

# *Mining Essential*

Mining Essential sarl  
www.essential-mining.com  
Tél /Fax : 01 46 80 42 71

85, rue Camille Groult  
94400 Vitry sur Seine  
France

CONTACT :  
Pierre-François Doucet  
06 64 38 39 19  
Abderrafih Lehman  
lehman@essential-mining.com  
06 70 33 10 82

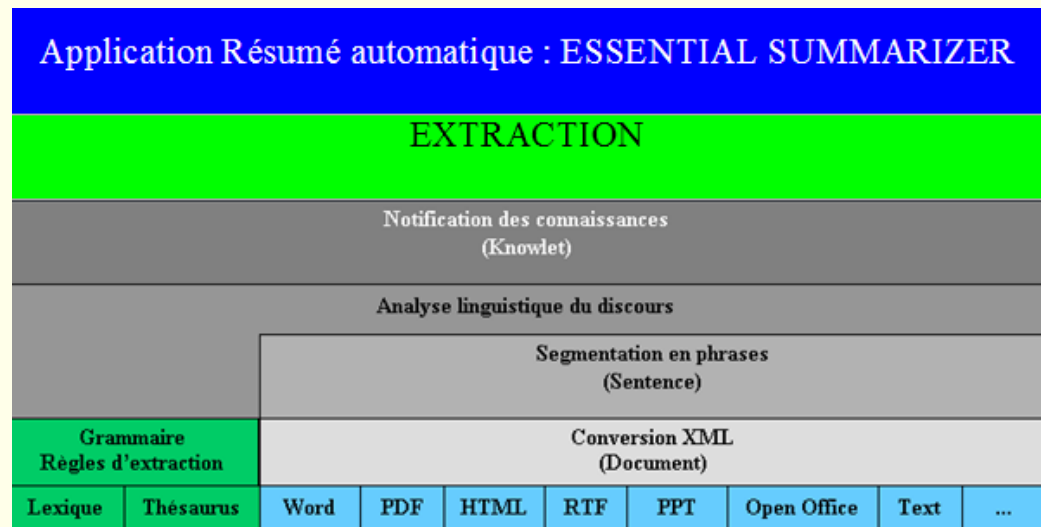
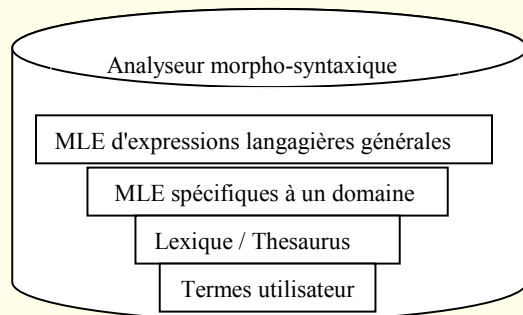
## **Solution de résumé automatique de textes multilingues**

## **Essential Summarizer**

- # La société commercialise la solution Essential Summarizer
  - L'unique application de résumé automatique de textes qui se fonde sur l'extraction intelligente de phrases en multilingue et en multi domaine.
  - Solution totalement flexible et personnalisable à la demande du client.
  - Fruit de 18 années de recherche et développements, elle permet le traitement de 20 langues:
    - français, anglais, espagnol, allemand, néerlandais, portugais, italien, japonais, chinois, coréen, arabe, grec, norvégien, russe, polonais, hébreu, suédois, turc, persan, hindi.

### Principe de fonctionnement

- Le principe est basé sur la seule analyse linguistique du discours en se référant au sens de l'énoncé



- Développée en JAVA/XML l'application utilise une technologie propriétaire de marqueurs linguistiques d'extraction (MLE) qui permettent d'extraire d'un texte les phrases les plus informatives.

### Solution facile à mettre en œuvre

---

- # Choix libre du niveau de contraction du texte source (0-100 %).
- # Possibilité de personnalisation par insertion d'expressions ou mots-clefs propres à l'utilisateur.
- # Spécialisation possible par domaine de traitement

### UNE CONCURRENCE PRATIQUEMENT INEXISTANTE

- ✦ Nécessitant un travail linguistique de longue haleine et un savoir-faire particulier:
  - Les recherches universitaires restent prototypiques et n'aboutissent pas à une commercialisation;
  - Les logiciels de résumé commerciaux ne sont pas nombreux.

### ✦ COMPARAISON CONCURRENTIELLE

Summarizer	Essential	Copernic	Sinope
Plate-forme	MS Windows, Linux, Unix, Solaris	MS Windows	MS Windows
Offre	Monoposte, Serveur, API, Webservice	Monoposte	Monoposte
Formats	.txt, .html, rtf, .doc, . pdf, .ppt, xls	.txt, .html	.txt, .html
Navigateur	Tous les navigateurs	MS IE	MS IE
Langues	18	4	3
Valeur ajoutée	4	1	0

# Mining Essential

## Avancée technologique et commercialisation

### # UNE TECHNOLOGIE EPPROUVEE ET VALIDEE PAR LE MARCHE

- Nous avons montré que la solution est vendable et recherchée - Clients :



- Partenariats :



- Démarches commerciales actuelles : Ministère de la Défense, OEB, IRSN, Secrétariat Général du Gouvernement (SGG), Ministère de l'Intérieur, Dassault, Total, Thalès...

## Poursuite des développements

---

### # DES DEVELOPPEMENTS TOTALEMENT INNOVANTS :

- Compléments à une avancée technologique reconnue et originale :
  - La synthèse automatique de plusieurs documents textuels
  - Le résumé automatique interlangue par traduction automatique

### # OBJECTIF

- Devenir un éditeur de solutions de gestion du résumé automatique reconnues pour leurs performances et leurs qualités fonctionnelles

Ce système multilingue est fondé sur:

- l'analyse linguistique du discours,
- la reconnaissance d'éléments phrastiques ou marqueurs linguistiques d'extraction (MLE) pour évaluer la pertinence de la phrase en vue de sa sélection pour la constitution du résumé,
- une spécialisation par domaine en vue de produire des résumés tenant compte du thème du texte,
- la prise en compte des termes importants pour les besoins d'utilisateur ou d'un groupe d'utilisateurs,
- l'exploitation des observations des utilisateurs permet une évaluation du système et conduire à des améliorations (marqueurs, pondération, compléments aux dictionnaires...).

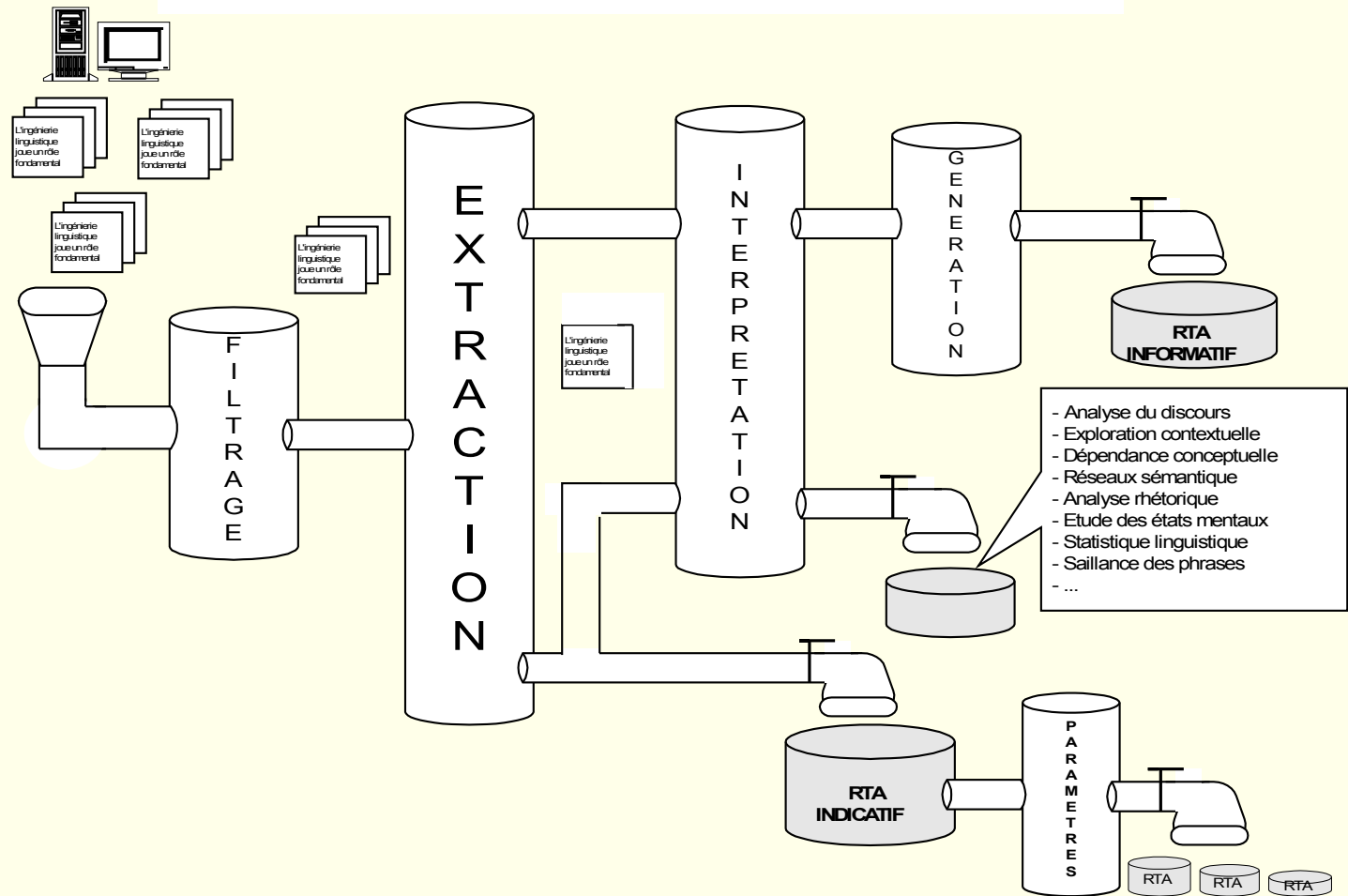


- Automatiser et systématiser l'exploitation de grosses quantités de données textuelles en vue d'un gain de temps significatif
  - Gérer un fonds documentaire spécialisé : Bibliothèque électronique
  - Extraire l'information pertinente liée à un certain domaine
  - Application complémentaire à un moteur de recherche, plate-forme de veille ou toute autre application de GED
  - Recherche et extraction de l'information textuelle mise à disposition par Internet, Intranet, Messagerie , Mobile, Wap, Palm Pilot mobile...
  - etc.

### Les moyens théoriques de reconnaissance :

- la statistique linguistique mots du texte en cooccurrence (Luhn, 1958)
- la combinaison des mots du titre et des sous-titres du texte en relation avec leur présence dans le texte source (Edmundson, 1968)
- l'extraction des seuls paragraphes contenant une concentration des meilleures phrases. Celles-ci sont mesurées par des moyens statistiques (Salton, 94)
- **l'analyse linguistique du discours et de sa structure couplée à la mesure de la pertinence des phrases au moyen de marqueurs linguistiques d'extraction (Lehman, 1995)**
- la composition d'une base de règles d'exploration contextuelle permettant l'extraction d'une liste de phrases (Berri, 1996)
- le calcul de saillance des phrases importantes à partir des éléments obtenus à la suite d'une combinaison de quelques paramètres grammaticaux, syntaxiques et contextuels (Boguraev and Kennedy, 97)
- l'utilisation de certains marqueurs du discours en vue de déterminer les meilleures phrases devant constituer le RA. Ces marqueurs représentent le pivot des relations rhétoriques comme la justification, la cause, la consécution, le contraste, la conséquence, l'antithèse... (Iino et al., 94) pour le japonais, (Marcu, 97) pour l'anglais.

### Le processus global de Résumé de Texte Automatique



- Base de Marqueurs Linguistiques d'Extraction (MLE)
- Base morphologique spécifique
- Base synonymique spécifique
- Spécialisation par domaine - Terminologie
- Personnalisation utilisateur
- Structuration de la base des MLE selon les domaines

## )Base de Marqueurs Linguistiques d'Extraction (MLE)

---

L'approche linguistique repose sur l'analyse du discours d'un domaine identifié et sur le repérage de marqueurs généraux du langage au moyen d'outils .informatiques de fouille de contenu textuel

L'utilisation de lexiques spécialisés du domaine peut améliorer la pertinence du .résumé en terme de thématisation

⋮ Exemples de MLE

- + \* à l'aide de+ \* analyser
- + \* évaluer+ \* qualité
- + \* pour réaliser \* confier
- + \* présenter+ \* en fonction de
- + adapter+ \* pour tenir compte de

La base des indicateurs d'extraction se démultiplie lors du traitement car chacun des mots des MLE est lié à plusieurs dérivations possibles (conjugaison, genre, nombre, nominalisation, synonymie ...), cela grâce à un dictionnaire structuré.

Pris en compte :

Modes : indicatif, conditionnel, participe et infinitif

Temps : tous sauf ceux du subjonctif et de l'impératif

Personnes : 1<sup>ère</sup> et 3<sup>ème</sup> au singulier et au pluriel.

Non pris en compte :

Modes : subjonctif et impératif

Temps : tous ceux du subjonctif et de l'impératif

Personnes : 2<sup>ème</sup> au singulier et au pluriel (tu, vous).

### Conjugaison

<entry> présenter

<word nb="s" pers="1" mode="cond" tps="pres">présenterais</word>

<word nb="s" pers="3" mode="cond" tps="pres">présenterait</word>

<word mode="inf" tps="pres">présenter</word>

<word nb="p" mode="part" tps="pass" gen="f">présentées</word>

<word mode="part" tps="pres">présentant</word>

<word nb="s" mode="part" tps="pass" gen="m">présenté</word>

.....

.....

<word nb="p" pers="3" mode="ind" tps="pass">présentèrent</word>

<word nb="p" pers="3" mode="ind" tps="fut">présenteront</word>

<word nb="p" pers="1" mode="ind" tps="fut">présenterons</word>

<word nb="s" pers="3" mode="ind" tps="fut">présentera</word>

</entry>

### Genre et nombre

<entry> positif

<word nb="s" gen="m"> positif </word>

<word nb="p" gen="m"> positifs </word>

<word nb="s" gen="f"> positive </word>

<word nb="p" gen="f"> positives </word>

</entry>

La base globale est augmentée par le traitement des synonymes adéquats de chacun des mots des MLE mais aussi par la "synonymie " des MLE eux-mêmes. Le but est ici d'éviter au maximum le silence car tous les rédacteurs n'utilisent forcément pas les mêmes expressions ni les mêmes mots pour exprimer une même idée essentielle.

### Synonymie simple

choisir  $\Leftrightarrow$  opter  
adopter  
préférer  
jeter son dévolu sur /contre

convenir  $\Leftrightarrow$  reconnaître  
s'entendre  
se mettre d'accord

### synonymie "composée"

choisir \* travailler  $\Leftrightarrow$  choisir \* étudier  
choisir \* œuvrer  
choisir d'entreprendre une étude

persister \* à l'avantage de  $\Leftrightarrow$  demeurer \* à l'avantage  
dépasser \* limite  $\Leftrightarrow$  franchir \* limite

résulter \* relation  $\Rightarrow$  relever \* lien  
 $\Uparrow$   $\Downarrow$   
résulter \* lien  $\Leftarrow$  relever \* relation



L'utilisation de lexiques terminologiques du domaine améliore la pertinence du résumé en terme de thématisation. Ces bases lexicales spécialisées par domaines (*Chimie, génétique, médecine, philosophie, Informatique, Télécom...*) peuvent servir à spécialiser Pertinence Summarizer pour chacun de ces domaines ou pour d'autres domaines à la demande.

### Médecine

Immunologie  
Immunomodulateur  
Immunosuppresseur  
Inflammation  
Inflammation cancéreuse  
Inflammation gangreneuse  
inflammatoire  
tumeur  
tumeur à malignité atténuée  
tumeur à malignité potentielle  
tumeur bénigne  
tumeur maligne

### Chimie

Acide  
Acide de Brönsted  
Acide carboxylique  
Acide de Lewis  
Alcanes  
Alcènes  
Alcool  
Aldéhyde  
Alkylation  
Alkyle  
Amine  
Arsenic

### Linguistique

consonne  
consonne mi-formée  
constrictive  
contexte  
dévoyellement  
diacritique  
dialectique  
dialogique  
ordre lexicographique  
ordre paradigmatique  
ordre référentiel  
ordre syntagmatique

### Économie / Finance

accompagner \* diversification  
accorder une attention  
adapter \* tenir compte de  
faire connaître \* rôle

favoriser \* processus  
influencer l'évolution  
influencer par la forte baisse  
incidence des opérations  
accélération de la reprise

avoirs en devises  
fonds de stabilisation  
flottement contrôlé  
chiffre d'affaires  
fonds commercial

### Médecine

accompagner \* diversification  
accorder une attention  
adapter \* tenir compte de  
faire connaître \* rôle

Base des MLE  
spécifiques au domaine  
de la MEDECINE

Inflammation  
inflammatoire  
tumeur  
tumeur maligne  
tumeur bénigne

### Chimie

accompagner \* diversification  
accorder une attention  
adapter \* tenir compte de  
faire connaître \* rôle

Base des MLE  
spécifiques au domaine  
de la CHIMIE

Alcool  
Aldéhyde  
Alkylation  
Alkyle  
Arsenic

Cette structuration permet de spécifier un système de résumé par rapport à un domaine donné et/ou par rapport l'attente de l'utilisateur

Base des Marqueurs Linguistiques d'Extraction  
dépendant d'expressions langagières courantes  
(MLE généraux)

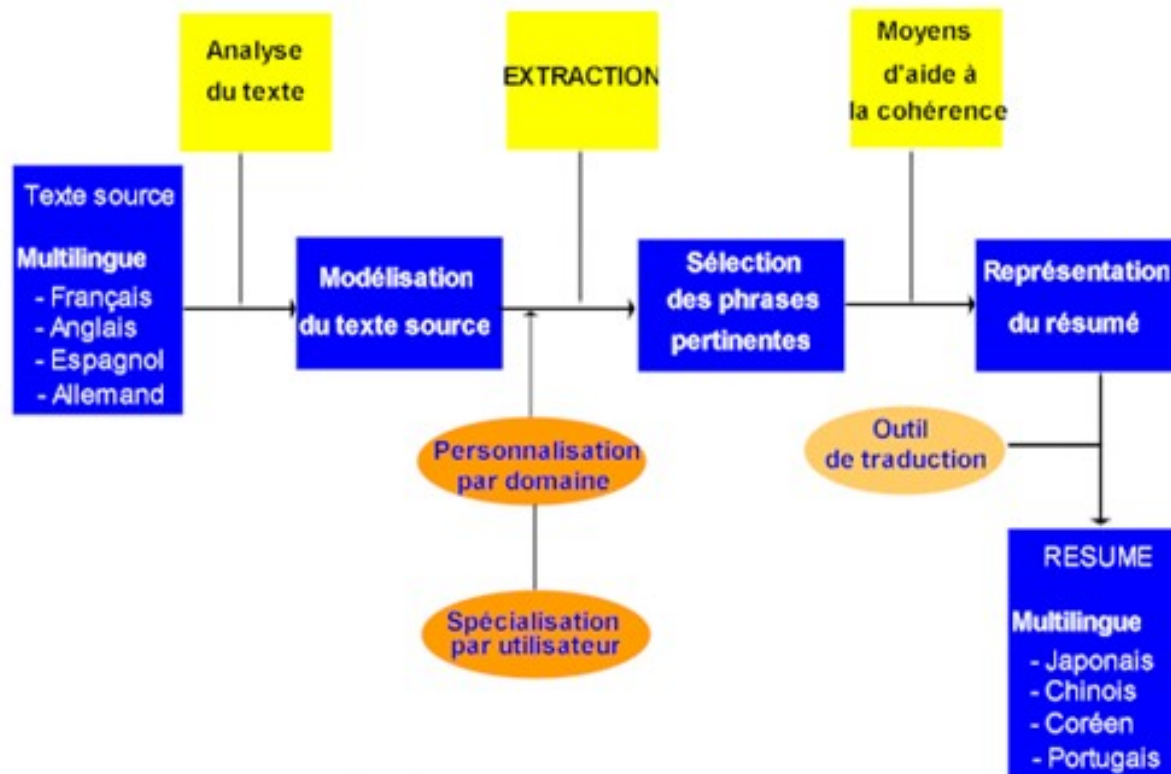
Base des MLE génériques  
d'un domaine particulier

Terminologie  
d'un domaine particulier

Mots ou expressions  
utilisateur

Mots ou expressions  
d'exclusion

## Le traitement multilingue



# *Mining Essential*

---



## Essential Summarizer

Démonstration en ligne sur Internet voir :

<http://www.essential-mining.com>



Interface en 7 langues en mode standard et avancé

Traitement de 20 langues

Traitement de 66 domaines en multilingue

Accueil Produits Solution Services Société Contact

Aide Login

Langue



ESSENTIAL SUMMARIZER

Moteur de résumé automatique de textes multilingues [Solution](#), [Produits](#) et [Services](#)  
Choix de la méthode : [résumé standard](#) ou [résumé avancé](#), en cas de difficultés voir [Aide](#)

Texte Page web Fichier

Langue du texte :

français

Texte source :

5 Tests

Résumer [Résumé avancé](#)

Texte Page web Fichier

Langue du texte : Affichage des phrases :

français

standard

<sup>A</sup> / <sup>A</sup> / Surlignage

Mots-clés de l'utilisateur :

Domaine :

sans

Page Web [http://](#)

Résumer

RÉSUMÉ AUTOMATIQUE STANDARD

RÉSUMÉ AUTOMATIQUE AVANCÉ

© ESSENTIAL SUMMARIZER

CHOISIR un % du résumé **60%** [295mots] ou un pourcentage libre **60.0** %

Récupérer le résumé

Soumettre un autre document

Texte source : [17] mots[493] octets[2651]

temps écoulé : 46ms

Résumé automatique :

PARIS (AP) — Le secrétaire général de la CFDT François Chérèque a estimé dimanche qu'on "pourrait arriver d'ici la fin de l'année" à engager des "choses sérieuses" sur les retraites mais "que c'est ce que le gouvernement ne veut pas justement". "Le courage politique aujourd'hui, c'est de travailler sur toutes les inégalités qui sont engendrées par le système et c'est de ça dont on veut parler", a-t-il résumé. Pour lui, "on pourrait arriver d'ici la fin de l'année ou au début de l'année prochaine pour engager des choses sérieuses", or "c'est ce que le gouvernement ne veut pas justement". Accusant de nouveau le gouvernement de préparer une "petite réforme", il a expliqué que "si on fait une petite réforme, on tape, on punit uniquement les gens sur les durées de cotisations ou l'âge de départ". Pour faire passer "un élément de justice", on ne peut répondre uniquement par la durée de cotisation ou l'âge du départ à la retraite, a-t-il estimé. "Quelles sont les conditions pour allonger la durée de cotisation", "est-ce qu'il faut à terme (...) rapprocher les différents systèmes de retraite par répartition" et "comment on fait pour financer la répartition (...) avec la taxation du capital"? "C'est globalement qu'on jugera la réforme du gouvernement" et "ne me faites pas dire aujourd'hui que j'ai déjà dit non à une réforme". "La réforme, elle sera jugée sur le contenu", a-t-il prévenu, "mais elle sera jugée aussi sur le contexte" des inégalités et de la crise économique.

Exemple de résumé automatique standard



## Exemple de la langue Arabe

### RÉSUMÉ AUTOMATIQUE STANDARD

### RÉSUMÉ AUTOMATIQUE AVANCÉ

© ESSENTIAL SUMMARIZER

summary **40% [136words]** or your own percentage **40.0** %

[Get the summary](#)

[Submit another document](#)

Source text : [11], words[341], bytes[2016] Processing time : 171ms

#### : Text Summarization

واتهم وودارغ ما وصفه بـ"لوبي شركات الصيدلة وتصنيع الأدوية" بخلق حالة من الذعر حيال المرض الذي يسببه فيروس H1N1 ، وقد تجاوزت منظمة الصحة العالمية مع هذا التهويل عبر رفع حالة التأهب إلى مستويات لا تتناسب مع حقيقة انتشار المرض وبحسب وودارغ، فإن منظمة الصحة العالمية استجابت لموجة التهويل هذه بسبب روابط مع مجموعة من الأشخاص في شركات صناعة الأدوية، مضيفاً أن التحقيق الذي وافق الاتحاد الأوروبي على فتحه في الملف بناء على طلبه سيحاول معرفة من الذي قرر تصنيف أنفلونزا الخنازير على أنه وباء ورفع درجة خطره. يذكر أن التحقيق الذي سيطلقه البرلمان الأوروبي سيسعى لمعرفة سبب توصيف المرض بأنه "وباء" من قبل منظمة الصحة منذ يونيو/حزيران 2009 ، بناء على اقتراح من خبراء اصطحهم على صلات بشركات الأدوية الكبيرة، ومنها "نوفارتس" و"روخ" و"غلاسكو سميث".

## Exemple de résumé automatique standard



## INNOVATION ESSENTIAL SUMMARIZER

ESSENTIAL SUMMARIZER propose une nouvelle manière de lire rapidement les documents : les phrases importantes du texte sont affichées en grands caractères, celles de moindre importance en petits caractères. Une gradation entre les deux affichages permet d'attirer visuellement l'attention de l'utilisateur sur les informations essentielles en vue d'une lecture en diagonale aisée. Cette façon de lire les documents peut aussi rendre service aux utilisateurs ayant un problème de vision en leur permettant une lecture confortable.

### Résumé automatique :

Les messages publiés par la sous-commission d'enquête permanente du Sénat suggère que cette grande banque new-yorkaise aurait misé sur une dévaluation des titres dérivés des crédits hypothécaires à haut risque, les fameux "subprimes", alors qu'elle a réaffirmé le contraire samedi. Dans l'un de ces courriers électroniques, le chef financier de la banque, David Viniar, écrit que l'établissement a gagné plus de 50 millions de dollars (37,4 millions d'euros) en un jour en misant sur la chute de l'immobilier, selon un communiqué du bureau du président de la sous-commission. **Nous avons perdu de l'argent, puis nous en avons gagné plus que nous n'en avons perdu grâce aux positions courtes", écrit pour sa part le PDG de Goldman Sachs, Lloyd Blankfein, dans un courriel daté du 18 novembre 2007.** Lorsque la bulle immobilière américaine a explosé, Goldman Sachs et de puissants fonds d'investissement spéculatifs à haut risque (hedge funds) ont pris des positions courtes sur le marché, beaucoup de ces paris reposant sur le fait que d'autres investisseurs misaient sur une hausse. Le marché immobilier s'est effondré et ceux qui avaient pris des positions courtes ont touché un pactole mais se sont fait accuser d'avoir alimenté la crise financière. Ces paris à la baisse sont au coeur des accusations d'escroquerie portées par le gendarme de la bourse américaine contre Goldman Sachs ce mois-ci. La SEC (Securities and Exchange Commission) soupçonne la banque de ne pas avoir dit à deux investisseurs qu'elle leur vendait un produit lié aux crédits subprimes composé avec l'aide du hedge fund new-yorkais Paulson & Co. qui pariait sur la baisse de valeur de ce portefeuille. **De son côté, Goldman Sachs a publié samedi des courriels accablants pour Fabrice Tourré, le courtier français au coeur de l'affaire soulevée par la SEC. Il y évoque notamment les placements à risque visés par l'enquête de la SEC.** "J'ai réussi à vendre quelques titres Abacus à des veuves et des orphelins que j'ai croisés à l'aéroport, apparemment ces Belges adorent" les investissements complexes, écrit le trader en avril 2007, alors qu'il table déjà sur l'effondrement du marché. Il s'adresse dans un mélange d'anglais et de français à une femme avec laquelle il a apparemment noué une relation sentimentale. La sous-commission d'enquête permanente du Sénat, qui enquête sur les causes de la crise financière depuis 18 mois, organise une quatrième et dernière audience mardi, au cours de laquelle elle entendra notamment MM. Blankfein et Tourré. Son porte-parole, Lucas Van Praag, a affirmé samedi dans un communiqué que la banque avait perdu 1,2 milliard de dollars (897.000 euros) dans le marché du crédit hypothécaire en 2007 et 2008. AP





## Fonctionnalités avancées de Text Mining en 20 langues

63 domaines en multilingue déjà intégrés dans Essential Summarizer

Domaines pouvant être personnalisés à la demande en multilingue

© ESSENTIAL S

CHOISIR un % du résumé  ou un pourcentage libre  %

Récupérer le résumé

Soumettre

Texte source : [164], mots[4011], octets[25419] ter

### Terms of the selected field:

[L'environnement](#) (26), [eau](#) (19), [Pollution](#) (13), [air](#) (6), [ozone](#) (6), [pollution de l'air](#) (5), [Pollution atmosphérique](#) (4), [flore](#) (4), [érosion](#) (4), [atmosphère](#) (3), [climat](#) (3), [cours d'eau](#) (3), [faune](#) (3), [gaz à effet de serre](#) (3), [Parc national](#) (2), [dioxyde de carbone](#) (2), [eau douce](#) (2), [environnementaux](#) (2), [gestion des ressources](#) (2), [irrigation](#) (2), [écologie](#) (2), [Antarctique](#) (1), [Arctique](#) (1), [CFC](#) (1), [Pollution de l'eau](#) (1), [Sciences de la terre](#) (1), [agriculture intensive](#) (1), [biodiversité](#) (1), [consommation d'eau](#) (1), [degré de pollution](#) (1), [dégradation de l'environnement](#) (1), [désertification](#) (1), [eau potable](#) (1), [eutrophisation](#) (1), [gestion de l'eau](#) (1), [marée noire](#) (1), [mer](#) (1), [monoculture](#) (1), [polluant](#) (1), [pollution marine](#) (1), [pollution organique](#) (1), [pollution radioactive](#) (1), [pollution thermique](#) (1), [protection de l'environnement](#) (1), [protection du paysage](#) (1), [précipitations](#) (1), [qualité de l'air](#) (1), [ressource naturelle](#) (1), [toxicité](#) (1), [écosystème](#) (1),

### Résumé automatique :

La notion d'environnement englobe aujourd'hui l'étude des milieux naturels, les impacts de l'homme sur l'environnement et les actions engagées pour les réduire. Les premières catastrophes industrielles et écologiques visibles (marées noires, pollution de l'air et des cours d'eau) sensibilisent l'opinion publique et certains décideurs à la protection des écosystèmes. Plus tard, dans les années 1970, les premier et deuxième chocs pétroliers font prendre conscience de l'importance stratégique de la bonne gestion des ressources et des conséquences de la hausse de la consommation matérielle[15]. La découverte et l'exploration de nouveaux milieux (Arctique, Antarctique, monde sous-marin) ont mis en évidence la fragilité de certains écosystèmes et la manière dont les activités humaines les affectent. Parmi les principaux, citons l'observation, puis l'analyse et la synthèse, photographie aérienne, puis satellitaire, et plus récemment, la modélisation prospective. Vers la fin du XXe siècle, la prise de conscience de la nécessité de protéger l'environnement devient mondiale, avec la première conférence des Nations unies sur l'environnement à Stockholm en juin 1972[16]. Dans les pays en voie de développement, où les préoccupations de la population sont très différentes de celles des pays développés, la protection de l'environnement occupe une place beaucoup plus marginale dans la société[18]. Par nécessité, le monitoring (programme de surveillance) environnemental se développe aujourd'hui à échelle planétaire[22], aidée par les avancées techniques, politiques et idéologiques. En 2030, en l'absence de mesures efficaces pour préserver les ressources en eau potable, il pourrait y avoir 3,9 milliards de personnes concernées par le stress hydrique, dont 80 % de la population du BRIC (Brésil, Russie, Inde, Chine). Les cours d'eau ne se limitant généralement pas à un seul état, ils sont devenus des enjeux géopolitiques stratégiques déterminants à la source de nombreux conflits. Parce que l'eau douce est une ressource précieuse, la pollution des nappes phréatiques, qui constituent une réserve importante d'eau douce relativement pure, et des lacs et des rivières, est sans doute la plus préoccupante. Provoquant un réchauffement significatif des cours d'eau concernés, elle peut avoir pour conséquence la disparition locale de certaines espèces animales ou végétales[40]. En l'absence de traitements spécifiques, elles se retrouvent dans les milieux naturels aquatiques, avec des conséquences pour l'environnement et la santé humaine encore mal connues[48].

**Exemple sur le domaine de « l'environnement »**