20 avantages à connaître sur la plateforme My_Knowledge

SEMDEE
The SmartData Company

Nous abordons dans ce document les aspects différenciants de notre solution par rapport à celles du marché.

Notons que nous abordons les choses dans ce document selon l'angle de l'utilisateur métier et des avantages concrets que « My_Knowledge » est susceptible de procurer aux entreprises et organisations qui l'utilisent.

1. **La proximité un principe inspiré de la cognition humaine**
2. **Approche vectorielle vs Approche dictionnaire**
3. **La prise en compte du contexte**
4. **Rapprochement de contenus de sens similaires exprimés avec des mots différents**
5. **Granularité réglable**
6. **Pas de plan de classements fixes / Découverte et Emergence de thèmes inattendus**
7. **Mise en évidence de relations entre des thèmes / Aide au raisonnement**
8. **Classement automatique par thématiques**
9. **Vision globale et exhaustive de toutes les thématiques**
10. **Thématiques classées par proportions**
11. **Navigation du général au détail**
12. **Recherche en mode « intuitif »**
13. **Classification supervisée : Isoler des contenus à valeur ajoutée dans des catégories**
14. **Exportez et retraitez vos données en interne ou réutilisez les dans d'autres outils**
15. **Algorithmes vs expérience « packagée »**
16. **Un outil simple à la portée de tous**
17. **Démarche intuitive et guidée**
18. **Accédez à une capacité de lecture illimitée**
19. **Mise à disposition immédiate et non intrusive**
20. **Détection et extraction de données personnelles et semi structurées**
21. **Sécurité et confidentialité**
22. **Différence entre une idée et une tendance**
23. **Apprentissage spécifique**
24. **Constance des résultats**
25. **Temps de traitement optimisé**

26. **Pas d'effet « black box »**
27. **Adapté aux gros volumes**
28. **Pas de temps de latence**
29. **Couts d'infrastructure, dépense energétique**

1. La proximité un principe inspiré de la cognition humaine

La technologie reposant sur l'algèbre linéaire permet d'utiliser **la proximité** comme **principe de fonctionnement** général.
Il est par ailleurs intéressant de noter que l'acquisition de connaissances de l'être humain fonctionne sur la base d'assemblage d'informations sur le même sujet. C'est lorsque nous disposons d'un maximum d'informations complémentaires sur le même sujet que nous accédons à la connaissance (savoir que le PSG est le club de foot de la capitale n'est pas suffisant pour prétendre être un « connaisseur » du PSG. En revanche savoir que c'est un club qui est capitalisé par le Qatar, qu'il compte deux stars mondiales, qu'il a gagné les titres de …, que ses sponsors sont…, que son entraineur est… etc.) là nous pouvons commencer à parler de connaissance du PSG.

Dans la pratique nous établissons des relations entre des informations liées les unes avec les autres. Notre intelligence fonctionne notamment comme cela : c'est lorsque nous arrivons à mettre en relation certaines informations que nous avançons dans nos raisonnements (un enquêteur de police cherche à faire des rapprochements, lorsque nous réfléchissons et arrivons à mettre en relation certaines informations dans un contexte donné nous nous exclamons que « nous n'avions pas fait le lien auparavant »).
La manipulation des informations par **la plateforme « My_Knowledge »** correspond donc à la **même démarche cognitive que celle de l'être humain**. Lorsqu'un opérateur utilise la plateforme, il n'y a donc pas de rupture dans la démarche intellectuelle.

Dans tous les cas d'accès à la connaissance ou de raisonnement, pour **produire de la connaissance**, le **préalable** est la **capacité à créer des liens** ou des mises en relations entre différentes informations ou contenus.

Technologiquement l'usage des vecteurs permet de disposer d'une proximité entre tous les contenus d'un Espace cognitif. Cette proximité exprimée en pourcentage qualifie le lien ou la relation entre un ou plusieurs contenus. La technologie de « My_Knowledge » sait donc faire des liens entre les contenus.

2. Approche vectorielle vs Approche dictionnaire

Un point fondamental est que notre technologie **ne fonctionne pas sur la base d'un dictionnaire** mais sur la base d'une vectorisation des mots et des contenus.
Il n'y a rien de révolutionnaire à cela mais il est cependant important de noter que 90% des outils dits « packagés » du marché sont des outils de type sémantiques qui fonctionnent sur des bases de la lemmatisation, de l'analyse syntaxique, lexicale, de retranscription de la grammaire et de contrôle de position des mots au sein d'une phrase, etc.
Pour notre part, **nous ne nous définissons pas comme un outil « sémantique » mais plutôt comme un outil « cognitif ». Il ne s'agit pas là d'une distinction marketing mais d'une réalité technologique.**

En effet, nous n'utilisons pas la grammaire et le vocabulaire pour indexer le texte mais les vecteurs et **donc la proximité entre les contenus**.

Quelques **avantages** techniques résultant de cette approche :

- Pas de temps ou d'étape préalable nécessaire à l'initialisation des dictionnaires
- Pas de nécessité à remettre constamment à jour les dictionnaires avec les nouveaux mots, nouvelles tournures...
- Pas d'approche projet consommatrice de ressources mais une approche orientée « user »
- Fonctionne instantanément
- Supporte directement toutes les langues

3. La prise en compte du contexte

Un autre **avantage** résultant de cette **indexation par proximité** est le rapport au contexte. On sait bien que l'intelligence humaine procède de manière contextualisée. L'usage des proximités et donc des mises en relation de tous les documents entre eux permet ce **rapport au contexte**.

La limite des outils qui gèrent des sens par phrases est ce que l'on entend souvent : « **il s'agit d'une phrase retirée de son contexte** ». Le rapport au contexte permet de « désambiguïser ». L'être humain fonctionne de cette manière aussi : c'est par le contexte que l'on arrive à comprendre le sens général d'un contenu ou d'un texte.

Le rapport au contexte consiste à **être capable d'associer le sens de chaque contenu élémentaire dans un sens plus général**.

4. Rapprochement de contenus de sens similaires exprimés avec des mots différents

Techniquement, cette association de sens se fait encore une fois grâce à **l'utilisation des liens et relations de proximité** existants **entre chaque contenu**. Pour imager cette capacité nous pouvons nous référer à l'utilisation de la plateforme « My_Knowledge » dans le cadre du « grand débat national ».

Dans la thématique « Organisation de l'état » **« My_Knowledge » a su rassembler** au sein du **même cluster** le verbatim **« *Moins de députés et de sénateurs* »** et le verbatim **« *Diminuer le nombre de parlementaires* »**. (Tout comme un opérateur humain l'aurait fait). Ces deux verbatims comme des centaines d'autres au sens proche (mais pas forcément tous rédigés de la même manière) sont rangés dans la même thématique. On constate par ailleurs qu'il n'y a **aucun mot en commun** dans ces deux verbatims qui ont **pour autant été classés dans la même thématique**.

La thématique est introduite de manière synthétique par la plateforme de la façon suivante : « députés, sénateurs, nombre ». Il s'agit bien là de la problématique générique du « nombre de députés et sénateurs ». Il est d'ailleurs très probable que si l'on avait demandé à un opérateur de résumer cette thématique il l'aurait exprimé de la même manière..

Cependant, les Français n'ont pas exprimé qu'une seule idée par commentaire. D'autres idées connexes ou complémentaires ont été également émises (il s'agit de la saisie d'un champ texte libre en bas du formulaire) c'est pourquoi la plateforme a généré une suite de termes (les plus signifiants, il ne s'agit pas de « comptage de mots ») permettant également d'introduire des sens complémentaires (par exemple : « réduire, élus, sénats, assemblée, supprimer, avantages, parlementaires, etc. »). Il s'agit bien la de **créations de sens supplémentaires** issus de l'ensemble des verbatims. A noter que l'on peut demander à la plateforme d'être plus ou moins prolixe (possibilité de définir plus ou moins de termes pour synthétiser le sujet).

5. Granularité réglable 

Le recours à l'algèbre linéaire permet de régler la granularité de chaque fonction pour **manipuler des contenus qui ont un sens plus ou moins proche**. Qu'il s'agisse de clustering non supervisé, de search, de clustering supervisé, de rapprochement ou autres, le recours a des réglages utilisant un pourcentage de proximité permet à l'utilisateur d'augmenter ou de restreindre la sélection selon les fonctions utilisées. Cela permet donc d'avoir des catégories « sur-mesure » selon que l'on veuille des verbatims de sens très proche (précision) ou au contraire dégager de grandes tendances (vision plus globale). Le degré d'analyse sera donc plus pointu et adapté selon le volume de données textuelles..

6. Pas de plan de classements fixes / Découverte et Emergence de thèmes inattendus

Un autre avantage lié à l'utilisation d'une technologie sans dictionnaire consiste à pouvoir effectuer une **classification de type non supervisée**. En d'autres termes quand beaucoup d'outils du marché fonctionnent à partir de « plans de classement fixe » c'est-à-dire de thématiques de classements fixées à l'avance, ils s'interdisent par la même occasion de découvrir certaines thématiques émergentes.

Comme le dit l'analyste d'un grand institut de sondage « Le problème avec cette méthode, c'est qu'on n'est jamais sûr de ne pas passer à côté de l'essentiel. « *En clair, on ne trouvera dans ce gigantesque corpus que ce qu'on y aura cherché, et la méthode choisie ne permettra pas de* ***détecter l'émergence d'un thème inattendu****. »*

Concrètement, **100% des clients** qui ont utilisés « My_Knowledge » ont **découvert des thématiques inattendus ou signaux faibles** en analysant leurs verbatims, créant ainsi instantanément de la valeur pour leur entreprise.

Par exemple, ENGIE a découvert plus de 25 M€ de leads (rassemblés sous la thématique « faire devis chaudière »), VINCI FACILITIES des situations à risques pour les personnes à cause d'une porte qui se ferme avec un aimant trop puissant, la CNAF a mis en évidence une demande jusque-là inconnu de 35.000 personnes qui souhaitaient obtenir un prêt relais pour leur fin de mois,...

Avec « My_Knowledge » **l'utilisateur trouve sans avoir eu à chercher.**

7. Mise en évidence de relations entre des thèmes / Aide au raisonnement

L'utilisation de la proximité entre toutes les thématiques permet également **d'aider l'utilisateur à établir des raisonnements** qui pourraient lui être impossible sur des gros volumes de données sans « My_Knowledge ». Le fait de disposer de **rapports de proximité** entre les différentes thématiques permet d'établir des **relations entre ces thèmes** (on peut par exemple découvrir que le thème « visites multiples chez les clients » est lié et de plus, très proche du thème « pièces détachées »). Ce qui pour un non expert du SAV ne signifie rien mais l'expert métier concerné établira tout de suite le lien de cause à effet.

Plus les contenus sont importants plus il sera non seulement difficile à un humain de les lire tous afin de les classer dans des thématiques communes et de surcroit établir des liens de proximité entre les différentes thématiques. Fort de cette connaissance, l'utilisateur va pouvoir **établir des connexions entre des sujets**, **découvrir des relations de causes à effet**, faire des rapprochements etc. Il s'agit la clairement d'une étape de **création de connaissances** non disponible initialement.

8. Classement automatique par thématiques

Face a un volume de données textuelles qui dépasse l'entendement humain la plateforme « My_Knowledge » va instantanément classer et organiser ces contenus en fonction de leur sens. Il est à noter que ce tri est effectué à partir de critères les plus signifiants déduits par la plateforme elle-même en fonction du contenu présent. Il ne s'agit donc, ni de comptages de mots ni de plans de classement fixe.

L'utilisateur n'a donc pas à chercher ou à supposer ce qu'il est susceptible de trouver dans ces données. **La donnée s'organise d'elle-même**. **Ce n'est plus à l'utilisateur d'aller vers la donnée c'est la donnée qui vient à lui !**

9. Vision globale et exhaustive de toutes les thématiques

Cette approche permet aussi d'appréhender toutes les thématiques présentes au sein du sujet analysé. Accéder au **maximum de facettes d'un sujet** rapproche de la **connaissance exhaustive** sur ce sujet.

10. Thématiques classées par proportion

Le **classement** par thématique ordonnancé **par proportions** permet à l'utilisateur de situer rapidement les tendances les plus importantes et d'évaluer la représentativité d'une tendance particulière par rapport aux autres.

11. Navigation du général au détail

La plateforme My_Knowledge propose à l'être humain une **navigation de type «Top-Down»** comme il a l'habitude de procéder instinctivement (d'un groupe de contenus ayant un sens proche regroupés dans une thématique jusqu'au contenu unique). Cela **facilite la prise de conscience** des principales **thématiques** sur un sujet donné et permet de **disposer rapidement** des liens existants entre les contenus au sein d'une thématique ou tendance.

A noter que l'expérience proposée la plus répandue consiste à présenter à l'utilisateur des contenus élémentaires disjoints sur quelque sujet que ce soit. C'est alors à l'être humain de les mettre en relation, de les hiérarchiser et de les consolider (dans sa tête...) pour construire sa connaissance du sujet (navigation de type « Bottom Up »).

12. Recherche en mode « intuitif »

Parfois nous souhaitons rechercher la présence d'une idée, d'un concept, d'une problématique sans savoir exactement comment elle est rédigée. Dans ce cas il est possible de saisir de manière intuitive quelques mots clés, une phrase, un groupe de

phrases, un paragraphe ou même un texte en entier. « **My_Knowledge » va alors sélectionner les contenus qui ont la plus grande proximité avec les éléments saisis**. Cette fonction permet aussi de vérifier les tendances découvertes lors du classement automatique des données et de décider s'il est intéressant de les organiser d'une manière supervisée.

13. Classification supervisée : Isoler des contenus à valeur ajoutée dans des catégories

Lorsque l'utilisateur est en présence de gros volumes de données texte il peut être intéressé par certaines idées qu'il considère comme ayant une valeur ajoutée mais qui ne sont pas forcément assez présentes pour ressortir dans les thématiques.

Aussi, la plateforme My_Knowledge lui permet de **créer des catégories** et de les nommer (comme un être humain ouvre un classeur ou pose une étiquette sur un tiroir pour classer des documents relatifs au même thème) en signalant à la plateforme le type de contenu qu'il souhaite voir se classer dans cette catégorie à partir de **contenus exemples** (que l'on appelle aussi patterns).

La plateforme va procéder selon une **approche de classification supervisée** (par opposition à la classification non supervisée initiale) et va donc isoler tous les contenus proches (évidemment l'exigence de proximité est réglable) du ou des contenus exemples dans ces catégories prédéfinies. Cette fonction permet de **mesurer la représentativité** de certaines idées et permet également d'isoler des contenus à valeur ajoutée afin de les réutiliser a d'autres endroits de l'entreprise ou dans d'autres sous-système (CRM, Business de Intelligence, ERP, Data-visualisation, etc…).

14. Exportez et retraitez vos données en interne ou réutilisez les dans d'autres outils

A n'importe quel étape de traitement dans la plateforme (Classification non supervisée, Exploration, Search, Contenus reliés, détection de données personnelles, détection de données semi-structurées, classification supervisée) il est possible de **réaliser un export de ces sélections** en interne pour effectuer un nouveau cycle d'affinage ou d'exporter ces données sous différents formats (CSV, JSON, XML) et formatages pour une réutilisation de ces données dans des outils externes (CRM, Business de Intelligence, ERP, Data-visualisation, etc…).

15. Algorithmes vs expérience « packagée »

Il est à noter qu'en dehors des outils packagés plus majoritairement orientés vers la sémantique et utilisant des dictionnaires il peut se trouver sur le marché de l'open

source des algorithmes libres de droits susceptibles de se positionner sur certaines fonctionnalités proposées par My_Knowledge (clustering, search, etc.).

Il faut alors se poser deux questions : la question de la performance (comment se comportent ces algorithmes si on leur demande de traiter 5 millions de contenus ? nécessitent-ils la mise en place de d'infrastructures lourdes de type serveur sparks, cloud, hadoop etc, fonctionnent-ils sur des serveurs standards ?) et sont-ils utilisables de manière autonome par les experts métiers ? (ou bien faut-il sans arrêt un soutien technique ?).

16. Un outil simple à la portée de tous

La valeur dégagée par l'analyse de la donnée textuelle concerne souvent les environnements métiers. De sorte que la nécessité pour les profils métiers de faire intervenir un technicien pour demander des extractions ou des rapprochements ne favorise pas une grande adaptabilité, une souplesse et une grande réactivité.

Avec « My_Knowledge », nul besoin d'être développeur, ingénieur ou technicien.

Après une **prise en main de 1 à 3 jour** un **profil métier est autonome** sur l'utilisation de la plateforme.

17. Démarche intuitive et guidée

Un autre intérêt de l'expérience « My_Knowledge » réside dans la démarche proposée par la plateforme : il s'agit véritablement d'une démarche de type « raffinage de données textuelle » organisée en **5 étapes majeures** :

- Collecte de la donnée
- Organisation des données
- Découverte automatisée
- Exploration
- Réutilisation

Au cours de ce cycle l'utilisateur va véritablement partir de l'ensemble des données, puis il va se voir proposer une vue générale organisée en thématiques. Evidemment il va pouvoir naviguer dans ces thématiques, les explorer pour mieux les comprendre en descendant jusqu'aux contenus élémentaires, manipuler les contenus élémentaires de proche en proche, un par un.

Si cette exploration a fait naitre en lui des questions sur la présence de certains sujets qu'il n'aurait pas vus par hasard dans le découpage automatique par thématique, il va pouvoir utiliser une fonctionnalité de recherche évoluée (recherche intuitive) qui lui permettra de rechercher une idée, un concept ou une problématique sans connaître les mots exacts mais à partir de mots clés, d'une phrase, d'un paragraphe ou d'un texte en entier.

Une fois qu'il aura validé la présence de certaines tendances ou thématiques qu'il considère comme ayant une valeur ajoutée, il pourra les isoler en donnant à la plateforme des contenus en exemple (patterns) et des proximités à respecter par rapport à ces patterns pour ainsi créer des catégories afin d'évaluer la représentativité de ces catégories et éventuellement les réutiliser.

18. Accédez à une capacité de lecture illimitée

Ce ne sont pas les contenus sur quelque sujet que ce soit qui manquent mais c'est bien la capacité à les lire et à les consolider qui fait plutôt défaut, contrainte réglée par **My_Knowledge qui offre à l'être humain une capacité de lecture illimitée**.

19. Mise à disposition immédiate et non intrusive

Pas d'installation, d'intégration, de projet, de charges en ressources humaines à prévoir, la **plateforme** est **instantanément utilisable** par les profils métiers concernés. La plateforme est déconnectée du système d'information donc non intrusive.

20. Détection et extraction de données personnelles et semi structurées

Grace à l'utilisation des REGEX, « My_Knowledge » permet de **détecter et d'extraire les données personnelles** (N° de client, passeport, téléphone, email etc.) ou **les données semi structurées** (N° de pièce détachée, référence de catalogue, etc.).

Cette fonctionnalité a permis au cabinet de conseil PwC de comprendre les demandes de support sur le chat live de **SCHNEIDER LECTRIC**. En isolant les commentaires des clients par référence produit, le consultant a pu reconstituer l'ensemble des sujets d'améliorations classés par ordre d'importance.

21. Sécurité et confidentialité

Grace à l'utilisation des REGEX, « My_Knowledge » permet de **détecter et d'extraire les données personnelles** (N° de client, passeport, téléphone, email etc.) ou **les**

22. Différence entre une idée et une tendance

Grace à l'utilisation des REGEX, « My_Knowledge » permet de **détecter et d'extraire les données personnelles** (N° de client, passeport, téléphone, email etc.) ou **les**

23. Apprentissage spécifique

Grace à l'utilisation des REGEX, « My_Knowledge » permet de **détecter et d'extraire les données personnelles** (N° de client, passeport, téléphone, email etc.) ou **les**

24. Constance des résultats

Grace à l'utilisation des REGEX, « My_Knowledge » permet de **détecter et d'extraire les données personnelles** (N° de client, passeport, téléphone, email etc.) ou **les**

25. Temps de traitement optimisé

Grace à l'utilisation des REGEX, « My_Knowledge » permet de **détecter et d'extraire les données personnelles** (N° de client, passeport, téléphone, email etc.) ou **les**

26. Pas d'effet « black box »

Grace à l'utilisation des REGEX, « My_Knowledge » permet de **détecter et d'extraire les données personnelles** (N° de client, passeport, téléphone, email etc.) ou **les**

27. Adapté aux gros volumes

Grace à l'utilisation des REGEX, « My_Knowledge » permet de **détecter et d'extraire les données personnelles** (N° de client, passeport, téléphone, email etc.) ou **les**

28. Pas de temps de latence

Grace à l'utilisation des REGEX, « My_Knowledge » permet de **détecter et d'extraire les données personnelles** (N° de client, passeport, téléphone, email etc.) ou **les**

29. Couts d'infrastructure, dépense energétique

Grace à l'utilisation des REGEX, « My_Knowledge » permet de **détecter et d'extraire les données personnelles** (N° de client, passeport, téléphone, email etc.) ou **les**