combiner l'orientation proposée par la solution et l'expertise métier permet de construire des modèles très pointus.

Traitement accéléré

Les procédures HPTMINE et HPTMSCORE exécutent le traitement des données textuelles en mode multithread. Dédiées aux environnements non distribués, elles optimisent l'utilisation des processeurs multicoeurs et permettent de réduire les temps de traitement de manière drastique dans de nombreux cas.

Filtrage intégré des documents

Des méthodes sophistiquées de réduction du nombre de dimensions permettent de filtrer précisément les documents grâce à des calculs de pondérations, de vérifier l'orthographe et de transformer des données textuelles en un

Principales caractéristiques

La génération automatique de règles booléennes facilite la classification des contenus.

- Aide à décrire et à prédire une variable cible en se basant sur les termes détaillés. Les règles résultantes peuvent servir à catégoriser les documents.
- Les règles peuvent être exportées comme des règles booléennes et utilisées comme jeux de départ dans SAS Enterprise Content Categorization.
- Les résultats de la modélisation sont présentés afin de pouvoir comparer les règles entre les phases d'apprentissage et de validation.
- Apprentissage actif (Active Learning) :
  - suggestions automatiques de catégories et de thèmes proposées par le système, reparamétrables par l'utilisateur,
  - modification de la cible affectée aux règles. Lorsque les règles sont régénérées sur la base d'une telle modification, le modèle est lui aussi mis à jour.

Interface conviviale

- La fusion de plusieurs thématiques en une thématique 'utilisateur' simplifie les résultats similaires.
- L'affichage des sujets permet de visualiser les termes associés à un document, soulignant ainsi son rattachement à un sujet en particulier.
- Le mode visualisation permet de représenter les termes au sein d'un document ou d'un sujet, ou encore de trier les documents.
- La liste de sentiments AFFIN comportant plus de 2000 termes (y.c. des pondérations de polarité pré assignées) permet d'analyser les sentiments au niveau du document.
- Les diagrammes comprenant les flux d'analyse textuelle peuvent être modifiés, sauvegardés et partagés.
- Des tables provenant d'analyses antérieures aux nœuds peuvent être ajoutées pour capitaliser sur la connaissance acquise.
- Chaque nœud peut être personnalisé (et déployé sous forme de score SAS) soit en paramétrant les algorithmes différemment, soit en ajoutant des règles métier pouvant s'appliquer aussi bien à la modélisation prédictive, qu'à la classification automatique, la représentation ou le reporting.
- La solution est conforme aux normes d'accessibilité de la plate-forme Windows.

Filtrage intégré des documents

- Les techniques sophistiquées de réduction des dimensions intègrent le filtrage avancé grâce aux pondérations, à la vérification orthographique automatique et à la transformation des données textuelles en format compact.
- Le nœud filtre propose de créer des tables de synonymes et d'importer des tables de synonymes déjà existantes.

Visualisation des résultats

- La représentation graphique des liens entre les termes permet d'analyser les résultats et d'explorer les liens visuellement.

format compact. Ces méthodes permettent de structurer des données non-structurées et des informations textuelles peuvent ainsi alimenter des modèles prédictifs aussi bien que descriptifs.

L'interface interactive pour l'importation de texte définit des termes pour les données web ou les fichiers texte.

Le nœud d'import de textes crée des tables dynamiquement, que ces textes proviennent d'un répertoire ou qu'ils soient extraits du web. De nombreux formats sont supportés — yc. Microsoft Word® et PDF. Le nœud d'import de texte collecte et convertit les données, filtre ou extrait le texte des fichiers, et enregistre les informations dans une table SAS. Lorsqu'une URL est mentionnée, les pages des sites web associés sont également collectées. La table ainsi générée peut être utilisée par le nœud d'analyse de texte à l'étape suivante de l'analyse. Par ailleurs, le nœud d'import identifie la langue de chaque document et le transforme dans le format d'encodage de la session.

Import facilité

La personnalisation de start list, stop list, liste de synonymes, ou dictionnaires est facilitée. Les options ajout/suppression de table permettent un meilleur contrôle dans l'interface d'import.

High Performance Text Mining

La solution peut s'exécuter en mode multithread sur un mono serveur en utilisant aussi bien les procédures que les nœuds de l'interface. Conçue pour traiter de très larges volumétries de textes, cette capacité va s'étendre à mesure que les collections de documents s'étendront, en distribuant la solution sur d'autres serveurs (license séparée). La haute performance permet de développer des modèles qui analysent des millions et des dizaines de millions de documents tout en exécutant les modèles en quelques minutes ou secondes.

Plus d'informations : **sas.com/hptextmining**

- Des graphiques interactifs permettent de communiquer les résultats aux personnes concernées, notamment les graphiques synthétisant l'information, présentant l'évaluation des thèmes ou les relations entre les termes.
- Le graphique de performance et la documentation générée sur la table de règles aident à naviguer dans les règles booléennes générées.

Traitements haute-performance

- L'analyse textuelle haute-performance, très consommatrice, permet de tirer profit de la puissance machine, réduisant ainsi les temps de traitement de manière conséquente.
- Les données textuelles peuvent être transformées en représentations structurées grâce aux SVDs.
- Les scores sont appliqués plus rapidement sur de très larges tables.

Sélection d'entités prédéfinies, création de nouvelles ou personnalisation pour l'extraction des faits et des événements

- Définition personnalisée de termes composés de plusieurs mots (ex : glisser-lâcher)
- Choix parmi 18 définitions d'entités pré spécifiées pour l'extraction des adresses, entreprises, dates, numéros de téléphone, numéros de cartes d'identité, heures, etc…
- Création d'entités personnalisées à extraire des textes, incluant une liste d'entités prédéfinies (départements, codes produits…) avec l'add-on SAS® Concept Creation for SAS Text Miner.

Interface interactive à l'import des données textuelles internes ou web

- Création dynamique de tables à partir de fichiers internes ou collectés sur le web.
- Nombreux formats supportés : Microsoft Word et PDF, ASCII, HTML, formats Office (feuilles Excel, présentations Powerpoint), emails, base de données…
- Les données textuelles sont extraites, transformées et chargées dans une table SAS pour l'analyse.
- Certains formats propriétaires peuvent être supportés. Dans ce cas le texte est filtré ou extrait et copié dans un fichier plat référencé dans la table SAS.
- La langue de chaque document est identifiée et est traduite dans la langue de la session.

Support natif de multiples langes

- Allemand, anglais, arabe, chinois, coréen, espagnol, français, italien, japonais, néerlandais, polonais, portugais, suédois, danois, finnois, grec, hébreu, hongrois, indonésien, norvégien, roumain, russe, slovaque, tchèque, thaï, turc et vietnamien. Parmi les dialectes figurent le chinois simplifié et traditionnel, le parisien et le québecquois, l'allemand ancien et moderne, les dialectes norvégiens Nunorsk and Bokmal, le brésilien, l'espagnol d'Amérique latine.

THE POWER TO KNOW®

SAS Institute s. a. s. – DOMAINE DE GREGY - GREGY-SUR-YERRES - 77257 BRIE COMTE ROBERT - FRANCE
TEL. : +33(0) 1 60 62 11 11 FAX : +33(0) 1 60 62 11 99 www.sas.com/france

02124FCE0114