Bases de données documentaires et distribuées
Cours NFE04
Analyse des textes avec Solr

Auteurs : Raphaël Fournier-S'niehotta, Philippe Rigaux, Nicolas Travers
prénom.nom@cnam.fr

Département d'informatique
Conservatoire National des Arts & Métiers, Paris, France

Effet de l'analyse

Quelques requêtes à faire sur la base des films au préalable :

- `Matrix`, `MATRIX` et `matrix`
- `title:Matrix` et `text:Matrix`
- `title:reloaded` et `text:reloaded`

Que se passe-t-il ?

Si vous avez bien suivi les explications sur le schéma vous devriez comprendre ce qui se passe. Réflechissons ensemble...

Rôle de l'analyse

L'analyse des textes permet d'effectuer une forme de normalisation / unification pour être moins dépendant de la forme du texte.

- un document parle de loup même si on y trouve les formes "loups", "Loup", "louve", etc.
- un document parle de travail quellle que soit la forme du verbe "travailler" ou de ses variantes.
- Jusqu'où va-t-on ? Traductions (loup = wolf = lupus) ? Synonymes (loup = prédateur) ? Recherche en cours.

Pour moins dépendre de la forme on applique une **transformation**.

Très important

La même transformation doit être appliquée aux documents **et** à la requête (ex. de la requête "MATRIX")

Impact de l'analyse

Bien comprendre.

Plus on normalise, plus on diminue la précision.
Car des mots distincts sont unifiés (cote, côte, côté, etc.)

Plus on normalise, plus on améliore le rappel.

Car on met en correspondance les variantes d'un même mot, d'une même signification (conjuquaisons d'un verbe).

Les phases de l'analyse

Avant d'insérer dans l'index, plusieurs phases.

Important : identification de quelques méta-données (la langue), prise en compte du contexte (quels documents pour quelle application). Puis, de manière générale :

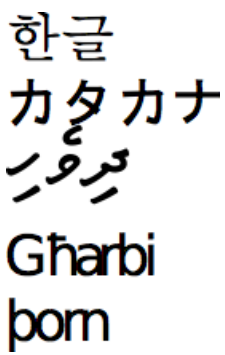
- **Tokenization** : découpage du texte en "mots"
- **Normalisation** : majuscules ? acronymes ? apostrophes ? accents ?
  Exemple : *Windows* et *window*, U.S.A vs USA, *l'auditeur* vs *les auditeurs*.
- **Stemming** ("racinisation"), **lemmatization**
  Prendre la racine des mots pour éviter le biais des variations (auditer, auditeur, audition, etc.)
- **Stop words**, quels mots garder ?
  Mots très courants peu informatifs (le, un à, de).

C'est de l'art et du réglage... Dans ce qui suit : introduction / sensibilisation aux problèmes.

Identification de la Langue

Comment trouver la langue d'un document ?

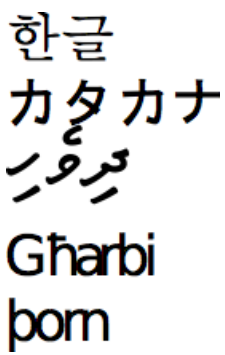
- Méta information (dans l'entête HTTP p.e.) : pas fiable du tout.
- **Par le jeu de caractères**, pas assez courant !



Identification de la Langue

Comment trouver la langue d'un document ?

- Méta information (dans l'entête HTTP p.e.) : pas fiable du tout.
- **Par le jeu de caractères**, pas assez courant !



  Respectivement : Coréen, Japonais, Maldives, Malte, Islandais.
- Par extension : **séquences de caractères fréquents**, (*n*-grams)
- Par techniques d'apprentissage (**classifiers**)

Des librairies font ça très bien (e.g., Tika, `http://tika.apache.org`)

Tokenisation

Principe

Séparation du texte en **tokens** (“mots”)

Pas du tout aussi facile qu'on le dirait !

- Dans certaines langues (Chinois, Japonais), les mots **ne sont pas** séparés par des espaces.
- Certaines langues s'écrivent de droite à gauche, de haut en bas.
- Que faire (et de manière **cohérente**) des acronymes, élisions, nombres, unités, URL, email, etc.
- **Mots composés** : les séparer en *tokens* ou les regrouper en un seul ?
  1. Anglais : *hostname*, *host-name* et *host name*, ...
  2. Français : Le Mans, aujourd'hui, pomme de terre, ...
  3. Allemand : *Levensversicherungsgesellschaftsangestellter* (employé d'une société d'assurance vie)

  Que faire si l'utilisateur cherche *hostname* et qu'on a normalisé en *host-name* ?

Majuscules, ponctuation ? Une solution simple est de normaliser (minuscules, pas de ponctuation).

Exemple pour notre petit jeu de données

On met en minuscules, on retire la ponctuation.

d_1	le loup est dans la bergerie
d_2	le loup et les trois petits cochons
d_3	les moutons sont dans la bergerie
d_4	spider cochon spider cochon il peut marcher au plafond
d_5	un loup a mangé un mouton les autres loups sont restés dans la bergerie
d_6	il y a trois moutons dans le pré et un mouton dans la gueule du loup
d_7	le cochon est à 12 euros le kilo le mouton à 10 euros le kilo
d_8	les trois petits loups et le grand méchant cochon

On considère que l'espace est le séparateur de tokens.

Stemming (racine), lemmatization

Principe

Confondre toutes les formes d'un même mot, ou de mots apparentés, en une seule **racine**.

Stemming Morphologique. Retire les pluriels, marque de genre, conjugaisons, modes, etc.

- Très dépendant de la langue : *geese* pluriel de *goose*, *mice* de *mouse*
- Difficile à séparer d'une analyse linguistique ("Les poules du couvent couvent", "la petite brise la glace" : où est le verbe ?)

Stemming lexical Fondre les termes proches lexicalement : "politique, politicien, police (?)" ou "université, universel, univers (?)"

Stemming phonétique. Correction fautes de frappes, fautes orthographes

Exemple de stemming

On retire les pluriels, on met le verbe à l'infinitif.

d_1	le loup etre dans la bergerie
d_2	le loup et les trois petit cochon
d_3	les moutons etre dans la bergerie
d_4	spider cochon spider cochon il pouvoir marcher au plafond
d_5	un loup avoir manger un mouton les autres loups etre rester dans la bergerie
d_6	il y avoir trois mouton dans le pre et un mouton dans la gueule du loup
d_7	le cochon etre a 12 euro le kilo le mouton a 10 euro le kilo
d_8	les trois petit loup et le grand mechant cochon

Suppression des *Stop Words*

Principe

On retire les mots porteurs d'une information faible afin de limiter le stockage.

articles : *le*, *le*, *ce*, etc.
verbes "fonctionnels" *être*, *avoir*, *faire*, etc.
conjunctions : *that*, *and*, etc.
etc.

Remarque

Maintenant moins utilisé car (i) espace de stockage peu coûteux et (ii) pose d'autres problèmes ("pomme de terre", "Let it be", "Stade de France")

Autres problèmes, en vrac

Majuscules / minuscules
Lyonnaise des Eaux, Société Générale, etc.

Acronymes
CAT = *cat* ou *Caterpillar Inc.* ? M.A.A.F ou MAAF ou Mutuelle ... ?

Dates, chiffres
Monday 24, August, 1572 – 24/08/1572 – 24 août 1572 10000 ou 10,000.00 ou 10,000.00

Accents, ponctuation
résumé ou résume ou resume...

Remarque

Dans tous les cas, les même règles de transformation s'appliquent aux documents ET à la requête.

Exemple avec suppression des *stop words*

Voici une solution possible.

d_1	loup etre bergerie
d_2	loup trois petit cochon
d_3	mouton etre bergerie
d_4	spider cochon spider cochon pouvoir marcher plafond
d_5	loup avoir manger mouton autres loups etre rester bergerie
d_6	avoir trois mouton pre mouton gueule loup
d_7	cochon etre 12 euro kilo mouton 10 euro kilo
d_8	trois petit loup grand mechant cochon

On a gardé les verbes fonctionnels (être, avoir).

Définir une chaîne d'analyse avec Solr

Un analyseur est associé à un champ :

- le tokenizer effectue le traitement lexical consistant à transformer le texte en un ensemble de tokens ;
- les filtres (filter) examinent les tokens un par un et décident de les conserver, de les remplacer par un ou plusieurs autres ;
- l'analyseur est une chaîne de traitement (pipeline) constituée de tokeniseurs et de filtres.

Exemple minimal :

```
   <fieldType name="text" class="solr.TextField">
<analyzer>
     <tokenizer class="solr.StandardTokenizerFactory" />
     <filter class="solr.LowerCaseFilterFactory" />
</analyzer>
   </fieldType>
```

tester les analyseurs

Très pratique : l'interface
`http://localhost:8983/solr/#/movies/analysis`.

- dans le champ field value (index), on saisit un texte à analyser ;
- dans le menu déroulant au dessous, on choisit un des champs du schéma de l'index ; c'est donc l'analyseur de ce champ qui sera appliqué.

Démonstration !