

Extraction d'information et indexation de documents

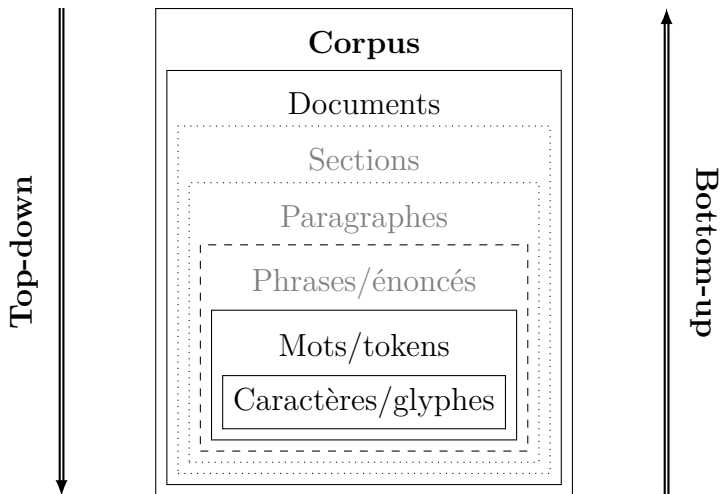
Damien Nouvel



Plan

1. Représenter des documents
2. Rechercher dans des documents

Approches pour la représentation



Indexs et documents

▶ Index

- Doigt
- Ouvrages censurés (XVI^{ème}, *index librorum prohibitorum*)
- Liste alphabétique de termes
- ⇒ **Référence, pointeur**
- ⇒ **Triangle sémiotique** [Ogden & Richards 1923]

▶ Objectif : accès rapide au contenu

- Du corpus (bibliothèque) vers les documents
- Depuis un document vers une partie de son contenu

Décomposition du document

- ▶ Métadonnées : auteur, date, liens, etc.
- ▶ Document comme agrégats
 - Chapitres, sections, sous-sections ...
 - Paragraphes
 - Phrases
 - **Mots**
 - Caractères, glyphes
- ▶ Segmentation selon des critères **explicites** ou **implicites**
 - Disposition du contenu (chapitres, paragraphes)
 - ⇒ Méta-information pour le document numérique
 - ⇒ Explicite, mais pas toujours renseignée
 - Contenu lui-même (phrases, mots)
 - ⇒ Selon la séquence des caractères
 - ⇒ Peu explicite, mais calculable
 - ⇒ Utilisation des espaces et ponctuations

Représentation des documents

- ⇒ Documents comme **séquences de caractères**
 - ▶ Manipulation aisée ... mais pas de **sens**
 - ▶ Unité minimale **sémantique** : le **mot**
 - Ou : mot-forme, lemme, lexème, **token**, morphème ...
 - Lien entre **forme** et **sens**
 - Attention aux expressions polylexicales (locutions)
 - ⇒ **Segmentation**
- ⇒ Documents comme **séquences de mots**
 - ▶ Problèmes de la **séquence**
 - Quel intérêt (position dans un document)
 - Combinatoire des séquences
- ⇒ Documents comme **ensembles de mots**
- ⇒ **Sacs de mots**
- ⇒ Tenir compte des collocations, homonymes, synonymes
- ⇒ « Normaliser » les mots

Plan

1. Représenter des documents
2. Rechercher dans des documents

Matrice termes / documents

- ▶ Représentation à l'aide de **matrices** (tableaux)
 - Lignes : documents
 - Colonnes : termes (mots)
 - Cellules : **occurrences** d'un terme dans un document
- ⇒ Matrice termes par documents
- ⇒ *Vector Space Model*
- ⇒ Représentation mathématique, opérations **algébriques**
- ▶ Exemple (sport vs politique)

	foot.	basket.	ballon	gouv.	ministre	aller
Euro 2016	3	0	2	0	0	3
Tony Parker	0	3	1	1	0	5
Présidentielles 2017	0	0	0	4	3	1
COP21	0	0	1	5	2	4
Affaire Blatter	5	0	0	2	1	7

Exploitation de la matrice

- ▶ Transformation de matrice (booléen, stopwords, TF.IDF, LSA)
- ▶ **Catégorisation** de documents
 - Comparer les documents deux-à-deux
 - Mesure de **distance** / **similarité** (euclide, cosinus)
- ▶ **Indexation** de documents
 - Comparer une requête à un document
 - Interprétation de la requête (mots-outils)
 - Mesure de **distance** / **similarité** (euclide, cosinus)
 - Contraintes d'accès (rapide)
- ▶ Problèmes **linguistiques**
 - **Proximité sémantique** (dispersion des termes)
 - **Synonymie**
 - **Homonymes**

Applications et évaluation

▶ Applications

- Moteurs de recherche
- ⇒ Google, Bing, Yahoo, Qwant, Baidu, Yandex, Naver...
- Similarités entre bases documentaires
- Textométrie (calcul de spécificités)

▶ Évaluations (pertinence pour une recherche)

- **Précision** : les documents sont-ils corrects? (bruit)
- **Rappel** : y-a-t-il tous les documents? (silence)
- F-mesure : moyenne harmonique des deux précédents
- Courbe rappel / précision
- Précision moyenne (interpolée)
- ...

⇒ Problème de la **temporalité** (infos d'actualité vs atemporel)