1 – Contexte

On dispose d'un fichier qui collecte des articles de structure identique. Ces articles peuvent être, accidentellement, identiques ou très semblables. On appelle « Doublon » deux articles identiques ou très semblables.

Pour le bon fonctionnement du fichier, il est indispensable de détecter et filtrer les doublons présents dans le fichier afin de les détruire si nécessaire.

2 – Critères de filtrage attendus

Par définition, deux articles sont des « Doublons » s'ils sont identiques ou très semblables.

Pour détecter rapidement des articles « identiques ou très semblables », on s'intéresse aux articles qui possèdent la même « signature » (c'est-à-dire la même trace lors d'une projection sur un hyperplan de référence).

3 – Méthode de filtrage préconisée

La méthode préconisée est le **modèle vectoriel** qui est une méthode algébrique de représentation d'un texte qui tient compte du vocabulaire (ou de l'alphabet) support du texte.

Le modèle vectoriel est classiquement utilisé en recherche d'information, notamment pour la recherche documentaire, la classification ou le filtrage de données. Ce modèle a été initialement conçu pour les documents textuels (même s'il a été depuis étendu à d'autres types de contenus plus riches).

Le modèle vectoriel consiste à calculer la signature d'un texte en le représentant sous forme vectorielle. La signature d'un texte est une représentation simplifiée de ce texte qui présente plusieurs avantages :

- La signature est beaucoup plus légère à manipuler que le texte initial.
- Deux textes distincts (mais très semblables) peuvent avoir des signatures voisines (voire identiques).
- On peut détecter rapidement des textes « identiques ou très semblables » à partir du moment où ils ont la même siganture.

Le modèle vectoriel consiste à décrire un texte sur un espace vectoriel **N**-dimensionnel dont les vecteurs de bases sont les **N** mots du vocabulaire (ou de l'alphabet) de référence. Ainsi, le texte est assimilé à un vecteur **V** de dimension **N** dont chaque composante **V[k]**, pour k=1..N, correspond au nombre d'occurrences du k-ième mot du vocabulaire de référence.

Par exemple, soit le texte **T** = « Le petit chat est mort » et le vocabulaire = {petit, gros, chat, blanc, noir, mort, vivant}. Alors, le cardinal du vocabulaire est **N** = 7 et le vecteur **V** représentant le texte **T** sur ce vocabulaire est **V** = [1, 0, 1, 0, 0, 1, 0]

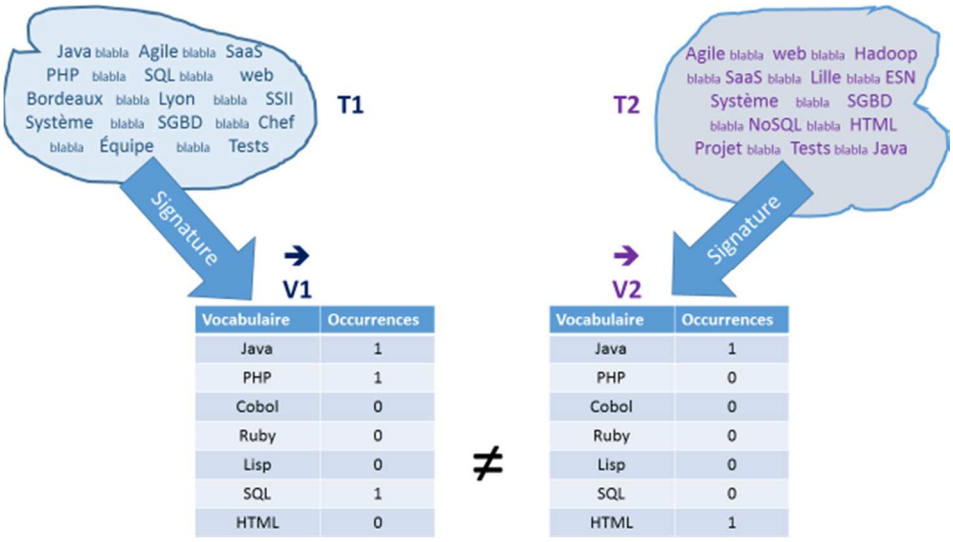
Dans ce même modèle vectoriel, le texte **T'** = « Un gros chat est plus lent qu'un petit chat » aurait pour représentation vectorielle **V'** = [1, 1, 2, 0, 0, 0, 0].

L'ensemble de représentation des textes est un vocabulaire formé avec les mots les plus significatifs du corpus considéré. Chaque texte **T** est ainsi représenté par un vecteur **V**, dont la dimension correspond au cardinal **N** du vocabulaire. Chaque élément **V[k]** du vecteur **V** correspond au poids associé au mot d'indice k dans le vocabulaire et pour le texte Indexé.

Avec le modèle vectoriel, l'ordre d'apparition des mots du vocabulaire dans le texte n'a aucun impact sur la signature.

Une fois que chaque texte est associé à sa signature vectorielle sur un vocabulaire de référence, il devient possible de tester des similarités entre deux textes.

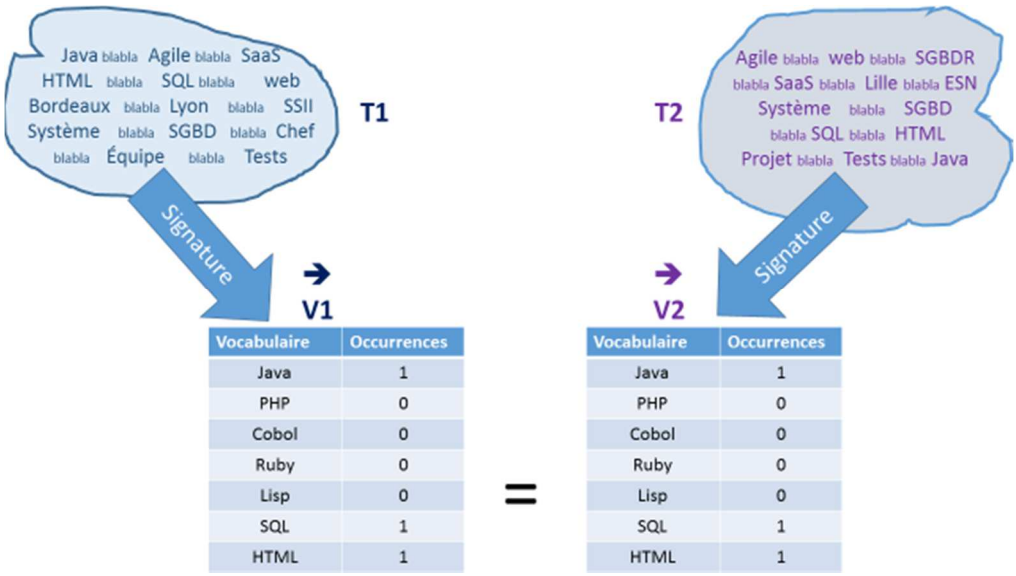
Par exemple, les deux textes suivants (T1 et T2) sont considérés comme différents parce que leurs signatures respectives (**V1** et **V2**) sont différentes.



Vocabulaire	Occurrences
Java	1
PHP	1
Cobol	0
Ruby	0
Lisp	0
SQL	1
HTML	0

Vocabulaire	Occurrences
Java	1
PHP	0
Cobol	0
Ruby	0
Lisp	0
SQL	0
HTML	1

A contrario, les deux textes suivants (T1 et T2) sont considérés comme similaires parce que leurs signatures respectives (**V1** et **V2**) sont identiques.



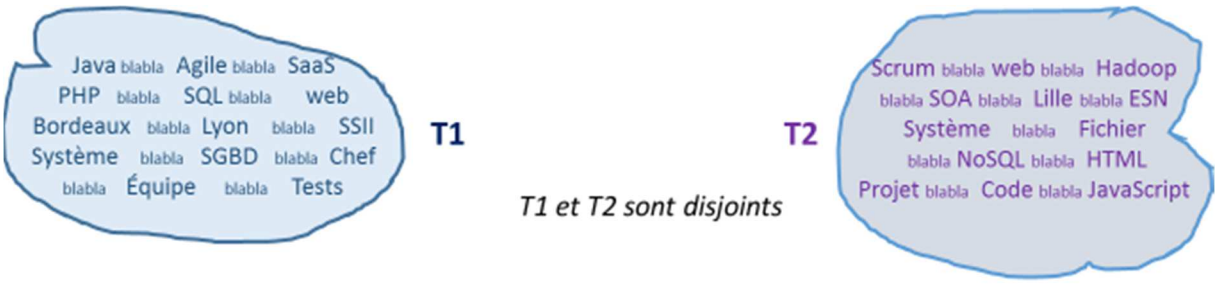
Vocabulaire	Occurrences
Java	1
PHP	0
Cobol	0
Ruby	0
Lisp	0
SQL	1
HTML	1

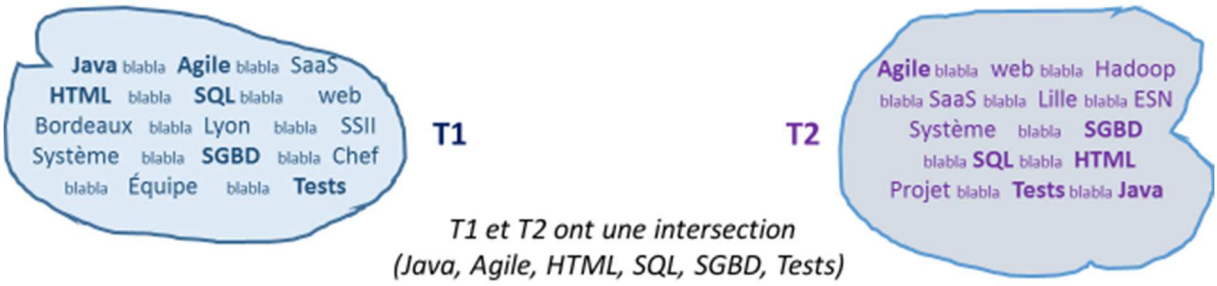
=

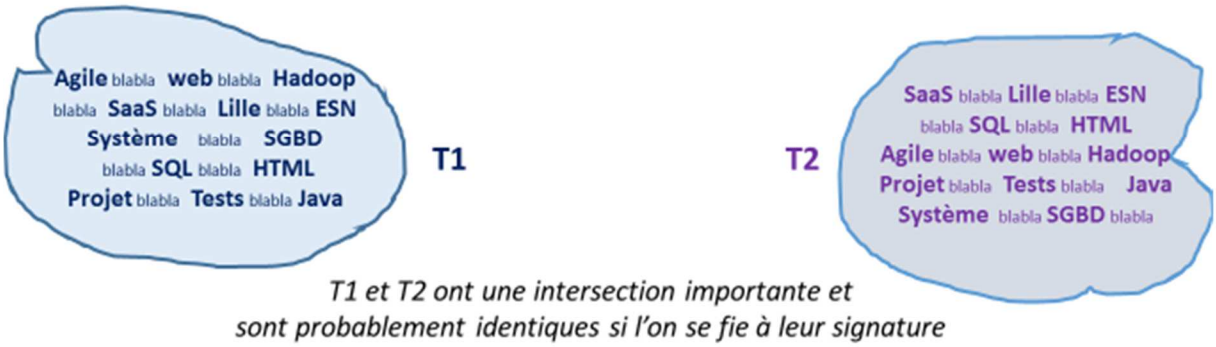
Vocabulaire	Occurrences
Java	1
PHP	0
Cobol	0
Ruby	0
Lisp	0
SQL	1
HTML	1

On note bien que « similaire » n'est pas la même chose que « identique ». Deux textes sont « similaires » si leurs signatures respectives sont « identiques ». Identique, pour une signature, signifie ici que les vecteurs sont strictement égaux.

Une fois que chaque texte est associé à sa signature vectorielle sur un vocabulaire de référence, il devient possible de tester des similarités sémantiques entre deux textes.







T1 et T2 ont une intersection importante et sont probablement identiques si l'on se fie à leur signature

Si deux textes **T1** et **T2** sont totalement disjoints, leurs signatures **V1** et **V2** sont orthogonales (et réciproquement).

Si deux textes **T1** et **T2** ont une intersection importante, leurs signatures **V1** et **V2** peuvent être colinéaires, voire identiques. En ce dernier cas, les deux textes seront considérés comme « similaires ».

Entre ces deux extrêmes, on a un éventail de situations possibles. On peut caractériser ces différentes situations par la proximité sémantique de deux textes.

4 – Mise en œuvre

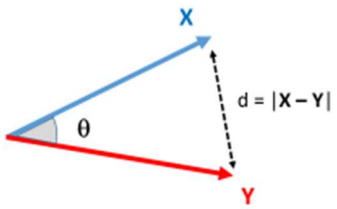
4.1 – Détection des doublons

On appelle « doublons » deux textes **Ti** et **Tj** qui sont identiques ou très semblables. On définit comme « identiques ou très semblables » deux textes **Ti** et **Tj** qui possèdent la même « signature » (c'est-à-dire le même nombre d'occurrences de mots appartenant au vocabulaire de référence. Autrement dit, la distance entre les deux vecteurs **Vi** et **Vj** est nulle, soit : |**Vi** – **Vj** | = 0

4.2 – Calcul de la distance inter-vecteurs

Soit **X** et **Y** deux vecteurs

avec **X** : (x[1], x[2], x[3], . . . , x[i], . . ., x[N]) et **Y** : (y[1], y[2], y[3], . . . , y[i], . . ., y[N]).



Les deux vecteurs **X** et **Y** appartiennent à un espace vectoriel à N dimensions et sont décrits dans un repère orthonormé. Dans le présent contexte d'utilisation de ces vecteurs, tous les x[i] et tous les y[i] sont positifs ou nuls. En conséquence, on peut supposer que les deux vecteurs X et Y forment un angle **Θ** qui est positif et inférieur à 90°. Autrement dit, cos **Θ** est compris entre 0 et +1.

Le produit scalaire **X**.**Y** = **Σ x[i] . y[i]**, pour i=1..N.

Le carré de la norme de **X** est $|\mathbf{X}|^2 = \mathbf{X}.\mathbf{X} = \mathbf{\Sigma\ x[i]\ .\ x[i]}$, pour i=1..N.

Le carré de la norme de **Y** est $|\mathbf{Y}|^2 = \mathbf{Y}.\mathbf{Y} = \mathbf{\Sigma\ y[i]\ .\ y[i]}$, pour i=1..N.

En conséquence,

$$\frac{\mathbf{X.Y}}{|\mathbf{X}|.|\mathbf{Y}|} = \frac{|\mathbf{X}|.|\mathbf{Y}|\cos\Theta}{|\mathbf{X}|.|\mathbf{Y}|} = \cos\Theta$$

Par ailleurs, **X – Y** représente le vecteur-écart entre **X** et **Y**.

avec **X – Y** : (x[1]-y[1], x[2]-y[2], x[3]-y[3], . . . , x[i]-y[i], . . ., x[N]-y[N])

En conséquence, la distance d est égale à la norme de **X – Y**, soit $|\mathbf{X} - \mathbf{Y}|$.

Autrement dit, $d^2 = |\mathbf{X - Y}|^2 = (\mathbf{X-Y}).(\mathbf{X-Y}) = \mathbf{X.X} - \mathbf{Y.X} - \mathbf{X.Y} + \mathbf{Y.Y} = |\mathbf{X}|^2 + |\mathbf{Y}|^2 - 2\ \mathbf{X.Y}$

D'où $d^2 = |\mathbf{X - Y}|^2 = (\mathbf{X-Y}).(\mathbf{X-Y}) = |\mathbf{X}|^2 + |\mathbf{Y}|^2 - 2\ \mathbf{X.Y}$

$$d^2 = |\mathbf{X - Y}|^2 = |\mathbf{X}|^2 + |\mathbf{Y}|^2 - 2\ \mathbf{X.Y} = |\mathbf{X}|^2 + |\mathbf{Y}|^2 - 2\,|\mathbf{X}|.|\mathbf{Y}|\cos\Theta$$

Ce qui permet de relier la distance inter-vecteur d = $|\mathbf{X-Y}|$ et cos **Θ** où **Θ** est l'angle inter-vecteur**.** Par la suite, chercher à minimiser $|\mathbf{X} - \mathbf{Y}|$ revient à maximiser cos **Θ.** Ce qui revient à minimiser **Θ.**

$$2\cos\Theta = \frac{|\mathbf{X}|}{|\mathbf{Y}|} + \frac{|\mathbf{Y}|}{|\mathbf{X}|} - \frac{|\mathbf{X} - \mathbf{Y}|^2}{|\mathbf{X}|.|\mathbf{Y}|}$$