

# RECHERCHE D'INFORMATION & TRAITEMENT AUTOMATIQUE DU LANGAGE Cours 1: RI - introduction et indexation

Monday 14<sup>th</sup> February, 2022

Laure Soulier









### Recherche d'Information

- 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

M1 DAC - RITAL 1/32

### Organisation du cours



- RI 1 Indexation / Modèle
- RI 2 PageRank / Evaluation
- RI 3 Deep learning pour la RI
- RI 4 Sujets d'ouverture : RI avec sources de connaissances externes
- RI 5 Intervention Qwant
- TAL Extraction d'information (Xavier Tannier)
- RI 6 Sujet d'ouverture : RI interactive

Outils: Pyterrier, ElasticSearch

#### Evaluation

- Contrôle continu 50%
  - En RI : mini-projet (ré-implémentation de papier)
  - En TAL : document compte-rendu de TME

• Examen terminal 50%

M1 DAC - RITAL 2/32

RI - Introduction



#### Définition

"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...

M1 DAC - RITAL 3/32

#### 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

M1 DAC - RITAL 4/32



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

M1 DAC - RITAL 5/32

### Des enjeux...



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

- ...

M1 DAC - RITAL 6/32

### ... Aux systèmes de RI



Problèmes de base pour construire un système d'accès à l'information

- Acquisition (e.g. crawling et prétraitement (diversité des types de documents))
- Représentation indexation (non structuré (texte, image), semi structuré e.g. video)
- Modèle de recherche (présenter des informations pertinentes à l'utilisateur, e.g. liste ordonnée selon un critère)
- Interaction utilisateur (feedback, recherche interactive, la RI est un processus centré utilisateur)
- Evaluation (Protocole d'évaluation, e.g. Cranfield, mesures e.g. rappel-précision)
- Et puis pour les données du web : Dynamicité, Performance, Passage à l'échelle (quantité de données (tera), stockage distribué), ...

M1 DAC - RITAL 7/32

# Exemples de tâches en RI classique



- RI ad-hoc : Trouver parmi un ensemble d'articles ceux qui concernent un sujet spécifique : pertinence d'un document ?
- Faire un résumé du contenu d'un document ou d'un ensemble de documents (éventuellement sur un sujet)
- Structuration (classification) automatique d'un ensemble de documents (groupes)
- Trouver dans un document les passages pertinents, les informations pertinentes concernant un sujet (mots - phrases)
- Suivre dans une collection d'articles l'évolution d'un sujet, Changements de sujets
- Guetter l'arrivée d'informations (appels d'offre, CFP, nouveaux produits, ...)
- Dialoguer avec les clients (e.g. Hot Line, réclamations, ...) ....

M1 DAC - RITAL 8/32

#### Campagnes d'évaluation TREC



- Centre Track: The overall goal of the track is to develop and tune a reproducibility evaluation protocol for IR.
- Common Core Track: The Common Core track uses an ad hoc search task over news documents.
   As such, it serves as a common task for a wide spectrum of IR researchers to attract a diverse run set that can be used to investigate new methodologies for test collection construction.
- Complex Answer Retrieval Track: The focus is on developing systems that are capable of answering complex information needs by collating relevant information from an entire corpus.
- Incident Streams Track: This is a new track for TREC 2018. The Incident Streams track is
  designed to bring together academia and industry to research technologies to automatically
  process social media streams during emergency situations with the aim of categorizing information
  and aid requests made on social media for emergency service operators.
- News Track: The News Track is a new track for 2018. It will feature modern search tasks in the
  news domain. In partnership with The Washington Post, we will develop test collections that
  support the search needs of news readers and news writers in the current news environment.
- Precision Medicine Track: This track is a specialization of the Clinical Decision Support track of
  previous TRECs. It will focus on building systems that use data (e.g., a patient's past medical
  history and genomic information) to link oncology patients to clinical trials for new treatments as
  well as evidence-based literature to identify the most effective existing
- Real-Time Summarization Track: The Real-Time Summarization (RTS) track explores techniques for constructing real-time update summaries from social media streams in response to users' information needs.

M1 DAC - RITAL 9/32

# Campagnes d'évaluation CLEF



- 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.

M1 DAC - RITAL 10/32

#### 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.

M1 DAC - RITAL 11/32

Schéma général et notions

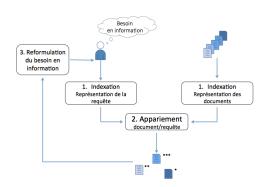


- Document: texte, abstract, passage de texte, texte + structure (e.g. balises HTML : titres, paragraphes, ...)...
- Corpus : ensemble de documents textuels (statique ou dynamique), éventuellement liens entre documents. Taille : 106, 109, ...
- Requête: expression en texte "libre" formulee par l'utilisateur (e.g., "textmining", "je voudrais trouver des documents qui parlent de ...", paragraphes entiers, requetes ambigues « apple », « java », « jaguar »...)
  - Requêtes informationnelles qui ont pour objectif de chercher de l'information en rapport à un sujet donné, sans aucun a priori sur la source d'information (e.g., "gaz à effet de serre").
  - Requêtes navigationnelles qui consistent à atteindre une page web particulière, connue par l'utilisateur (e.g. "Sorbonne université").
  - Requêtes transactionnelles qui souhaitent réaliser des transactions ou bénéficier de services en ligne (e.g., "Avion pas cher Toulouse-Paris").

M1 DAC - RITAL 12/32

#### Schéma général





- Processus en U comporte 3 étapes :
  - Indexation permet d'extraire le contenu d'un document dans un index
  - Appariement permet de mettre en relation la collection de documents, indexée au préalable, avec la requête, également pré-traitée, afin d'identifier les documents pertinents.
  - Reformulation du besoin en information permet de redéfinir le besoin de l'utilisateur au fur et à mesure de la session de recherche.

M1 DAC - RITAL 13/32

# Pertinence système vs. Pertinence utilisateur

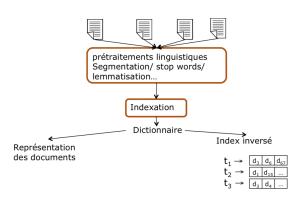


- La pertinence est issue de la mise en correspondance de trois elements :
  - La requete
  - Le document
  - Le besoin utilisateur <u>∧</u> <u>∧</u>dépend de la tache, du contexte, du temps, de la fraicheur, ...
- La pertinence systeme : mesure algorithmique basee sur le calcul de l'adequation entre la representation de la requete et celle de la collection de documents
- La pertinence utilisateur : pertinence subjective que l'utilisateur aurait donne a chacun des documents.

M1 DAC - RITAL 14/32







- 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

M1 DAC - RITAL 15/32

### Lois de distribution des termes pour les collections de documents



#### • Loi de Zipf

- Stipule que la frequence d'occurrence d'un mot est inversement proportionnelle a celle de son rang
- Principe du "moindre effort": plus facile pour un auteur de repeter des mots que d'en chercher des nouveaux (Quelques mots communs representent la plus grande partie des textes (stopwords))
- Le 1er mot est environ 2 fois plus frequent que le 2nd qui est 2 fois plus frequent que le 3e etc...
- Brown Corpus ( > 1 M mots)

Mot	Rang	Occurrence	Fréquence
the	1	69971	7%
of	2	36411	3.5%
and	3	28852	2.6%

M1 DAC - RITAL 16/32

# Lois de distribution des termes pour les collections de documents



• Loi de Zipf - Expression formelle

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

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

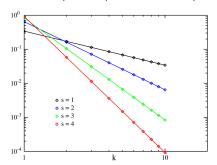


Figure 1: Analyse log fréquence vs. log rang. k = s et N = 10 (source Wikipedia)

M1 DAC - RITAL 17/32

#### Lois de distribution des termes pour les collections de documents



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

- ....

M1 DAC - RITAL 18/32



- 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

$$V = Kn^{\beta} \tag{2}$$

V : taille du vocabulaire, N : taille du texte, K et  $\beta$  : paramètres dépendant du texte (en Anglais : K entre 10 et  $100 - \beta$  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 superieure (noms propres, erreurs de typos, etc.)
  - \* Les nouveaux mots apparaissent moins frequemment quand le vocabulaire croit

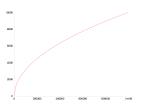
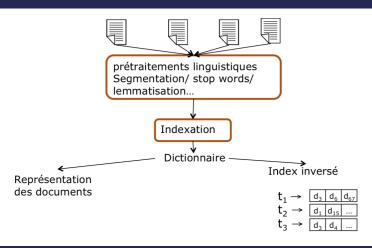


Figure 2: The x-axis: text size, and the y-axis: the number of distinct vocabulary

#### Chaîne d'indexation





#### Deux types d'index

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

M1 DAC - RITAL 20/32



- Identification des segments: mots simples (unigrammes), mots composés (bi-grammes), segments de 3 mots (tri-grammes), ..., N-grammes
- Analyse lexicale (segmentation tokenisation)
- Conversion du texte en un ensemble de termes
  - Unite lexicale ou radical
  - Espaces, chiffres, ponctuations, etc
  - Depend de la specificite des langues traitees
    - e.g. langues asiatiques (pas de signe de separation) vs indo-europeennes
    - Meme pour les langues d'une meme famille, nombreuses specificites
      - e.g. aujourd'hui constitue un seul mot :
        Arbeiterunfallversicherungsgesetz (33 lettres) = Loi sur
      - l'assurance des accidents du travail
  - Logiciel « TreeTagger » pour les langues indo-europeennes

M1 DAC - RITAL 21/32



- Quelles unités conserver pour l'indexation ? stopword/anti-dictionnaires
  - Les mots les plus frequents de la langue "stop words" n'apportent pas d'information utile e.g. prepositions, pronoms, mots « athematiques »,.. (peut representer jusqu'a 30 ou 50
  - Ces "stop words" peuvent etre dependants d'un domaine ou pas L'ensemble des mots elimines est conserve dans un anti-dictionnaire (e.g. 500 mots).
  - Les mots les plus frequents ou les plus rares dans un corpus (frequency cut-off)
  - Les connaissances semantiques permettent egalement d'eliminer des mots
  - Techniques de selection de caracteristiques

M1 DAC - RITAL 22/32



# Stopword list

а	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
	~~	_

M1 DAC - RITAL 23/32

# 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

M1 DAC - RITAL 24/32



- 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

 $- s \rightarrow : cats \rightarrow cat$ 

M1 DAC - RITAL 25/32

- 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 representer au mieux le contenu d'un document ?
  - Considerations statistiques, parfois linguistiques
  - Loi de Zipf : elimination des termes trop frequents ou trop rares
  - Facteurs de ponderation
    - \* e.g.tf (ponderation locale), idf (ponderation globale)
    - \* Normalisation:prise en compte de la longueur des documents, etc

M1 DAC - RITAL 26/32

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



- Term Frequency tf(t<sub>i</sub>, d): nombre occurrences de t<sub>i</sub> 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 ferequency idf

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

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

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

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

M1 DAC - RITAL 27/32



• Index : représentation simple des documents

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

M1 DAC - RITAL 28/32

#### 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);

M1 DAC - RITAL 29/32

ambitious

with

#### • Index inversé : point d'entrée par les mots

Doc 1

	i' the Capitol; Brutus killed me.			hath told you Caesar was ambitious:		
term	docID	term	docID			
I	1	ambitiou	s 2	term doc. freq.	$\rightarrow$	postings lists
did	1	be	2		$\rightarrow$	
enact	1	brutus	1	ambitious 1	$\rightarrow$	2
julius	1	brutus	2	be 1	$\rightarrow$	2
caesar	1	capitol	1	brutus 2	$\rightarrow$	1 - 2
I	1	caesar	1	capitol 1	$\rightarrow$	1
was	1	caesar	2	caesar 2	$\rightarrow$	1 → 2
killed	1	caesar	2	did 1	<i>→</i>	1
i'	1	did	1			-
the	1	enact	1	enact 1	$\rightarrow$	1
capitol	1	hath	1	hath 1	$\rightarrow$	2
brutus	1	I	1	I 1	$\rightarrow$	1
killed	1	I	1	i' 1	$\rightarrow$	1
me	1 _	→ i′	1	it 1	$\rightarrow$	2
so	2	it	2	iulius 1	<i>-</i>	1
let	2	julius	1	,		-
it	2	killed	1	killed 1	$\rightarrow$	1
be	2	killed	1	let 1	$\rightarrow$	2
with	2	let	2	me 1	$\rightarrow$	1
caesar	2	me	1	noble 1	$\rightarrow$	2
the	2	noble	2	so 1	-	2
noble	2	so	2			
brutus	2	the	1		$\rightarrow$	
hath	2	the	2	told 1	$\rightarrow$	2
told	2	told	2	you 1	$\rightarrow$	2
you	2	you	2	was 2	$\rightarrow$	1 → 2
caesar	2	was	1	with 1	$\rightarrow$	2
was	2	was	2		,	-

Doc 2 I did enact Julius Caesar: I was killed So let it be with Caesar. The noble Brutus

Figure 3: source: Manning et al. 2008

M1 DAC - RITAL 30/32



• Index inversé : point d'entrée par les mots

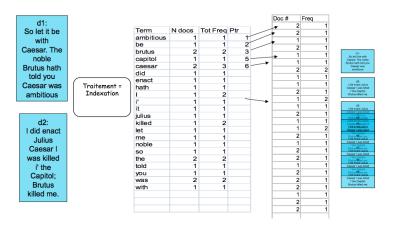


Figure 4: source: Mohand Boughanem

M1 DAC - RITAL 31/32



- Ressources pédagogiques
  - Christopher D. Manning, Prabhakar Raghavan and Hinrich Schutze, Introduction to Information Retrieval, Cambridge University Press. 2008
  - Massih-Reza Amini, Eric Gaussier, Recherche d'information, Applications, modeles et algorithmes, Eyrolles 2013-2018
  - W. Bruce Croft, Donald Metzler, Trevor Strohman, Search Engines Information Retrieval in Practice, Addison Wesley, 2009
- Ressources scientifiques
  - Conférences : SIGIR. CIKM. ECIR. ICTIR. WSDM. WWW. CORIA
  - Journaux : ACM-TOIS, JASIST, IP&M, JIR

M1 DAC - RITAL 32/32