Natural Language Processing (NLP)

Baptiste, Jamal, Jérémy

Sommaire

- Définition Data mining,Text mining et la différence entre les deux.
- Processus du Text Mining (Étapes)
- Définition d’un document structuré et non structuré
- Définition du NLP
- Exemple d’applications du Text mining
- Représentation Bag Of Words (BOW)
- Explication de la réduction de dimensionnalité avec le retrait des stop words, lemmatization et le stemming.

Définition

Data mining

En règle générale, le terme Data Mining désigne **l'analyse de données depuis différentes perspectives et le fait de transformer ces données en informations utiles, en établissant des relations entre les données ou en repérant des patterns**. Ces informations peuvent ensuite être utilisées par les entreprises pour augmenter un chiffre d'affaires ou pour réduire des coûts.

lebigdata.fr

Text mining

Le text mining, peut être défini comme étant un ensemble de techniques issues de l'intelligence artificielle, alliant plusieurs domaines : **la linguistique, la sémantique, le langage, les statistiques et l'informatique**. Combinées ensemble, ces techniques permettent d'extraire des données pour recréer de l'information à partir de corpus de textes en les classifiant et les analysant de manière à **établir des tendances**.

ia-data-analytics.fr

Data Mining vs Text Mining

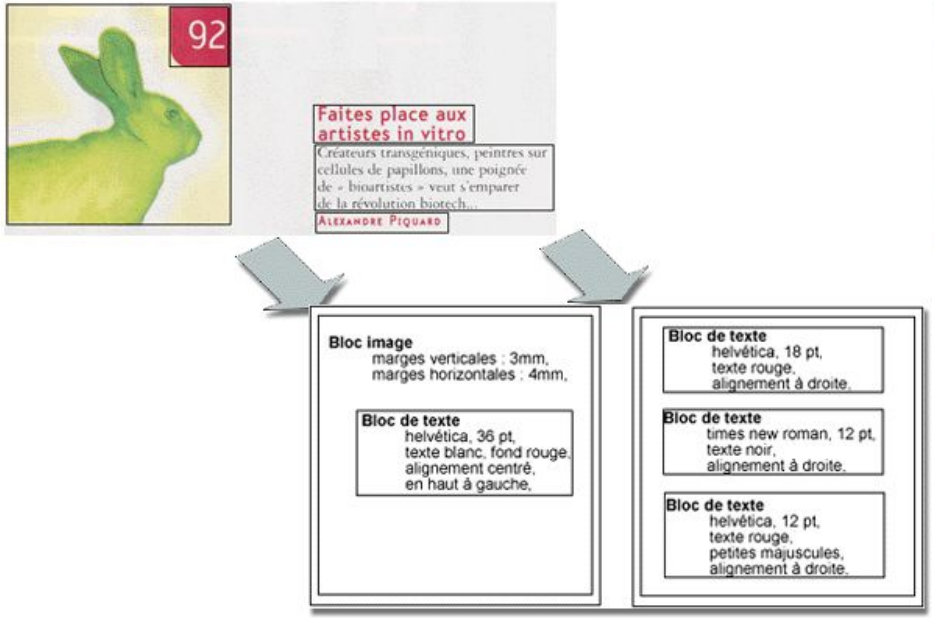
	Data Mining	Text Mining
1.	C’est une technique statistique de traitement des données brutes sous une forme structurée.	C’est une partie du data mining qui implique le traitement de texte à partir de documents.
2.	Des bases de données et des feuilles de calcul préexistantes sont utilisées pour recueillir des informations.	Le texte est utilisé pour recueillir des informations de haute qualité.
3.	Le traitement des données se fait directement.	Le traitement des données se fait de manière linguistique.
4.	Des techniques statiques sont utilisées pour évaluer les données.	Les principes linguistiques informatiques sont utilisés pour évaluer le texte.
5.	Les données sont stockées dans un format structuré.	les données sont stockées dans un format non structuré.
6.	Les données sont homogènes et faciles à récupérer.	Les données sont hétérogènes et ne sont pas si faciles à récupérer.
7.	Il prend en charge l'extraction de données mixtes.	Dans le text mining, l'extraction de texte est uniquement effectuée.
8.	Il combine intelligence artificielle, apprentissage automatique et statistiques et l'applique aux données.	Il applique la reconnaissance de modèles et le traitement du langage naturel aux données non structurées.

Processus du Text Mining (Étapes)

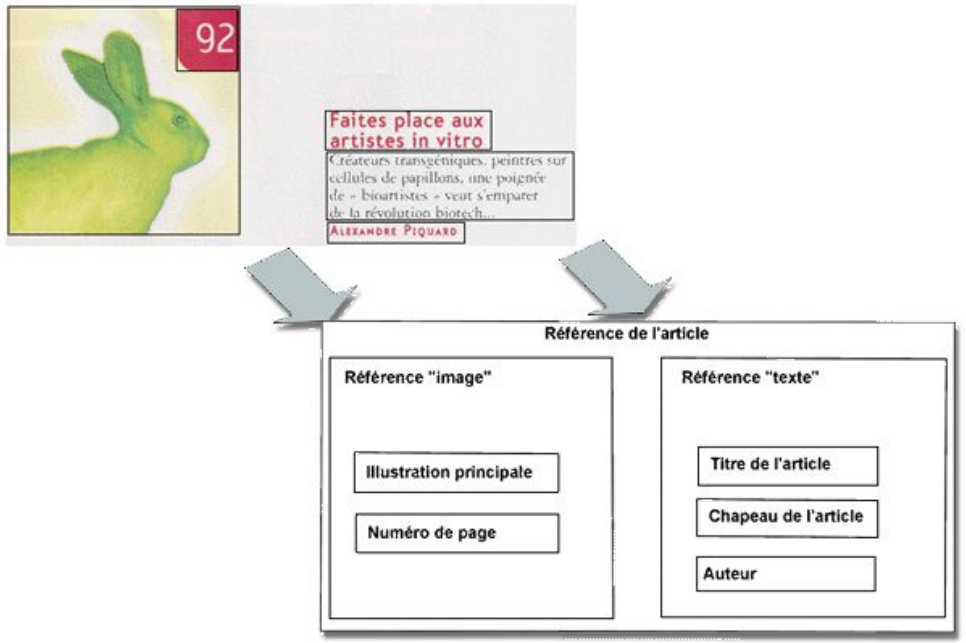
Le texte mining, ou fouille de textes, respecte deux étapes principales:

- La première étape, l'analyse, consiste à **analyser les corpus de textes** de manière à en reconnaître les mots, les phrases, les rôles grammaticaux ainsi que les relations et les sens de ces derniers entre eux.
- La seconde étape : **l'interprétation de l'analyse**. Cette étape permet de sélectionner des textes en particuliers parmi d'autres. Un exemple d'application concret de cette seconde étape étant la classification de courriers mails en spam, c'est-à-dire dans la catégorie des mails non sollicités, ou bien en non spam, c'est-à-dire en mails devant être lus par le destinataire.

Document structuré



structure physique



structure logique

Balisage procédural du texte

```
<FONT COLOR="red" FACE="Helvetica" SIZE="18pt"
ALIGN="left">Faites place aux artistes in vitro </FONT>

<FONT COLOR="black" FACE="Times" SIZE="12pt"
ALIGN="left">Créateurs transgéniques, peintres sur cellules de papillons,
une poignée de "bioartistes" veut s'emparer de la révolution
biotech... </FONT>

<FONT COLOR="red" VARIANT="small-caps" FACE="Helvetica"
SIZE="12pt" ALIGN="left">Alexandre Piquard </FONT>
```

Affichage, par un navigateur, du document balisé

Faites place aux artistes in vitro

Créateurs transgéniques, peintres sur cellules de papillons, une poignée de "bioartistes" veut s'emparer de la révolution biotech...

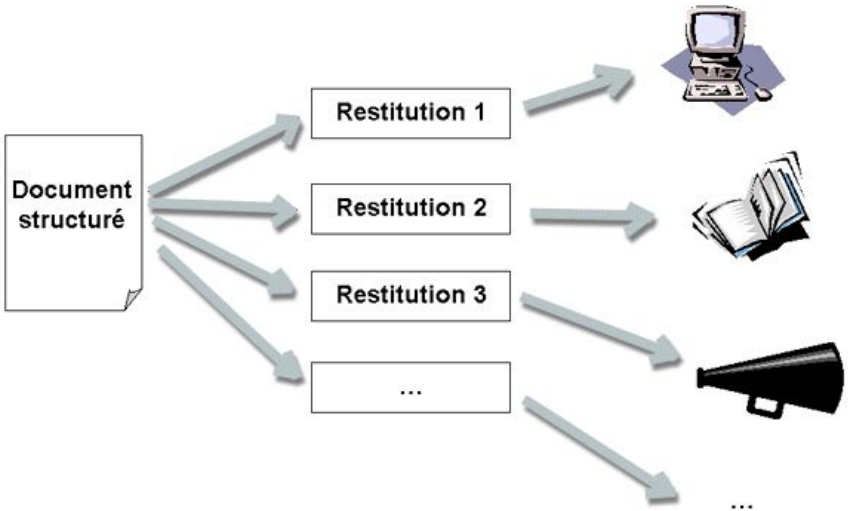
ALEXANDRE PIQUARD

Balisage descriptif du texte

```
<REFERENCE>
    <TITRE>Faites place aux artistes in vitro </TITRE>
    <CHAPEAU>Créateurs transgéniques, peintres sur cellules de
    papillons, une poignée de "bioartistes" veut s'emparer de la
    révolution biotech... </CHAPEAU>
    <AUTEUR>Alexandre Piquard </AUTEUR>
</REFERENCE>
```

Structuration Html : Les balises ne contiennent donc que des informations de nature "typographique".

Structuration SGML / XML : on lui appliquera un marquage qualifié de descriptif. Les balises décrivent le rôle de chaque élément du texte.



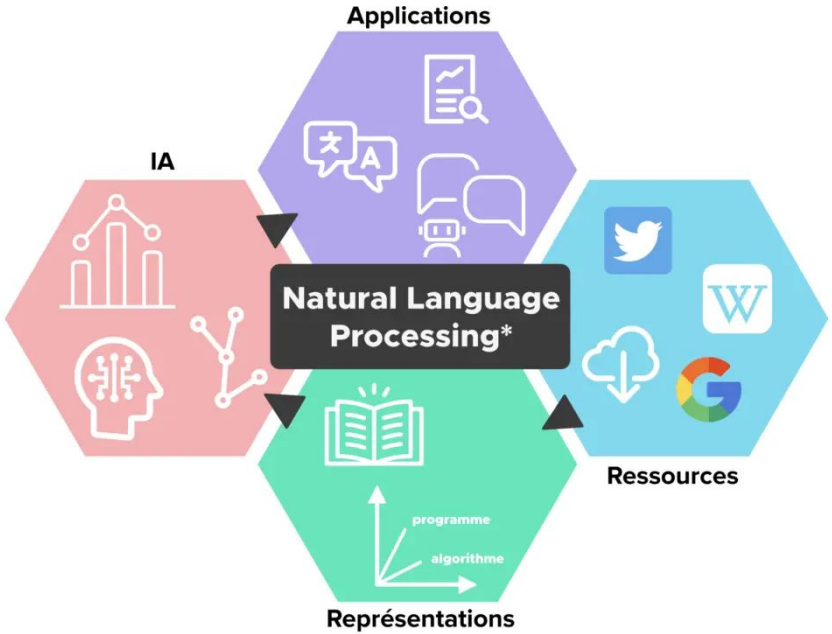
le contenu du document est stocké à part de toute information liée à la restitution. Un affichage écran, une impression papier, voire une restitution orale, ne diffèrent finalement que par la feuille de style utilisée pour la restitution du document qui, lui, reste inchangé.

L'utilisation d'une structure permettra ainsi de mieux cibler *a priori* les interrogations faites sur les corpus de documents.

Définition de la NLP

Traitement du langage naturel PNL est un domaine de l'intelligence artificielle qui donne aux machines la capacité de lire, de comprendre et de tirer du sens des langages humains.

La NLP est présente dans plusieurs champs d'applications comme la Traduction automatique, le Sentiment analysis, le Marketing, les Chatbots, ou encore les Résumé automatique.



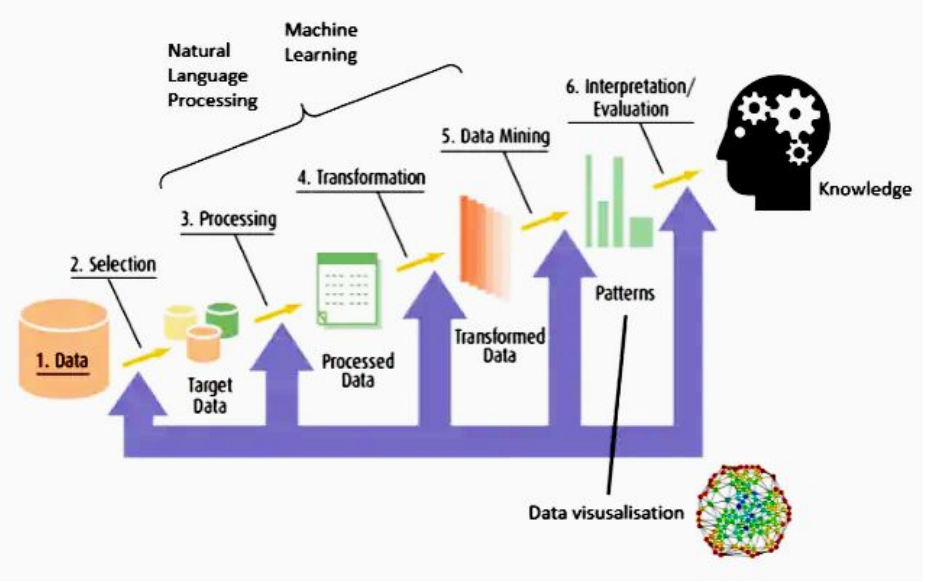
https://datascientest.com/introduction-au-nlp-natural-language-processing

Définition de la NLP

Globalement, deux aspects essentiels à tout problème de NLP peuvent être distingués :

- La partie « linguistique », qui consiste à prétraiter et transformer les informations en entrée en un jeu de données exploitable.

- La partie « apprentissage automatique » qui porte sur l'application de modèles de Machine Learning ou Deep Learning à ce jeu de données.



Exemple d’application du Text Mining

Dans les utilisations courantes nous avons l’**analyse d’emails**, **des sentiments exprimés par les internautes** sur les réseaux sociaux (opinion mining), d’**actions marketings**, la **gestion de la relation client** ou encore l’**amélioration du référencement naturel**.

Le nouveau challenge du text mining, est relatif à **la détection des données sensibles issues des zones de texte libre** présentes dans les applications métier (les CRM ou Gestion de Relation Client par exemple). La détention de ces données sensibles (opinions politiques, origines raciales, convictions religieuses, orientation sexuelle, santé, etc) est interdite par la CNIL. Une problématique fondamentale depuis l’entrée en vigueur du Règlement Général sur la Protection des Données (RGPD), pouvant être en partie résolue par les algorithmes de text mining développés dans ce contexte.

Apprentissage supervisé : l'intérêt de la fouille d'opinion

La fouille d'opinions, en particulier à partir de données des réseaux sociaux, est un excellent substitut nettement moins coûteux des enquêtes d'opinions.

- Evaluation des produits, d'une politique, d'une personnalité
- En appui des systèmes de recommandations (ex. ne pas proposer des produits qui ont des mauvaises notes)
- Analyse de la popularité, des tendances
- Positionnements par rapport à un sujet délicat (ex. mariage pour tous)

Mais les réseaux sociaux permettent d'aller plus loin....

- Susciter des réactions (ex. loi travail)
- Identifier des leaders d'opinions et/ou des spammeurs d'opinions
- Détecter des communautés (ex. Affaire Brown à Ferguson)

Exemple de fouille d’opinions sur les RS

(1) I bought an iPhone a few days ago. (2) It was such a nice phone. (3) The touch screen was really cool. (4) The voice quality was clear too. (5) However, my mother was mad with me as I did not tell her before I bought it. (6) She also thought the phone was too expensive, and wanted me to return it to the shop . . .

(1) relate un fait **objectif**. (2), (3) et (4) expriment une **opinion subjective** plutôt **positive**; (5) et (6) une opinion **négative**.

L’entité iPhone en général est le sujet de (2) ; (3), (4) et (6) sont relatifs respectivement aux **aspects** «touch screen», «voice quality» et «price» de l’iPhone ; «me» est le sujet de (5).

«I» (je) est le **titulaire** des opinions (2), (3) et (4) ; pour (5) et (6), il s’agit de «mother».

Concepts de la fouille d'opinions:

Entité est la cible de l'opinion (objet, produit, évènement, personne…)

Aspect de l'entité ciblée par l'opinion. Parfois, l'opinion peut porter sur l'entité en général (ex. *nice phone*).

Holder (titulaire) de l'opinion c.-à-d. celui qui l'exprime. Important à distinguer lorsqu'ils sont plusieurs et que des phénomènes de communautés peuvent apparaître (coopérant ou s'opposant)

(1) I bought an iPhone a few days ago. (2) It was such a nice phone. (3) The touch screen was really cool. (4) The voice quality was clear too. (5) However, my mother was mad with me as I did not tell her before I bought it. (6) She also thought the (price of the) phone was too expensive, and wanted me to return it to the shop . . .

Orientation de l'opinion, qui peut être de polarité positive ou négative. Une opinion peut être *régulière* (une appréciation) ou *comparative* (en comparaison avec une autre entité). Seul le premier cas nous intéresse ici.

Définition d'une opinion:

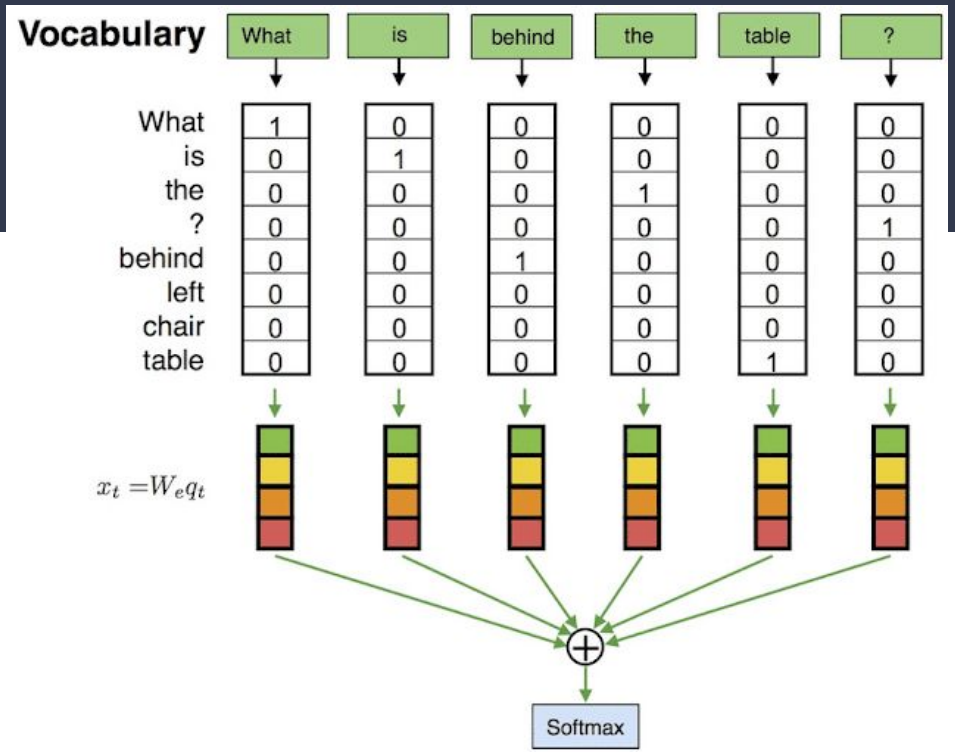
Une opinion est un quintuplet(ei,aij,oijkl,hk,tl), où ei est l'entité, aij est un des aspects de ei, hk est le titulaire, oijkl est son orientation (polarité), et tl est la date (time) où elle a été exprimée. Parce que bien sûr, une opinion peut être fluctuante dans le temps.

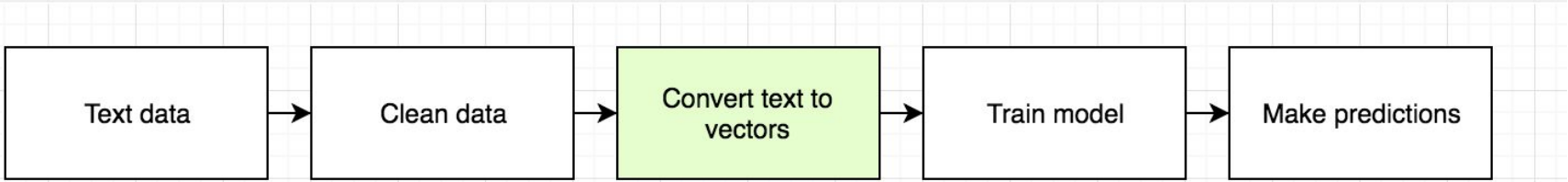
Étapes pour identifier le quintuplet:

1. **Extraction des entités et regroupement**. Identification des entités, regroupement des éventuels synonymes.
2. **Extraction des aspects**. Association avec les entités, regroupement des éventuels synonymes.
3. **Identification du titulaire**, **datation**. Un titulaire exprime une opinion à une date donnée, qui peut être déterminante dans l'analyse.
4. Détermination de l'**orientation** de l'opinion. Elle peut être positive, neutre, ou négative. Parfois, il est nécessaire de faire la distinction au préalable entre une expression objective (fait) et subjective (opinion).
5. Énumération de l'ensemble des tuples(ei,aij,oijkl,hk,tl) dans le corpus suite aux étapes ci-dessus.

Bag of Words

Le modèle de Bag of Words est un moyen de représenter des données textuelles lors de la modélisation de texte avec des algorithmes d'apprentissage automatique. C'est un moyen d'extraire des caractéristiques du texte pour les utiliser dans la modélisation.





https://machinelearningmastery.com/gentle-introduction-bag-words-model/

Explication de la réduction de dimensionnalité

Quelles pistes pour réduire la dimensionnalité (réduire la taille du dictionnaire) ?

- Certains mots ne sont pas intrinsèquement porteurs de sens (ex. most) “stopwords”
- Certains mots sont issus de la même forme canonique (lemmatisation) ou partagent la même racine (stemming) (ex. imaging vs. images)

Les stopwords sont en très grande partie composée de mots qui n’ont pas de sens en eux mêmes, mais qui sont utilisés dans la construction des phrases (ex. prépositions, pronoms, verbes auxiliaires, articles). Ils sont caractéristiques d’une langue et peuvent être utilisés pour les identifier.

Exemple de mot vides : avoir, bon, car, ce, cela, ces, ceux, chaque, ci, comme, comment, dans, des, du, dedans, dehors, depuis, devrait, doit, donc, dos, début, elle, elles, en, encore, essai, est, et, eu, fait, faites, fois, font, hors, ici, il, ils, je, juste, la, le, les, leur, là, ma, maintenant, mais, mes, mine, moins, mon, mot, même, ni, nommés, notre, nous, ou, où, par, parce, pas, peut, peu, plupart, pour, pourquoi, quand, que, quel, quelle, quelles, quels, qui, sa, sans, ses, seulement, si, sien, son, sont, sous, soyez, sujet, sur, ta, tandis, tellement, tels, tes, ton, tous, tout, trop, très, tu, voient, vont, votre, vous, vu, ça, étaient, état, étions, été, être

Exemple de stopwords

Texte original : 139 termes

Ohio Mattress Co said its first nquarter , ending February 28, profits may be below the 2.4 mln ndlrs , or 15 cts a share, earned in the first quarter of fiscal n1986. \n The company said any decline would be due to expenses n\related to the acquisitions in the middle of the current nquarter of seven licensees of Sealy Inc , as well as 82 pct of \nthe outstanding capital stock of Sealy. \n Because of these acquisitions, it said, first quarter sales \nwill be substantially higher than last year's 67.1 mln dlrs. \n Noting that it typically reports first quarter results in \nlate march, said the report is likely to be issued in early \nApril this year. \n It said the delay is due to administrative considerations, \nincluding conducting appraisals, in connection with the \nacquisitions \n Reuter

Texte nettoyé 125 termes :
Après : harmonisation de la casse, retrait des ponctuations et de \n, retrait des nombres, retrait des espaces en trop

ohio mattress co said its first quarter ending february profits may be below the mln dlrs or cts a share earned in the first quarter of fiscal the company said any decline would be due to expenses related to the acquisitions in the middle of the current quarter of seven licensees of sealy inc as well as pct of the outstanding capital stock of sealy because of these acquisitions it said first quarter sales will be substantially higher than last years mln dlrs noting that it typically reports first quarter results in late march said the report is likely to be issued in early april this year it said the delay is due to administrative considerations including conducting appraisals in connection with the acquisitions reuter

Texte après retrait des stopwords en anglais : 76 termes

ohio mattress co said first quarter ending february profits may mln dlrs cts share earned first quarter fiscal company said decline due expenses related acquisitions middle current quarter seven licensees sealy inc well pct outstanding capital stock sealy acquisitions said first quarter sales will substantially higher last years mln dlrs noting typically reports first quarter results late march said report likely issued early april year said delay due administrative considerations including conducting appraisals connection acquisitions reuter

$\frac{125-76}{125} = 39.2\%$ des termes ont été retirés.

Lemmatisation

La lemmatisation consiste à analyser les termes de manière à identifier sa forme canonique (lemme), qui existe réellement . L'idée est de réduire les différentes formes (pluriel, féminin, conjugaison, etc.) en une seule.

Ex. am, are, is → be

car, cars, car's, cars' → car

Ainsi, la phrase « the boy's cars are different colors » devient « the boy car be differ color ».

La technique fait à la fois référence à un dictionnaire, et à l'analyse morphosyntaxique des mots (ex. Weiss et al., 2005 ; page 24). Elle est spécifique à chaque langue. Des erreurs sont toujours possibles !

Exemple : Mignonne , pourquoi es tu partie si loin de nos avenirs radieux, ceux où nous devions regarder le soleil se lever au milieu des chants des boeufs?

Le mot mignon regroupe : mignon, mignonne

Le mot etre regroupe : etre, es

Le mot taire regroupe : taire, tu

Le mot partir regroupe : partir, partie

Le mot notre regroupe : notre, nos

Le mot celui regroupe : celui, ceux

Le mot devoir regroupe : devoir, devions

Le mot du regroupe : du, des

Le mot chant regroupe : chant, chants

Le mot boeuf regroupe : boeuf, boeufs

Stemming

Le stemming consiste à réduire un mot à sa racine (stem), qui peut ne pas exister!

L'algorithme de Porter est un des plus connus pour la langue anglaise. Il applique une succession de règles (mécaniques) pour réduire la longueur des mots c.à.d. supprimer la fin des mots.

Page de Martin Porter

Avant stemming : 549 caractères

ohio mattress co said first quarter ending february profits may mln dlrs cts share earned first quarter fiscal company said decline due expenses related acquisitions middle current quarter seven licensees sealy inc well pct outstanding capital stock sealy acquisitions said first quarter sales will substantially higher last years mln dlrs noting typically reports first quarter results late march said report likely issued early april year said delay due administrative considerations including conducting appraisals connection acquisitions reuter

Après stemming : 477 caractères

ohio mattress co said first quarter end februari profit may mln dlrs cts share earn first quarter fiscal compani said declin due expens relat acquisit middl current quarter seven license seali inc well pct outstand capit stock seali acquisit said first quarter sale will substanti higher last year mln dlrs note typic report first quarter result late march said report like issu earli april year said delay due administr consider includ conduct apprais connect acquisit reuter