



## *Université Sidi Mohamed Ben Abdellah* *Faculté des Sciences Dhar Mahraz*

### Fouille de Texte (Text Mining)

**Chakir LOQMAN**

27/09/2023

# Plan

## 1 Introduction

## 2 Notions de base

- Processus de Text Mining
- Représentation des textes
- Pondération des termes
- Similarité entre documents
- Réduction des dimensions

## Objectifs et Organisation du cours

### Objectifs

- Acquérir une connaissance approfondie de certaines techniques considérées comme des méthodes de base en Text Mining
- Introduire les techniques d'apprentissage supervisé et non supervisé utilisées le plus couramment en fouille de textes.
- Savoir appliquer la fouille de textes dans :
  - Recherche d'information
  - Résumé de documents

### Matériel nécessaire

- Un poste informatique sous Windows.
- Installez eclipse.



### Organisation du cours

- 25 **Cours+TD+TP**

## Le Text Mining ?

- La quantité d'information non structurée double tous les deux mois dans les grandes entreprises.
- Aujourd'hui, les entreprises ayant mis en place un système de gestion des données non structurées ont accru leur productivité de 15 % en moyenne.
- L'employé qualifié moyen passe 2h30 / jour à rechercher des documents.
- Documents électroniques :
  - Structurés (10%) et non-structurés (90%)
  - Grand volume, croissance exponentielle

### Problèmes

- Recherche d'information
- Correction orthographique/grammaticale
- Filtrage/classification d'information : courrier électronique, flux d'actualité, document métier...
- Traduction automatique et Résumé automatique

## Fouille de textes (Text mining)

### Définition

- Le text mining est une technique permettant d'automatiser le traitement de gros volumes de contenus textuels.
- C'est une spécialisation de la fouille de données et fait partie du domaine de l'intelligence artificielle

### Domaines d'application

- Recherche d'information
- Applications biomédicales
- Filtrage des communications
- Applications de sécurité
- Intelligence économique
- Marketing
- ....

# Positionnement IR/IE

## Recherche d'informations (IR)

- Ensemble des méthodes, procédures et techniques permettant, en fonction de critères de recherche propres à l'utilisateur, de sélectionner l'information dans un ou plusieurs fonds de documents plus ou moins structurés
- Exemple :
  - Trouver les documents qui traitent la réussite scolaire

## Extraction d'informations (IE)

- L'extraction d'information (EI) est une technologie visant à reconnaître dans un corpus de documents textuels un ensemble d'informations spécifiques, à les extraire et à les structurer dans un format prédéfini.
- Exemple :
  - Etablir une base de données où l'on peut retrouver les noms des entreprises informatiques cédées en 2003

# Pièges et difficultés du langage naturel

## Caractéristiques du langage naturel

- L'implicite
  - La pragmatique : Liée au contexte du message, aux connaissances sur le monde, à l'usage
    - Ex : Il donna le billet à la jeune femme
    - spectacle → billet d'entrée      transaction commerciale → billet de banque
- La redondance
  - La synonymie : Mots ou expressions différents ayant le même sens.
    - Ex : voiture et automobile; train et chemin de fer...
  - La paraphrase : Expressions équivalentes mais de structure ou de termes différents
    - Ex : Mon fils a cessé de fumer ⇔ Jean a renoncé au tabac
- L'ambiguïté
  - L'homonymie : Mots ayant la même forme, la même graphie mais des sens différents.
    - Ex : Je porte la porte
  - La polysémie : Mots ou expressions ayant plusieurs sens
    - Ex : Mémoire humaine, mémoire d'ordinateur
  - L'homotaxie : Une même syntaxe recouvrant des réalités différentes
    - Ex : Jean est facile à convaincre ≠ Jean est habile à convaincre

# Morphologie

## Définition

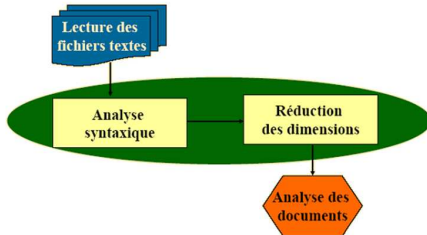
- La **morphologie** est la branche de la **linguistique** qui étudie les types et la forme des **mots** en interne ou en externe.
- L'étude des mots en **interne** rend compte des relations qui existent entre différentes formes d'un même mot
  - Ex : Toutes les formes d'un même verbe entretiennent mutuellement un certain nombre de relations
- L'étude des mots en **externe** rend compte des relations qui existent entre différents mots du lexique
  - Ex : Une étude rapide de certains mots contenant le suffixe -eur met en évidence différents sens à attribuer à ce morphème. Le suffixe -eur peut donc revêtir plusieurs sens différents

# Processus de Text Mining

## Vue simplifiée

Au moins quatre grands niveaux d'analyse linguistique du texte intégral :

- **niveau morpho-lexical** : reconnaissance du mot ;
- **niveau syntaxique** : étiqueter les séquences de mots (niveau d'utilisation de la grammaire)
- **niveau sémantique** : niveau de la reconnaissance des concepts
- **niveau pragmatique** : niveau contextualiser



# Lemmatisation

## Définition

**La lemmatisation** désigne l'analyse lexicale du contenu d'un texte regroupant les mots d'une même famille. Chacun des mots d'un contenu se trouve ainsi réduit en une entité appelée lemme (**forme canonique**).

- La lemmatisation regroupe les différentes formes que peut revêtir un mot
- La lemmatisation d'une forme d'un mot consiste à en prendre sa forme canonique. Celle-ci est définie comme suit :
  - Pour un **verbe** : ce verbe à **l'infinitif**,
  - pour les autres **mots** : le mot au **masculin singulier**.

## Exemples

- L'adjectif petit existe sous quatre formes : petit, petite, petits et petites. La forme canonique de tous ces mots est petit.
- Il existe beaucoup plus de formes du verbe avoir : ai, as, a, avons, ais, avons eu, ayez eu, eussions eu, aurions eu, etc. La forme canonique de eussions eu est avoir.

# Racinisation (Stemming)

## Définition

La **racinisation** ou désuffixation (anglais : **stemming**) est un procédé de transformation des flexions en leur radical ou racine (anglais : stem)

- La racine d'un mot correspond à la partie du mot restante une fois que l'on a supprimé son (ses) préfixe(s) et suffixe(s)
- La racinisation est un procédé fréquent dans les applications de traitement automatique du langage naturel :
  - Traduction automatique
  - Recherche d'information
  - Indexation des moteurs de recherche

## Exemples

- Les mots continu, continua, continuait, continuant sont tous ramenés au stem **continu**
- cheval, chevaux, chevalier, chevalerie, chevaucher  $\implies$  cheva

# Racinisation (Stemming)

## Les différents algorithmes

Ces divers algorithmes de racinisation procèdent en deux étapes : un pas de désuffixation qui consiste à ôter aux mots des terminaisons prédéfinies les plus longues possibles, et un pas de recodage qui ajoute aux racines obtenues des terminaisons prédéfinies.

- Deux principales familles de stemmers sont présentes dans la littérature : les stemmers algorithmiques et ceux utilisant un dictionnaire :
  - **Un stemmer algorithmique** va être souvent plus rapide et va permettre d'extraire des racines de mots inconnus. Il va cependant avoir un taux d'erreur plus élevé.
  - **L'approche par dictionnaire** quant à elle ne fait pas d'erreur sur les mots connus, mais en produit sur ceux qu'elle ne liste pas. Elle est aussi plus lente.

## Exemples

- Algorithme de Lovins
- Algorithme de Porter
- Algorithme de Carry (français)

## Représentation en sac de mots (Bag Of Words)

### Principe général

- L'idée est de transformer les textes en vecteurs dont chaque composante représente un terme. Ces termes sont les mots qui constituent un texte, et ils possèdent un sens explicite.
- Méthode fondée sur :
  - **l'élimination des mots-vides** (articles, prépositions, mots grammaticaux...), à partir d'un dictionnaire de termes (appelé **stop list**)
  - **la constitution d'un index des termes non éliminés**, considérés comme des chaînes de caractères.

### Exemple du texte :

" L'histoire n'a pas seulement légué ses monuments à la Bretagne. Elle lui a aussi donné ses paysages ruraux, lentement façonnés par des générations de paysans anonymes."

anonymes	Bretagne	donné	façonnés	générations	histoire
légué	lentement	monuments	paysages	paysans	ruraux

## Représentation en sac de mots (Bag Of Words)

### Stop Words (de liaison)

- Liste de mots (ex. ceux listés par Oracle text) sont les 200 suivants :
  - a , beaucoup, comment, encore, lequel, moyennant, près, ses, toujours, afin, ça, concernant, entre, les, ne, puis, sien, tous, ailleurs, ce, dans, et, lesquelles, ni, puisque, sienne, toute, ainsi, ceci, de, étaient, lesquels, non, quand, siennes, toutes, alors, cela, dedans, était, leur, nos, quant, siens, très, après, celle, dehors, étant, leurs, notamment, que, soi, trop, attendant, celles, déjà, etc, lors, notre, quel, soi-même, tu, au, celui, delà, eux, lorsque, notres, quelle, soit, un, aucun, cependant, depuis, furent, lui, nôtre, quelqu'un, sont, une, aucune, certain, des, grâce, ma, nôtres, quelqu'une, suis, vos, au-dessous, certaine, desquelles, hormis, mais, nous, quelque, sur, votre, au-dessus, certaines, desquels, hors, malgré, nulle, quelques-unes, ta, vôtre, auprès, certains, dessus, ici, me, nulles, quelques-uns, tandis, vôtres, auquel, ces, dès, il, même, on, quels, tant, vous, aussi, cet, donc, ils, mêmes, ou, qui, te, vu, aussitôt, cette, donné, jadis, mes, où, quiconque, telle, y, autant, ceux, dont, je, mien, par, quoi, telles, autour, chacun, du, jusqu, mienne, parce, quoique, tes, aux, chacune, duquel, jusque, miennes, parmi, sa, tienne, auxquelles, chaque, durant, la, miens, plus, sans, tiennes, auxquels, chez, elle, laquelle, moins, plusieurs, sauf, tiens, avec, combien, elles, là, moment, pour, se, toi, à, comme, en, le, mon, pourquoi, selon, ton.

## Représentation en sac de mots (Bag Of Words)

يعتبر من أبرز عباقرة الموسيقى في جميع العصور، و أبدع اصعالا موسيقية خالدة. له الفضل الاعظم في تطوير الموسيقى الكلاسيكية

يعتبر	1	الفضل	1	اصعالا	1	الموسيقى	2
ابرز	1	الاعظم	1	موسيقية	1	العصور	1
عباقرة	1	ابدع	1	خالدة	1		
الأوروبية	1						
تطوير	1						
الكلاسيكية	1						

Figure – Représentation du texte Arab par un vecteur de mots

### Inconvénients :

- tous les mots non vides mis sur le même plan :
  - pas de prise en compte de l'ordre des mots
  - apparition des différentes formes d'un mot : par ex. un verbe va apparaître plusieurs fois sous des formes différentes
- l'analyse porte seulement sur des mots isolés (des unitermes), et délaisse toutes les expressions (les syntagmes), souvent porteurs de sens

# Représentation des textes avec les racines lexicales

## Principe général

- Contrairement au modèle précédent (Représentation en sac de mots), chaque flexion est considérée comme descripteur différent et donc une dimension de plus. La représentation lexicale cherche à résoudre cette difficulté en considérant uniquement la racine des mots plutôt que les mots entiers
- Plusieurs algorithmes ont été proposés pour substituer les mots par leur racine :
  - Algorithme le plus connu pour la langue anglaise : **Porter stemmer**
  - Algorithme le plus connu pour la langue arabe : **Khoja**

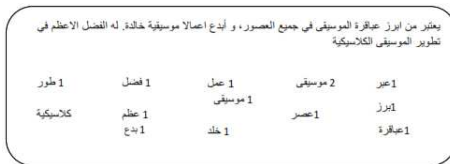


Figure – Représentation du texte Arab par un vecteur de racines

# Représentation des textes avec les lemmes

## Principe général

- **Réduction des mots à leur forme canonique** : toutes les formes d'un verbe par exemple sont regroupées à l'infinitif ; tous les mots au pluriel sont ramenés au singulier...
- Plusieurs algorithmes ont été proposés pour substituer les mots par leur forme canonique :
  - Algorithme le plus connu pour la langue anglaise, française, allemande et italienne : **TreeTagger**
  - Pour la langue arabe reste toujours un sujet de recherche

## Exemple du texte :

" L'histoire n'a pas seulement légué ses monuments à la Bretagne. Elle lui a aussi donné ses paysages ruraux, lentement façonnés par des générations de paysans anonymes."

anonyme	Bretagne	donner	façonner	génération	histoire
léguer	lentement	monument	paysage	paysan	rural

## Représentation des textes avec les lemmes

يعتبر من أبرز عباقرة الموسيقى في جميع العصور، و أبدع اعمالا موسيقية خالدة. له الفضل الاعظم في تطوير الموسيقى الكلاسيكية				
1 عب	2 موسيقى	1 عمل	1 فضل	1 طور
1 يبرز	1 عصر	1 موسيقى	1 عظم	كلاسيكي
1 عقري		1 خلد	1 بدع	

Figure – Représentation du texte Arab par un vecteur de lemmes

### Inconvénients :

- La grande dimension de l'espace de représentation : Les documents sont représentés par des vecteurs de dimension égale à la taille du vocabulaire, qui est en général assez grande :
  - rend la plupart des algorithmes de classification difficiles à utiliser
- La perte d'information lors de la construction de la représentation de textes, à cause de l'ignorance de toute relation entre les termes.

# Représentation conceptuelle

## Principe général

- Elle se base sur le formalisme vectoriel pour représenter les documents. Les éléments du vecteur ne sont plus associés directement à des termes d'indexation mais plutôt à des concepts.
- Pour permettre une telle représentation des documents, il est nécessaire de pouvoir projeter les termes dans une ressource sémantique :
  - WordNet, Arabic WordNet
  - Cet outil, ayant pour objectif de représenter des aspects sémantiques d'un lexique.

## Exemple du texte :

**" L'histoire n'a pas seulement légué ses monuments à la Bretagne. Elle lui a aussi donné ses paysages ruraux, lentement façonnés par des générations de paysans anonymes."**

anonyme	Bretagne	donner	Façonnage de paysage	génération
histoire	Legs de monument	lentement	paysan	

# Pondération des termes

## Pondération des termes

- La pondération de termes a pour but de déterminer de manière quantitative la représentativité d'un terme.
- Pour permettre une meilleure catégorisation, il faut refléter ce degré d'importance dans un algorithme, en attribuant à chaque terme un poids.
- Il existe différentes méthodes pour calculer ce poids. Ces méthodes sont basées sur les observations suivantes :
  - Plus le terme  $t$  est fréquent dans un document  $d$ , plus il est en rapport avec le sujet de ce document.
  - Plus le terme est fréquent dans une collection, moins il sera utilisé comme discriminant entre documents.

## TF-IDF (Term Frequency-Inverse Document Frequency)

**TF-IDF** est une méthode de pondération souvent utilisée en recherche d'information et en particulier dans la fouille de textes. Cette mesure statistique permet d'évaluer l'importance d'un terme contenu dans un document, relativement à une collection ou un corpus.

## Fréquence du terme TF

### Définition

Un terme qui apparaît plusieurs fois dans un document est plus important qu'un terme qui apparaît une seule fois

$$TF_{ij} = \frac{w_{ij}}{|d_j|}$$

- $w_{ij}$  : Nombre d'occurrences du terme  $t_i$  dans le document  $d_j$
- $|d_j|$  : Nombre total de termes dans le document  $d_j$

$TF_{ij}$  = Fréquence du terme  $t_i$  dans le document  $d_j$

### Remarque

Plus le terme est fréquent dans un document, plus il est important dans la description de ce document. C'est le nombre d'occurrences de ce terme dans le document considéré.

## Fréquence inverse de document IDF

### Définition

Un terme qui apparaît dans peu de documents est un meilleur discriminant qu'un terme qui apparaît dans tous les documents

$$IDF_i = \log\left(\frac{|D|}{|\{d_j : t_i \in d_j\}|}\right)$$

- $|D|$  : Nombre total de documents dans le corpus
- $|\{d_j : t_i \in d_j\}|$  : Nombre de documents où le terme  $t_i$  apparaît

$IDF_i$  = Inverse document frequency

### Remarque

$IDF_i$  mesure en quelque sorte la rareté du terme  $t_i$  dans la collection.

## Mesure TF-IDF

### Définition

La mesure  $TF - IDF$  sert à mesurer l'importance d'un terme dans toute la collection. Alors le poids d'un terme  $t_i$  dans un document  $d_j$  est :

$$w_{ij} = TF_{ij} \times IDF_i$$

- $TF_{ij}$  = Fréquence du terme  $t_i$  dans le document  $d_j$
- $IDF_i$  = Inverse document frequency

### Remarque

Mesure  $TF - IDF$  prend en compte la fréquence locale d'un terme, c'est-à-dire relative à un document (term frequency, TF) et sa fréquence globale, relative à un corpus (inverse document frequency, IDF).

## Exemple

### Exemple

L'exemple porte sur les trois documents ( $d_1, d_2, d_3$ ) et le terme analysé est « qui » (soit  $t_1 = \text{qui}$ ). La ponctuation et l'apostrophe sont ignorées.

- Document 1 :
  - Son nom est célébré par le bocage **qui** frémit, et par le ruisseau **qui** murmure, les vents l'emportent jusqu'à l'arc céleste, l'arc de grâce et de consolation que sa main tendit dans les nuages.
- Document 2 :
  - À peine distinguait-on deux butts à l'extrémité de la carrière : des chênes ombrageaient l'un, autour de l'autre des palmiers se dessinaient dans l'éclat du soir.
- Document 3 :
  - Ah ! le beau temps de mes travaux poétiques ! les beaux jours que j'ai passés près de toi ! Les premiers, inépuisables de joie, de paix et de liberté ; les derniers, empreints d'une mélancolie **qui** eut bien aussi ses charmes.

# Exemple

## Exemple

L'exemple porte sur les trois documents ( $d_1, d_2, d_3$ ) et le terme analysé est « qui » (soit  $t_1 = \text{qui}$ ). La ponctuation et l'apostrophe sont ignorées.

- Document 1 :

- $TF_{11} = \frac{w_{11}}{|d_1|} = \frac{2}{38}$

- $IDF_1 = \log\left(\frac{|D|}{|\{d_j: t_1 \in d_j\}|}\right) = \log\left(\frac{3}{2}\right)$

- $w_{11} = TF_{11} \times IDF_1 = \frac{2}{38} \cdot \log\left(\frac{3}{2}\right) \approx 0.0092$

- Document 2 :

- $TF_{12} = \frac{w_{12}}{|d_2|} = 0$

- $IDF_1 = \log\left(\frac{|D|}{|\{d_j: t_1 \in d_j\}|}\right) = \log\left(\frac{3}{2}\right)$

- $w_{12} = TF_{12} \times IDF_1 = 0 \cdot \log\left(\frac{3}{2}\right) = 0$

- Document 3 :

- $TF_{13} = \frac{w_{13}}{|d_3|} = \frac{1}{40}$

- $IDF_1 = \log\left(\frac{|D|}{|\{d_j: t_1 \in d_j\}|}\right) = \log\left(\frac{3}{2}\right)$

- $w_{13} = TF_{13} \times IDF_1 = \frac{1}{40} \cdot \log\left(\frac{3}{2}\right) \approx 0.0044$

# Représentation des documents

## Représentation des documents

Le principal intérêt de ramener les données textuelles à des tableaux de nombres est de faciliter le calcul de distances entre textes.

- Vecteurs de document
- Matrice Terme/Document ou Document/terme

$$\begin{matrix} & d_1 & d_2 & \dots & \dots & \dots & d_d \\ \begin{matrix} t_1 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ t_n \end{matrix} & \left( \begin{array}{cccccc} w_{1,1} & w_{1,2} & \dots & \dots & \dots & w_{1,d} \\ & & \vdots & \vdots & & w_{2,d} \\ w_{2,1} & & \vdots & \vdots & & \\ \vdots & & & & & \vdots \\ \vdots & & \dots & w_{i,j} & \dots & \vdots \\ \vdots & & & & & \vdots \\ \vdots & & \dots & & \dots & \vdots \\ \vdots & & & \vdots & \vdots & \vdots \\ \vdots & & & \vdots & \vdots & \vdots \\ w_{n,1} & \dots & & & \dots & w_{n,d} \end{array} \right) ,
 \end{matrix}$$

$$w_{ij} = TF_{ij} \times IDF_i$$

## Mesures de distances et de similarité

### Similarité entre documents

- Le principal intérêt de ramener les données textuelles à des tableaux de nombres est de faciliter le calcul de distances entre textes.
- Etre en mesure d'évaluer des proximités entre données est en effet un pré-requis fondamental de beaucoup de programmes de fouille de textes, notamment pour les tâches de RI et de classification.
- Permet de ranger les documents par pertinence
- Le cosinus de l'angle est souvent utilisé

$$\cos(d_1, d_2) = \frac{d_1^T \bullet d_2}{\|d_1\| \cdot \|d_2\|}$$

### Remarque

En règle générale, pour mesurer finement la similarité entre des séquences de texte, les vecteurs sont construits d'après un calcul de type TF-IDF (term frequency-inverse document frequency)

## Réduction des dimensions

- La réduction des dimensions est un problème crucial pour la catégorisation de textes et l'apprentissage en général :
  - Le coût du traitement car le nombre des termes intervient dans l'expression de la complexité de l'algorithme ; plus ce nombre est élevé, plus le volume de calcul est important ;
  - La faible fréquence de certains termes : on ne peut pas construire des règles fiables à partir de quelques occurrences dans l'ensemble d'apprentissage.
- Les techniques de la réduction des dimensions sont issues de la théorie de l'information et de l'algèbre linéaire :
  - Seuillage de fréquence (Document Frequency Thresholding)
  - Test du  $\chi^2$  : Détermine les termes les plus caractéristiques de chaque catégorie
  - Latent Semantic Indexing (LSI)

### Remarque

Cependant, même après la suppression des mots plus fréquents et des mots très rares, le nombre de candidats reste encore élevé, et il faut utiliser une méthode statistique pour choisir les mots utiles pour discriminer entre documents pertinents et non pertinents.

## Seuillage de fréquence

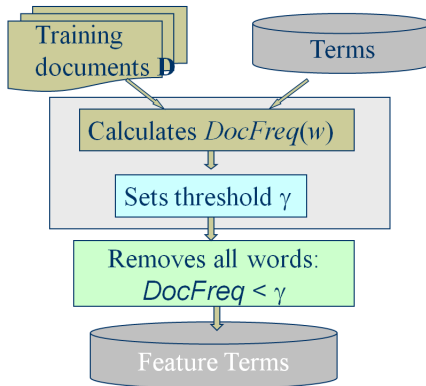


Figure – Seuillage de fréquence

# LA STATISTIQUE DU CHI2 (CHI-SQUARE (CHI))

## Définition

La statistique du  $\chi^2$  mesure l'écart à l'indépendance entre une caractéristique  $t$  et une classe  $c_i$ . C'est une mesure statistique bien connue, elle s'adapte bien à la sélection d'attributs, car elle évalue le manque d'indépendance entre un mot et une classe.

$$\chi^2(t_j, c_k) = \frac{N.[P(t_j, c_k).P(\bar{t}_j, \bar{c}_k) - P(\bar{t}_j, c_k).P(t_j, \bar{c}_k)]^2}{P(t_j).P(\bar{t}_j) - P(c_k).P(\bar{c}_k)}$$

- $P(t_j, c_k)$  : représente la probabilité des documents contenant le terme  $t_j$  dans la catégorie  $c_k$ .
- $P(\bar{t}_j, \bar{c}_k)$  : représente la probabilité des documents contenant le terme  $\bar{t}_j$  dans la catégorie  $\bar{c}_k$ .
- $P(t_j, \bar{c}_k)$  : représente la probabilité des documents contenant le terme  $t_j$  dans la catégorie  $\bar{c}_k$ .
- $P(\bar{t}_j, c_k)$  : représente la probabilité des documents contenant le terme  $\bar{t}_j$  dans la catégorie  $c_k$ .

# LA STATISTIQUE DU CHI2 (CHI-SQUARE (CHI))

## Estimation d'indépendance entre termes et catégories

La statistique du  $\chi^2$  mesure l'écart à l'indépendance entre un terme  $t$  et une classe  $c_i$  :

$$\chi^2(t_j, c_k) \approx \frac{N.[A.D - B.C]^2}{(A+B)(A+C)(B+D)(C+D)}$$

- $A := |d_i \in c_k : t_j \in d_i|$
- $B := |d_i \notin c_k : t_j \in d_i|$
- $C := |d_i \in c_k : t_j \notin d_i|$
- $D := |d_i \notin c_k : t_j \notin d_i|$
- $N$  : Nombre total des documents

## Term categorical score

- $\chi_{max}^2(t_j) = \max_k \chi^2(t_j, c_k)$

## LA STATISTIQUE DU CHI2

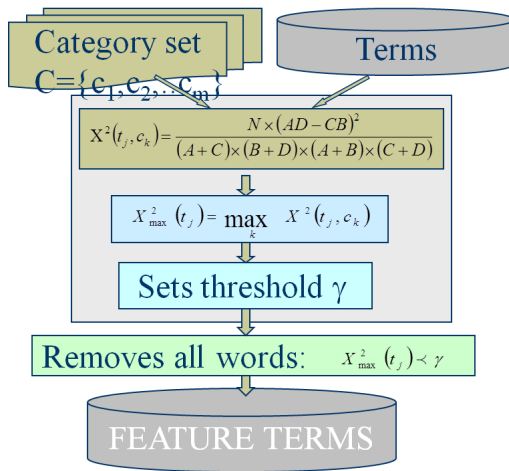


Figure – LA STATISTIQUE  $\chi^2$

# Indexation sémantique latente(LSI)

## Définition

- Indexation sémantique latente (ou LSI, de l'anglais : Latent semantic indexation) est un procédé de traitement des langues naturelles, dans le cadre de la sémantique vectorielle.
- Elle permet d'établir des relations entre un ensemble de documents et les termes qu'ils contiennent, en construisant des **concepts** liés aux documents et aux termes.

## Remarques

- La LSI utilise une matrice qui décrit l'occurrence de certains termes dans les documents. C'est une matrice creuse dont les lignes correspondent aux **termes** et dont les colonnes correspondent aux **documents**.
- la LSI permet de trouver une matrice de rang plus faible, qui donne une approximation de cette matrice des occurrences.

# Indexation sémantique latente(LSI)

## Principe de LSI

### ① Construction de la matrice des occurrences

- Soit  $X$  la matrice où l'élément  $(i, j)$  décrit les occurrences du terme  $i$  dans le document  $j$  ( par exemple la fréquence). Alors  $X$  aura cette allure :

$$w_{ij} = TF_{ij} \times IDF_i$$

$$\begin{matrix} & d_1 & d_2 & \dots & \dots & \dots & d_d \\ \begin{matrix} t_1 \\ \vdots \\ \vdots \\ \vdots \\ t_n \end{matrix} & \begin{pmatrix} w_{1,1} & w_{1,2} & \dots & \dots & \dots & w_{1,d} \\ & & & & & \\ & w_{2,1} & & & & w_{2,d} \\ & \vdots & & & & \vdots \\ & \vdots & \dots & w_{i,j} & \dots & \vdots \\ & w_{n,1} & \dots & & \dots & w_{n,d} \end{pmatrix} \end{matrix},$$

### ② Décomposition en valeurs singulières :

- On effectue alors une décomposition en valeurs singulières sur  $X$ , qui donne deux matrices orthonormales  $U$  et  $V$  et une matrice diagonale  $S$ .

$$X = USV^T$$

### ③ Espace des concepts

- Lorsqu'on sélectionne les  $k$  plus grandes valeurs singulières, ainsi que les vecteurs singuliers correspondants dans  $U$  et  $V$ , on obtient une approximation de rang  $k$  de la matrice des occurrences.

**FIN**