

RECHERCHE D'INFORMATION & TRAITEMENT AUTOMATIQUE DU LANGAGE

Cours 1 : RI - introduction et indexation

Monday 14th February, 2022

Laure Soulier



Organisation du cours

- Indexer et interroger une collection de documents
→ Développer un moteur de recherche
- évaluer un moteur de recherche
- Découvrir les avancées récentes dans le domaine sous l'angle du deep learning

RI - Introduction

"Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers)." C. Manning

- Application la plus courante : Moteurs de recherche



- Mais aussi dans : les entreprises, les bibliothèques numériques, domaines d'application (médecine, droit, ...), nos ordinateurs...

La donnée, l'or noir de l'information

Données

- By 2020, new information generated/second/person 1.7 MB
- Users on Facebook/minute: 31.25 million messages and watch 2.77 million videos

Informations

- 40,000 search queries on Google/day
 - 16% to 20% of queries that get asked every day have never been asked before.
 - In 2012, 1min to index 50million pages
- RI : Collecter, organiser et identifier la bonne donnée au bon moment pour le bon utilisateur pour lui donner de l'information

Connaissances

Data Mining, Machine Learning

- Texte : articles, livres (pdf, ps, ebook, html, xml, ...)
- Images, Vidéos, Son, Musique
- Pages/sites Web dynamiques
- Médias sociaux (blogs, Twitter, ...) : dynamisme, structure relationnelle
- Messageries - fils de discussion
- Information majoritairement peu structurée, mais structures exploitables (HTML, XML), relations (web réseaux sociaux), hiérarchies, ...

Des enjeux...

- Liés à la diversité des demandes d'accès à l'information
 - Consultation (browsing)
 - Requêtes booléennes, mots-clés
 - Recherche automatique (e.g., robots)
 - Suivi d'évènements, analyse de flux
 - Extraction d'information de texte, de web caché, ...
 - ...

- **Centre:** The overall goal of the track is to develop and tune a reproducibility evaluation protocol for IR.
- **CheckThat!:** CheckThat! aims to foster the development of technology capable of both spotting and verifying check-worthy claims in political debates in English and Arabic.
- **LifeCLEF:** LifeCLEF lab aims at boosting research on the identification of living organisms and on the production of biodiversity data in general. LifeCLEF is intended to push the boundaries of the state-of-the-art in several research directions at the frontier of multimedia information retrieval, machine learning and knowledge engineering.
- **PAN:** Scientific events on digital texts: Author identification –Author obfuscation –
- **CLEF eHealth:** Medical content is available electronically in a variety of forms ranging from patient records and medical dossiers, scientific publications and health-related websites to medical-related topics shared across social networks. This lab aims to support the development of techniques to aid laypeople, clinicians and policy-makers in easily retrieving and making sense of medical content to support their decision making.

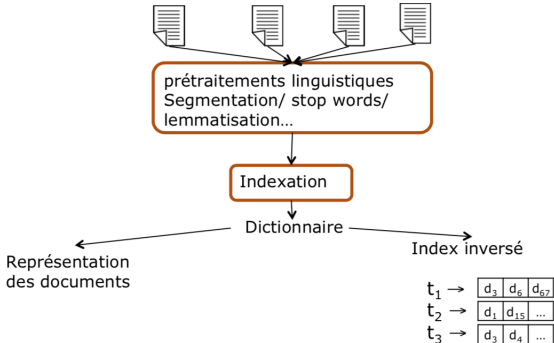
Campagne d'évaluation CLEF

- **ImageCLEF** Interpreting and summarizing the insights gained from medical images The task addresses the problem of bio-medical image caption prediction from large amounts of training data.
- **PIR-CLEF**: The primary aim of the PIR-CLEF laboratory is to provide a framework for evaluation of Personalised Information Retrieval
- **Early risk prediction on the Internet**: eRisk explores the evaluation methodology, effectiveness metrics and practical applications (particularly those related to health and safety) of early risk detection on the Internet.
- **Dynamic Search for Complex Tasks**: The goal of the CLEF Dynamic Search lab is to propose and standardize an evaluation methodology that can lead to reusable resources and evaluation metrics able to assess retrieval performance over an entire session, keeping the “user” in the loop.
- **Multilingual Cultural Mining and Retrieval**: Developing processing methods and resources to mine the social media sphere surrounding cultural events such as festivals. This requires to deal with almost all languages and dialects as well as informal expressions.

Schéma général et notions

Indexation

Chaîne d'indexation



- Indexation peut être manuelle, automatique, semi-automatique
- Elle peut aussi reposer sur un langage libre (issu du texte) ou contrôlé (lexiques, ressources sémantiques, ...)
- L'objectif est d'identifier la distribution des termes pour représenter les documents

Lois de distribution des termes pour les collections de documents

- Loi de Zipf - Expression formelle

$$f(r, s, N) = \frac{\frac{1}{r^s}}{\sum_{n=1}^N \frac{1}{n^s}} \propto \frac{1}{r^s} \text{ ainsi } \log(f) = -s \cdot \log(r) + cst \quad (1)$$

où r : rang, N : taille du corpus, s : paramètre du corpus

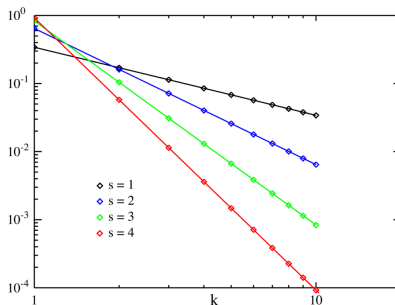


Figure 1: Analyse log fréquence vs. log rang. $k = s$ et $N = 10$ (source Wikipedia)

- Autres phénomènes suivant une loi puissance similaire à la Loi de Zipf
 - Fréquence d'accès des pages web n Population des villes
 - Trafic internet par site
 - Noms dans une population
 - ...

Lois de distribution des termes pour les collections de documents

- Loi de Heaps

- Stipule que le nombre de mots distincts dans le vocabulaire d'une collection est proportionnel au nombre de mots dans la collection

$$V = Kn^{\beta} \quad (2)$$

V : taille du vocabulaire, N : taille du texte, K et β : paramètres dépendant du texte (en Anglais : K entre 10 et 100 – β entre 0.4 et 0.6)

- Croissance sous-linéaire du vocabulaire en fonction de la taille du texte ou de la collection
 - * L'index n'a pas borne supérieure (noms propres, erreurs de typos, etc.)
 - * Les nouveaux mots apparaissent moins fréquemment quand le vocabulaire croît

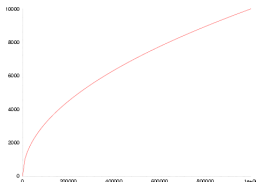


Figure 2: The x-axis: text size, and the y-axis: the number of distinct vocabulary elements present in the text

prétraitements linguistiques
Segmentation/ stop words/
lemmatisation...

Indexation

Dictionnaire

Représentation des documents

Index inversé

$t_1 \rightarrow$	d_3	d_6	d_{67}
$t_2 \rightarrow$	d_1	d_{15}	...
$t_3 \rightarrow$	d_3	d_4	...

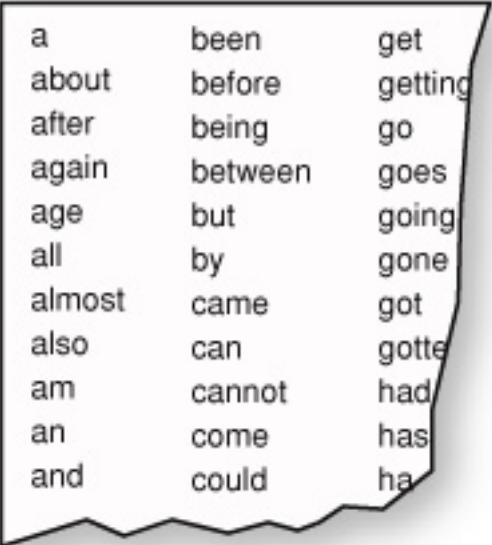
- Index : représentation directe des documents
- Index inversé : représentation avec les termes pour point d'entrée.

Prétraitement et représentation des textes - stopwords

- Quelles unités conserver pour l'indexation ? stopword/anti-dictionnaires
 - Les mots les plus fréquents de la langue "stop words" n'apportent pas d'information utile e.g. prepositions, pronoms, mots « athématiques »,... (peut représenter jusqu'à 30 ou 50)
 - Ces "stop words" peuvent être dépendants d'un domaine ou pas
L'ensemble des mots éliminés est conservé dans un anti-dictionnaire (e.g. 500 mots).
 - Les mots les plus fréquents ou les plus rares dans un corpus (frequency cut-off)
 - Les connaissances sémantiques permettent également d'éliminer des mots
 - Techniques de sélection de caractéristiques

Prétraitement et représentation des textes - stopwords

Stopword list



a	been	get
about	before	getting
after	being	go
again	between	goes
age	but	going
all	by	gone
almost	came	got
also	can	gotte
am	cannot	had
an	come	has
and	could	ha

Prétraitement et représentation des textes - Normalisation

- Normalisation textuelle : transformations superficielles pour obtenir la forme canonique (ponctuations, casse, symboles spéciaux, accents, dates et valeurs monétaires)
- Normalisation linguistique :
 - Racinisation : regrouper les différentes variantes morphologiques d'un mot (cheval, chevalier, chevaux → cheva ; amusing, amusement, and amused → amus)
 - Lemmatisation : analyse linguistique e.g. infinitif pour les verbes, singulier pour les noms (amusement, amusing, and amused → amuse)
- Regroupement de mots similaires au sens d'un critère numérique

Prétraitement et représentation des textes - Porter Stemmer

- Largement utilisé en anglais
- 5 phases de réduction des mots appliquées séquentiellement
- Règles de réécriture avec priorité d'application
- Exemple (Manning et al. 2008)
 - sses → ss : caresses → caress
 - ies → i : ponies → poni
 - ss → ss : caress → caress
 - s → : cats → cat

Prétraitement et représentation des textes

- Représentations d'un document
 - Booléenne : présence/absence
 - Réelle : un indicateur numérique qui pondère le terme
 - Sélection de caractéristiques
 - Projections : réduction supplémentaire (SVD, ACP, NMF, Word2Vec, ...)
- Pondération des termes
 - Mesure l'importance d'un terme dans un document : Comment représenter au mieux le contenu d'un document ?
 - Considerations statistiques, parfois linguistiques
 - Loi de Zipf : élimination des termes trop fréquents ou trop rares
 - Facteurs de pondération
 - * e.g.tf (pondération locale), idf (pondération globale)
 - * Normalisation: prise en compte de la longueur des documents, etc

Pré-traitement et représentation - Une méthode de référence : TF-IDF

- Term Frequency $tf(t_i, d)$: nombre occurrences de t_i dans le document d .
Remarque : varie en fonction de la taille des documents. Si on double la taille des documents, tf double. Le document sera considéré comme plus pertinent.
- Inverse Document frequency idf

$$idf(t_i) = \log\left(\frac{1 + N}{1 + df(t_i)}\right) \quad (3)$$

- $df(t_i)$: nombre de documents contenant t_i
- $idf(t_i)$: fréquence inverse, décroît vers 0 si t_i apparait dans tous les documents
- N : nombre de documents

- TF-IDF

$$x_i = tf(t_i, d) * idf(t_i) \quad (4)$$

- Il existe plusieurs variantes de ces poids (lissage, logarithme, ...)

Modèles d'indexation - index

- Index : représentation simple des documents

d1	(t1, n11) ; (t2,n12) ;
d2	(t1, n21) ; (t2,n22) ;
...	
dk	(t1, nk1) ; (t2,nk2) ;

Modèles d'indexation - index inversé

- Index inversé : point d'entrée par les mots

t1	(d1, n11) ; (d3,n13) ;
t2	(d4, n24) ; (d5,n25) ;
...	
tj	(d1, dj1) ; (d7; d72) ;

Modèles d'indexation - index inversé

- Index inversé : point d'entrée par les mots

Doc 1

I did enact Julius Caesar: I was killed
i' the Capitol; Brutus killed me.

Doc 2

So let it be with Caesar. The noble Brutus
hath told you Caesar was ambitious:

term	docID	term	docID	term	doc.	freq.	→	postings lists
I	1	ambitious	2	ambitious	1	→	2	
did	1	be	2	be	1	→	2	
enact	1	brutus	1	brutus	2	→	1	→ 2
julius	1	brutus	2	capitol	1	→	1	
caesar	1	capitol	1	caesar	2	→	1	→ 2
I	1	caesar	1	did	1	→	1	
was	1	caesar	2	enact	1	→	1	
killed	1	caesar	2	hath	1	→	2	
i'	1	did	1	I	1	→	1	
the	1	enact	1	i'	1	→	1	
capitol	1	hath	1	it	1	→	2	
brutus	1	I	1	julius	1	→	1	
killed	1	I	1	killed	1	→	1	
me	1	i'	1	let	1	→	2	
so	2	it	2	me	1	→	1	
let	2	julius	1	noble	1	→	2	
it	2	killed	1	so	1	→	2	
be	2	killed	1	the	2	→	1	→ 2
with	2	let	2	told	1	→	2	
caesar	2	me	1	you	1	→	2	
the	2	noble	2	was	2	→	1	→ 2
noble	2	so	2	with	1	→	2	
brutus	2	the	1					
hath	2	the	2					
told	2	told	2					
you	2	you	2					
caesar	2	was	1					
was	2	was	2					
ambitious	2	with	2					

Figure 3: source : Manning et al. 2008

- Index inversé : point d'entrée par les mots



References

