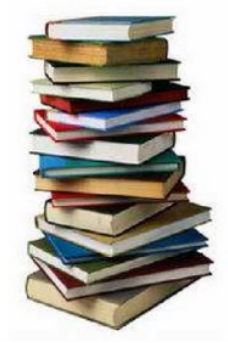
université PARIS-SACLAY

Recherche d'information

Recherche et Extraction d'Information

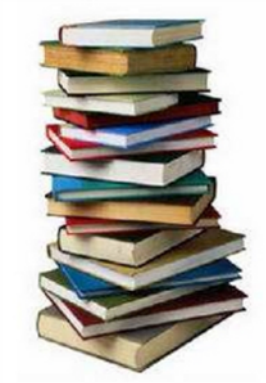


Anne-Laure Ligozat/Xavier Tannier

Rappels des épisodes précédents

Les acteurs de la Recherche d'Information

Collection :
un ensemble de documents



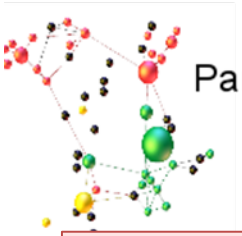
Utilisateur :
un besoin d'information et/ou une tâche à accomplir

Les **systèmes de RI** doivent pouvoir traiter :

- De **grandes masses** d'information
- En **langage naturel** (et créée pour des humains)
- De façon **rapide** et **pertinente**

Indexation
-
Normalisation

Recherche d'Information



Parcours du Web
(robots)

Analyse des
documents

Indexation
(modèle de document)

Collections dynamiques
vs. *statiques*

Modèle de recherche

Fonction de similarité

Évaluation

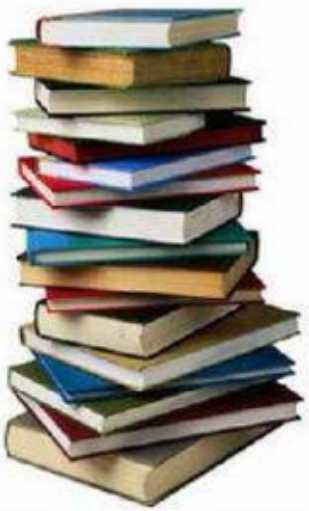
Requête

Analyse du
besoin

Indexation : pourquoi ?

- L'idée principale du moteur de recherche est de retrouver les documents qui « parlent de » la requête.
- On utilise ce qu'on a sous la main : les **mots**
  - Qu'est-ce qu'un mot ?
  - Que faire lorsqu'un mot est « proche » d'un mot de la requête ?
- Le parcours complet de l'ensemble des documents avec les termes d'une requête est impossible : trop de documents et temps de réponse prohibitif.
- On passe par un traitement préalable : *l'indexation* :

  Le but de l'indexation automatique : "*transformer des documents en substituts capables de représenter le contenu de ces documents*" (Salton et McGill, 1983)

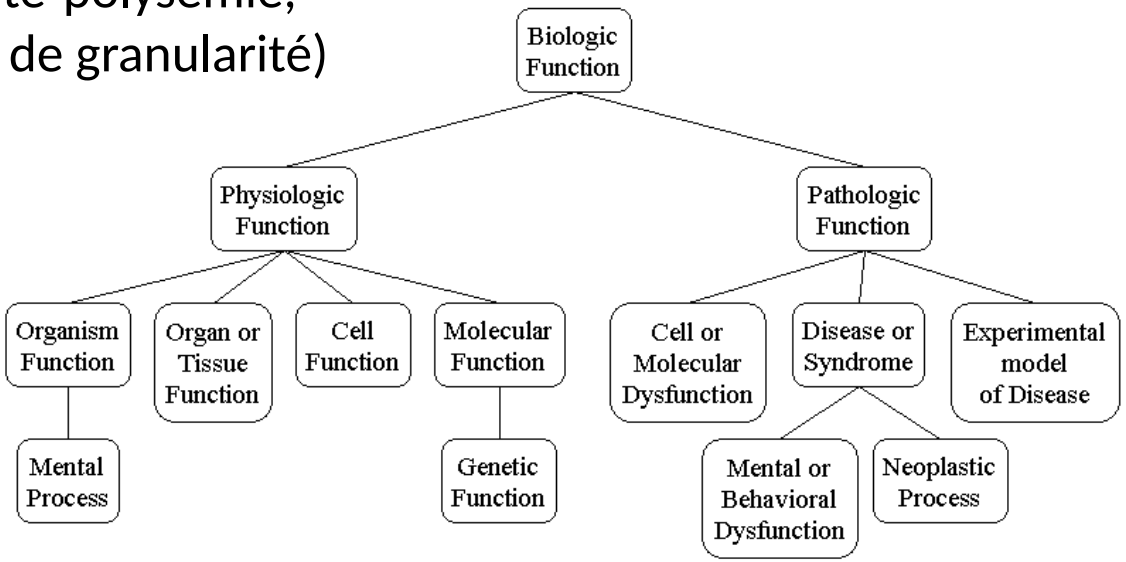


Indexation libre et contrôlée

- Indexation **libre** :
  - Mots, termes des documents

- Indexation **contrôlée**
  - Listes de termes **prédéfinies**
  - Vocabulaire **contrôlé** (évite polysémie, synonymie et problèmes de granularité)
  - Thésaurus

exemple : thésaurus UMLS



Construction de l'index : vue générale

DOCUMENTS

TEXTE

Rien ne sert de courir; il faut partir à point :
Le lièvre et la tortue en sont un témoignage.
«Gageons, dit celle-ci, que vous n'atteindrez point
Sitôt que moi ce but. - Sitôt? Êtes-vous sage ?
Repartit l'animal léger :
Ma commère, il vous faut purger
Avec quatre grains d'ellébore.)
- Sage ou non, je parie encore."
Ainsi fut fait; et de tous deux
On mit près du but les enjeux :
Savoir quoi, ce n'est pas l'affaire,
Ni de quel juge l'on convint.
Notre lièvre n'avait que quatre pas à faire,
J'entends de ceux qu'il fait lorsque, prêt d'être atteint,
Il s'éloigne des chiens, les renvoie aux calendes,
Et leur fait arpenter les landes.
Ayant, dis-je, du temps de reste pour brouter,
Pour dormir et pour écouter
D'où vient le vent, il laisse la tortue
Aller son train de sénateur.
Elle part, elle s'évertue,
Elle se hâte avec lenteur.
Lui cependant méprise une telle victoire,
Tient la gageure à peu de gloire,
Croit qu'il y a de son honneur
De partir tard. Il broute, il se repose,
Il s'amuse à toute autre chose
Qu'à la gageure. A la fin, quand il vit
Que l'autre touchait presque au bout de la carrière,
Il partit comme un trait; mais les élans qu'il fit
Furent vains : la tortue arriva la première.
"Eh bien! lui cria-t-elle, avais-je pas raison ?
De quoi vous sert votre vitesse ?
Moi l'emporter! et que serait-ce
Si vous portiez une maison ?"

TERMES

Rien ne sert de
courir il faut
partir à point

TERMES NORMALISÉS

rien sert
courir faut
partir point

INDEX

Dans quels documents cherche-t-on ?

- **Formats** :
  - HTML (menus, tableaux, publicité, rendu)
  - Texte brut (structure ?)
  - pdf (problèmes d'encodage, rendu)
  - Word (format propriétaire, structure)
  - Excel (gestion des tableaux)
  - OpenOffice (XML)
  - …

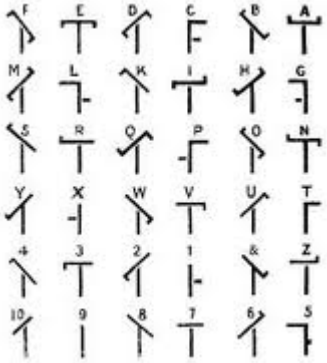
- Il est assez simple de détecter le type d'un document
- Des heuristiques spécifiques à chaque format pour extraire le texte
- Les moteurs de recherche utilisent très rarement la structure des documents

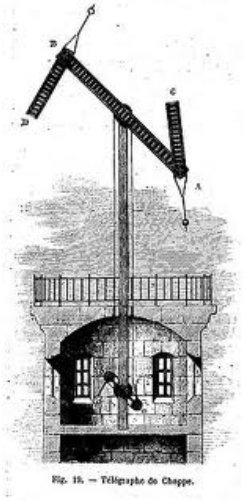
Dans quels documents cherche-t-on ?

- **Langues**
  - Identification de langues, un problème difficile
  - Des documents multilingues
  - De la recherche d'information multilingue

- **Encodages**
  - Des erreurs dans la gestion de l'encodage peuvent conduire à des résultats erronés

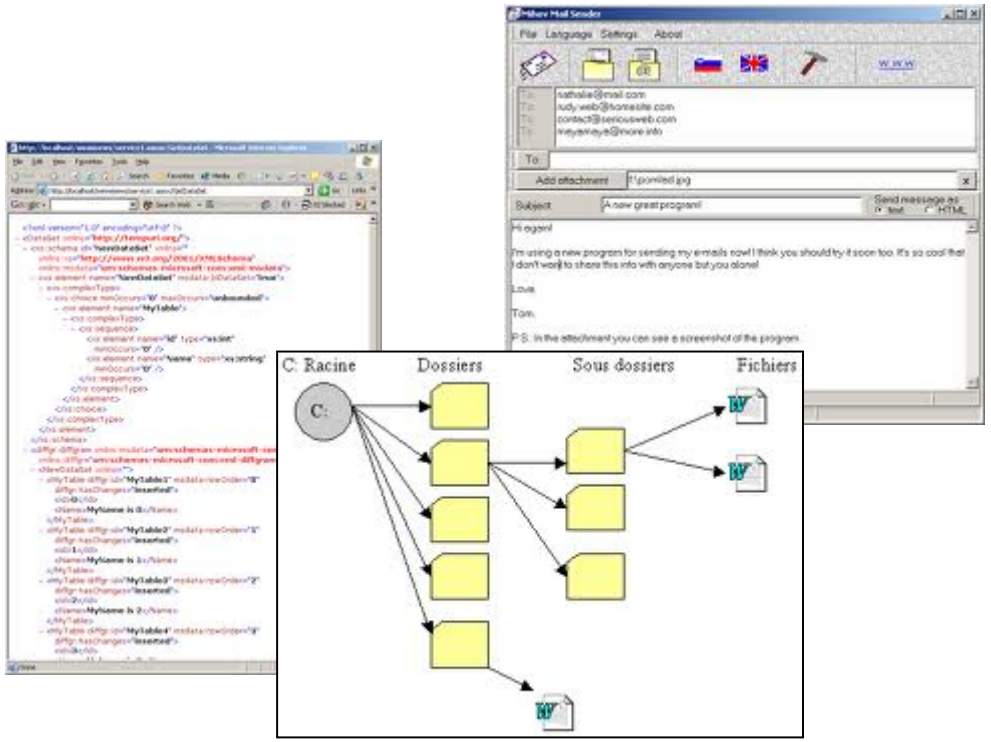
« prÃ©sident du PÃ©rou »





Dans quels documents cherche-t-on ?

- **« Unité » document**
  - Un fichier ?
  - Un e-mail ?
    - Avec ses entêtes ?
    - Avec ses attachements ?
  - Un groupe de fichiers ?
    - Site Web
    - Document en plusieurs fichiers
  - Etc.



Du texte aux termes

DOCUMENTS

TEXTE

Rien ne sert de courir; il faut partir à point :
Le lièvre et la tortue en sont un témoignage.
«Gageons, dit celle-ci, que vous n'atteindrez point
Sitôt que moi ce but. - Sitôt? Êtes-vous sage ?
Repartit l'animal léger :
Ma commère, il vous faut purger
Avec quatre grains d'ellébore.)
- Sage ou non, je parie encore."
Ainsi fut fait; et de tous deux
On mit près du but les enjeux :
Savoir quoi, ce n'est pas l'affaire,
Ni de quel juge l'on convint.
Notre lièvre n'avait que quatre pas à faire,
J'entends de ceux qu'il fait lorsque, prêt d'être atteint,
Il s'éloigne des chiens, les renvoie aux calendes,
Et leur fait arpenter les landes.
Ayant, dis-je, du temps de reste pour brouter,
Pour dormir et pour écouter
D'où vient le vent, il laisse la tortue
Aller son train de sénateur.
Elle part, elle s'évertue,
Elle se hâte avec lenteur.
Lui cependant méprise une telle victoire,
Tient la gageure à peu de gloire,
Croit qu'il y a de son honneur
De partir tard. Il broute, il se repose,
Il s'amuse à toute autre chose
Qu'à la gageure. A la fin, quand il vit
Que l'autre touchait presque au bout de la carrière,
Il partit comme un trait; mais les élans qu'il fit
Furent vains : la tortue arriva la première.
"Eh bien! lui cria-t-elle, avais-je pas raison ?
De quoi vous sert votre vitesse ?
Moi l'emporter! et que serait-ce
Si vous portiez une maison ?"

TERMES

Rien ne sert de
courir il faut
partir à point

aujourd'hui
d'un
S.N.C.F.
Le Mans
14/07/1789
…

TERMES NORMALISÉS

rien sert
courir faut
partir point

INDEX

La segmentation

- Identification des unités élémentaires (phonèmes, morphèmes, mots, etc.). Pour l'écrit, des mots et des phrases.
- Un problème très complexe dans certaines langues (chinois…)
- L'étape initiale indispensable pour tout travail sur le texte
- On obtient des *mots*, ou des *termes*, ou des *tokens*

- Ces unités seront les candidats à l'indexation et à la recherche dans une requête

La segmentation

- Dans les langues "européennes" :
  - Les délimiteurs de mots et de phrases peuvent être ambigus
    - *T.A.L.*, *www.sncf.com*, *l'illusion*, *aujourd'hui*, *Jean-Louis*, *donne-t-il*, *1914-1918*
  - Les mots (noms propres en particulier) peuvent avoir des variantes :
    - *Etats-Unis = États-Unis*, *France Inter = France-Inter*
  - Même l'espace n'est pas toujours un bon délimiteur
    - San Francisco ?, « Ni putes ni soumises » ?
  - Les nombres, les dates
    - 14/07/1789, Mardi 12 mars, B-52, (+33) 6 45 65 13 95
  - Les anciens systèmes de RI retiraient tout simplement les nombres
  - Toujours source de beaucoup d'erreurs dans les systèmes de RI modernes

- Les langues agglutinantes
  - ***Lebensversicherungsgesellschaftsangestellter*** (employé d'une compagnie d'assurance-vie)
  - Un segmenteur de mots composés est alors utile

La segmentation

- En Japonais, Chinois, etc. il n'y a pas d'espace entre les mots
  - 學而不思則罔，思而不學則殆。
  - La segmentation n'est pas toujours unique

- En Japonais, Coréen, on manipule plusieurs types d'alphabets !

  東京、マドリード、イスタンブール（トルコ）が争う２０２０年夏季五輪の開催地は７日（日本時間８日）、ブエノスアイレスでの国際オリンピック委員会（ＩＯＣ）総会で、ＩＯＣ委員約１００人の投票で決まる。

- En Arabe ou en Hébreu, on écrit de droite à gauche, mais certains éléments sont écrits de gauche à droite

  يرتقب توزيع ما مجموعه 3026 وحدة سكنية جديدة موجهة لامتصاص السكن الهش عبر ولاية عنابة وذلك قبل نهاية السنة الجارية 2013

Normalisation

DOCUMENTS

TEXTE

Rien ne sert de courir; il faut partir à point :
Le lièvre et la tortue en sont un témoignage.
«Gageons, dit celle-ci, que vous n'atteindrez point
Sitôt que moi ce but. - Sitôt? Êtes-vous sage ?
Repartit l'animal léger :
Ma commère, il vous faut purger
Avec quatre grains d'ellébore.)
- Sage ou non, je parie encore."
Ainsi fut fait; et de tous deux
On mit près du but les enjeux :
Savoir quoi, ce n'est pas l'affaire,
Ni de quel juge l'on convint.
Notre lièvre n'avait que quatre pas à faire,
J'entends de ceux qu'il fait lorsque, prêt d'être atteint,
Il s'éloigne des chiens, les renvoie aux calendes,
Et leur fait arpenter les landes.
Ayant, dis-je, du temps de reste pour brouter,
Pour dormir et pour écouter
D'où vient le vent, il laisse la tortue
Aller son train de sénateur.
Elle part, elle s'évertue,
Elle se hâte avec lenteur.
Lui cependant méprise une telle victoire,
Tient la gageure à peu de gloire,
Croit qu'il y a de son honneur
De partir tard. Il broute, il se repose,
Il s'amuse à toute autre chose
Qu'à la gageure. A la fin, quand il vit
Que l'autre touchait presque au bout de la carrière,
Il partit comme un trait; mais les élans qu'il fit
Furent vains : la tortue arriva la première.
"Eh bien! lui cria-t-elle, avais-je pas raison ?
De quoi vous sert votre vitesse ?
Moi l'emporter! et que serait-ce
Si vous portiez une maison ?"

TERMES

Rien ne sert de
courir il faut
partir à point

aujourd'hui d'un S.N.C.F. Le Mans 14/07/1789 …

INDEX

TERMES NORMALISÉS

rien sert
courir faut
partir point

*Mots **vides***

- Les mots « **outils** » n'apportent pas de sens au texte

  déterminants : « le », « la », pronoms : « je », « nous »,
  prépositions : « sur », « contre », …
- Ce sont les mots les plus fréquents de la langue
  - Les 30 mots les plus fréquents représentent 30 % des occurrences de mots
  - Les supprimer permet d'économiser beaucoup de place dans l'index
- Mais :
  - On en a besoin pour des requêtes multi-termes

    « pomme de terre », « les Chevaliers du Zodiaque »
  - Ils sont parfois porteurs de sens dans des cas particuliers

    « Let it be », « The Who », « ça », « être ou ne pas être »
  - La compression permet finalement de conserver les mots vides dans peu d'espace (nous verrons cela plus tard)

Normalisation de mots « identiques »

- Dans les documents comme dans la requête
- On veut par exemple normaliser :
  - « U.S.A. » et « USA » → USA
  - « morpho-syntaxe » et « morphosyntaxe » → morphosyntaxe
  - « Tuebingen », « Tübingen » et « Tubingen » → Tubingen
  - « Gorbatchov » et « Gorbatchev » → Gorbatchev
- Mais pas :
  - « sur » et « sûr »,
  - « pêche » et « péché »
  - En allemand, « mit » (avec) et « MIT »
  - En anglais, « C.A.T. » (Caterpillar) et « cat »

- Sans oublier les fautes de frappe / d'orthographe (voir plus tard)

Formes d'un mot, famille d'un mot

- **Flexion**
  - Verbale : *montrer, montreras…*
  - Nominale : *cheval, chevaux…*
  - forme canonique (lemme) et formes fléchies
- **Dérivation**
  - *penser/V + able = pensable*
  - *in + pensable/A = impensable*
  - base et dérivé
- **Composition**
  - *appendice + ectomie = appendicectomie*
  - éléments de formation, mot composé

La morphosyntaxe

- Des analyses différentes pour des besoins différents :
  - **Lemmatisation**
    - obtention de la forme canonique (*chevaux* → *cheval*)
    - long à mettre en œuvre et dépendant de la langue
    - pour rechercher/extraire de l'information, accéder au sens d'un lemme
  - **Racinisation** (*stemming*)
    - obtention de la racine (*chevaux*, *chevalier* → *cheva*)
    - pour agréger les dérivations morphologiques à peu de frais
  - **Étiquetage**
    - catégorisation morphosyntaxique (cheval → nom commun)
    - pour appliquer des techniques de TAL sur les catégories grammaticales (suppression mots vides, désambiguïsation (*or*), termes)
- Des techniques assez bien maîtrisées : un pourcentage d'erreurs faible mais difficilement compressible

Racinication : algorithme de Porter

- 5 phases de réduction par règles (pour l'anglais, adapté ensuite au français)
- Si deux règles de réduction s'appliquent, on choisit celle qui supprime le plus long suffixe
    - *sses* → *ss*
    - *ies* → *i*
    - *ational* → *ate*
    - *tional* → *tion*
    - *Si m > 1 alors cement* → ""

      replacement → *replac*

      cement → *cement*

Indexation

DOCUMENTS

TEXTE

Rien ne sert de courir; il faut partir à point :
Le lièvre et la tortue en sont un témoignage.
«Gageons, dit celle-ci, que vous n'atteindrez point
Sitôt que moi ce but. - Sitôt? Êtes-vous sage ?
Repartit l'animal léger :
Ma commère, il vous faut purger
Avec quatre grains d'ellébore.)
- Sage ou non, je parie encore."
Ainsi fut fait; et de tous deux
On mit près du but les enjeux :
Savoir quoi, ce n'est pas l'affaire,
Ni de quel juge l'on convint.
Notre lièvre n'avait que quatre pas à faire,
J'entends de ceux qu'il fait lorsque, prêt d'être atteint,
Il s'éloigne des chiens, les renvoie aux calendes,
Et leur fait arpenter les landes.
Ayant, dis-je, du temps de reste pour brouter,
Pour dormir et pour écouter
D'où vient le vent, il laisse la tortue
Aller son train de sénateur.
Elle part, elle s'évertue,
Elle se hâte avec lenteur.
Lui cependant méprise une telle victoire,
Tient la gageure à peu de gloire,
Croit qu'il y a de son honneur
De partir tard. Il broute, il se repose,
Il s'amuse à toute autre chose
Qu'à la gageure. A la fin, quand il vit
Que l'autre touchait presque au bout de la carrière,
Il partit comme un trait; mais les élans qu'il fit
Furent vains : la tortue arriva la première.
"Eh bien! lui cria-t-elle, avais-je pas raison ?
De quoi vous sert votre vitesse ?
Moi l'emporter! et que serait-ce
Si vous portiez une maison ?"

TERMES

Rien ne sert de
courir il faut
partir à point

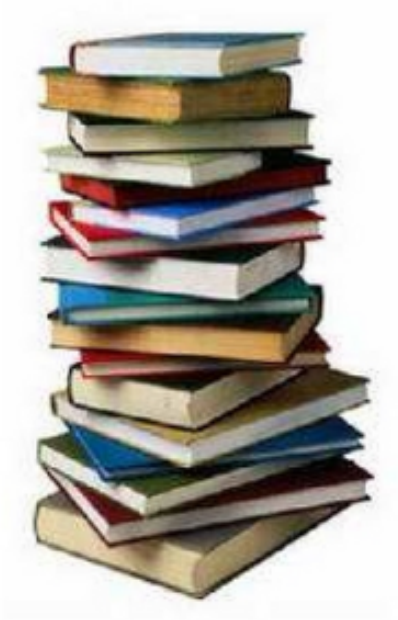
aujourd'hui
d'un
S.N.C.F.
Le Mans
14/07/1789
…

TERMES NORMALISÉS

rien sert
courir faut
partir point

INDEX

Matrice d'incidence



	Antoine & Cléopâtre	Jules César	La Tempête	Hamlet	Othello	Macbeth
Antoine	1	1	0	0	0	0
Brutus	1	1	0	1	0	0
César	1	1	0	1	1	1
Calpurnia	0	1	0	0	0	0
Cléopâtre	1	0	0	0	0	0
pitié	1	0	1	1	1	1
pire	1	0	1	1	1	0

Brutus ET Cléopâtre ET PAS Calpurnia

	Antoine & Cléopâtre	Jules César	La Tempête	Hamlet	Othello	Macbeth
Antoine	1	1	0	0	0	0
Brutus	1	1	0	1	0	0
César	1	1	0	1	1	1
Calpurnia	0	1	0	0	0	0
Cléopâtre	1	0	0	0	0	0
pitié	1	0	1	1	1	1
pire	1	0	1	1	1	0

Vecteurs d'incidence

¬Calpurnia	1	0	1	1	1	1

ET "bit à bit"

1	0	0	0	0	0

Matrice d'incidence

- On ne peut pas utiliser une telle matrice d'incidence en pratique

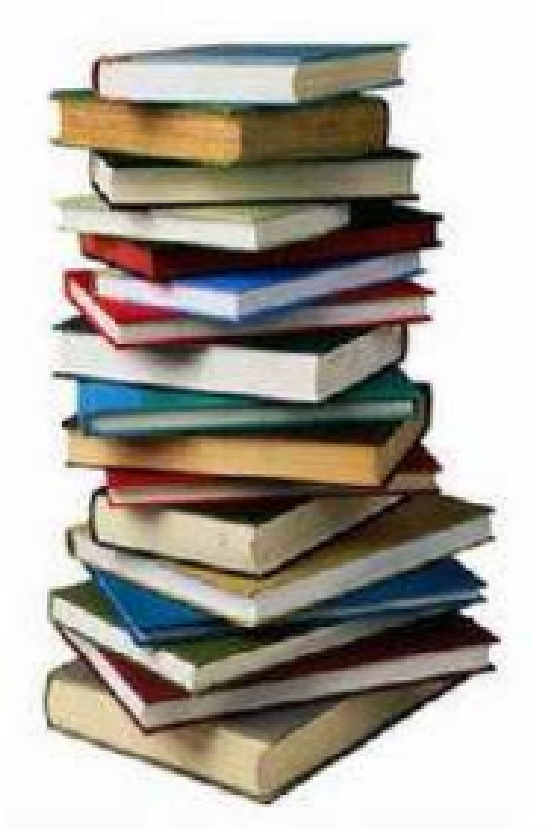
Pourquoi ?

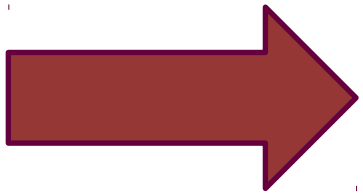
Indices

- Considérons une collection **d'un million** de documents
- Environ **1000** mots par document en moyenne
- Un vocabulaire total **de 500 000** mots distincts.

- Combien y a-t-il de cases dans la matrice ?
- Combien de 1 ?
- Combien de 0 ?

Fichier inverse





Index

Indexation : le fichier inverse

- Notion "classique" de l'index
- Un fichier inverse associe des index aux documents qui les contiennent. Chaque document possède un identifiant unique.
  - a ► d1, d2, d3, d4, d5…
  - à ► d1, d2, d3, d4, d5…
  - abaissa ► d3, d4…
  - abaissable ► d5
  - abandon ► d1, d5
  - abandonna ► d2
  - abasourdi ► d1
  - …

- Quelle structure de données pour cet index ?
- Que se passe-t-il si on ajoute le mot « *abandon* » au document d3 ?

Sac de mots

- Modèles « **sac de mots** » pour l'indexation et la recherche :
  - On oublie l'ordre des mots

    (« *Jean est plus rapide que Marie* » = « *Marie est plus rapide que Jean* »)
  - On raisonne en termes de présence / absence des termes dans un document, ou en terme de fréquence de ces termes

Pondération des termes

Taille du vocabulaire

- Le vocabulaire grandit quand la collection grandit. Pourquoi ?

- **Loi de Heaps** : $M = kT^b$
  - M : taille du vocabulaire
  - T : nombre de tokens dans la collection
  - b et k : constantes (typiquement, $b = 0{,}5$ et $k = 30$ à 100)
  - Loi empirique

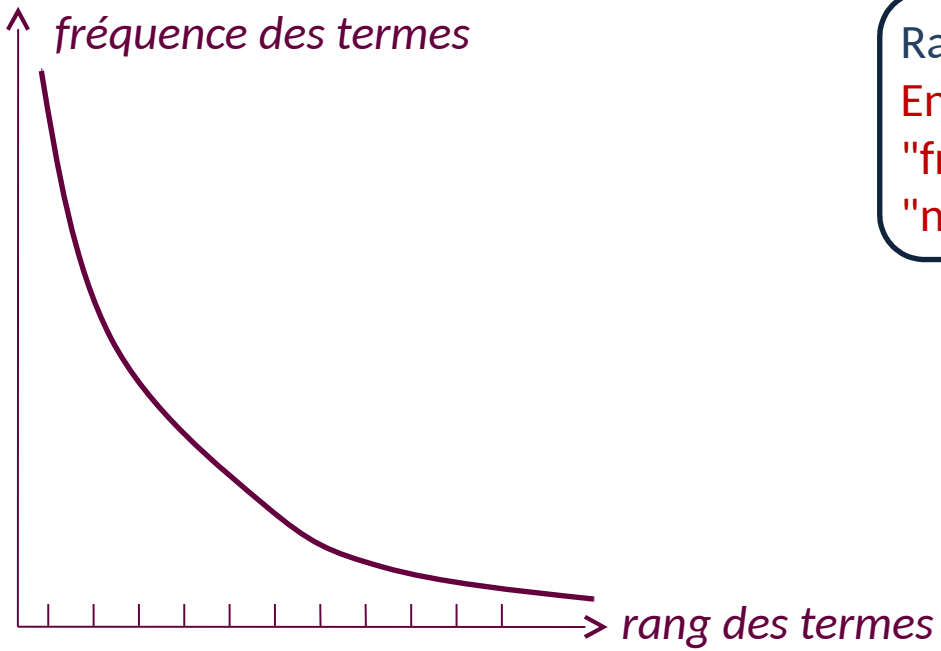
- Et c'est bien pire pour le Web !

Pourquoi ?

Fréquence des termes

- Peu de mots fréquents, et beaucoup de mots rares

- **Loi de Zipf** : le $n^{ème}$ mot le plus fréquent a une fréquence proportionnelle à $1/n$

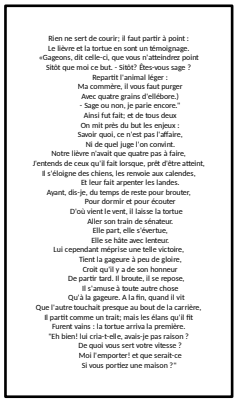


Rappel
En RI,
"fréquence" =
"nb d'occurrences"

Le tf

- Dans une requête comme dans un document, les termes n'ont pas tous la même importance

- **Intuition #1** : plus un document contient d'occurrences d'un terme, plus il est "à propos" de ce terme
(plus il sera pertinent par rapport
à une requête contenant ce terme)


Rien ne sert de courir; il faut partir à point :
Le lièvre et la tortue en sont un témoignage.
«Gageons, dit celle-ci, que vous n'atteindrez point
Sitôt que moi ce but. - Sitôt? Êtes-vous sage ?
Repartit l'animal léger :
Ma commère, il vous faut purger
Avec quatre grains d'ellébore.)
- Sage ou non, je parie encore."
Ainsi fut fait; et de tous deux
On mit près du but les enjeux :
Savoir quoi, ce n'est pas l'affaire,
Ni de quel juge l'on convint.
Notre lièvre n'avait que quatre pas à faire,
J'entends de ceux qu'il fait lorsque, prêt d'être atteint,
Il s'éloigne des chiens, les renvoie aux calendes,
Et leur fait arpenter les landes.
Ayant, dis-je, du temps de reste pour brouter,
Pour dormir et pour écouter
D'où vient le vent, il laisse la tortue
Aller son train de sénateur.
Elle part, elle s'évertue,
Elle se hâte avec lenteur.
Lui cependant méprise une telle victoire,
Tient la gageure à peu de gloire,
Croit qu'il y a de son honneur
De partir tard. Il broute, il se repose,
Il s'amuse à toute autre chose
Qu'à la gageure. A la fin, quand il vit
Que l'autre touchait presque au bout de la carrière,
Il partit comme un trait; mais les élans qu'il fit
Furent vains : la tortue arriva la première.
"Eh bien! lui cria-t-elle, avais-je pas raison ?
De quoi vous sert votre vitesse ?
Moi l'emporter! et que serait-ce
Si vous portiez une maison ?"

$$tf_{t,d} = \text{nombre d'occurrences du terme } t \text{ dans le document } d$$

- On va donc conserver dans l'index le nombre d'occurrences de chaque terme dans le document

La matrice des fréquences

	Antoine & Cléopâtre	Jules César	La Tempête	Hamlet	Othello	Macbeth
Antoine	157	73	0	0	0	0
Brutus	4	157	0	1	0	0
César	232	227	0	2	1	1
Calpurnia	0	10	0	0	0	0
Cléopâtre	57	0	0	0	0	0
pitié	2	0	3	5	5	1
pire	2	0	1	1	1	0

Chaque document est un vecteur v dans $\mathbb{N}^{|v|}$

idf

- **Intuition #2** : des termes très fréquents dans tous les documents ne sont pas si importants (ils sont moins ***discriminants***)
- On compense donc la fréquence des termes dans les documents (tf) en prenant en compte leur fréquence dans la collection (df)

$$df_t = \text{nombre de documents qui contiennent le terme } t$$

$$idf_t = log_{10} \frac{N}{df_t}$$

(N = nombre total de documents)

tf.idf

- Le **poids** d'un terme (*tf.idf*) est la combinaison de ces deux intuitions pour rendre compte du caractère discriminant d'un terme dans un document

$$w_{t,d} = tf_{t,d} \times idf_t$$

$$= tf_{t,d} \times log_{10} \frac{N}{df_t}$$

ou $$w_{t,d} = log\_tf_{t,d} \times log_{10} \frac{N}{df_t}$$

- Le poids d'un terme t :
  - augmente avec sa fréquence dans le document
  - augmente avec sa rareté dans la collection

La matrice des poids

	Antoine & Cléopâtre	Jules César	La Tempête	Hamlet	Othello	Macbeth
Antoine	13,1	11,4	0	0	0	0
Brutus	3,0	8,3	0	1	0	0
César	2,3	2,3	0	0,5	0,3	0,3
Calpurnia	0	11,2	0	0	0	0
Cléopâtre	17,7	0	0	0	0	0
pitié	0,5	0	0,7	0,9	0,9	0,3
pire	1,2	0	0,6	0,6	0,6	0

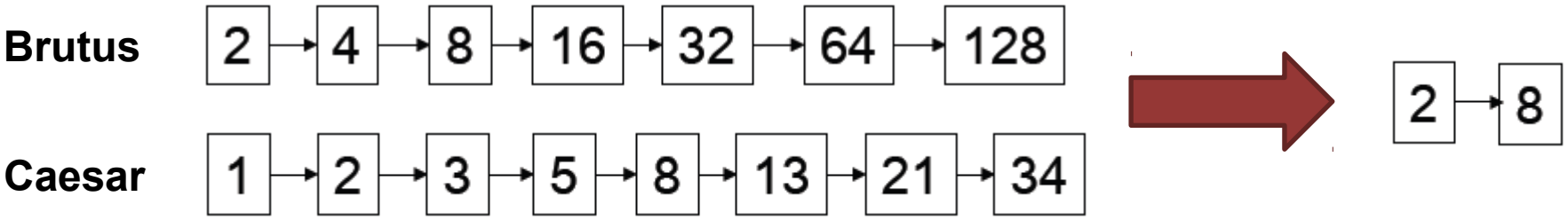
Chaque document est un vecteur v dans $\mathbb{R}^{|v|}$

Recherche dans un index

Retrouver les documents

Brutus AND Caesar

- On **recherche** « **Brutus** » dans le dictionnaire
  → On récupère la liste de documents
- On **recherche** « **Caesar** » dans le dictionnaire
  → On récupère la liste de documents
- On **fusionne** les deux listes



La notion de n-gramme

- ***n-gramme*** : une sous-séquence de *n* éléments extraite d'une séquence donnée.

(cf. modèles de Markov)

Ici, n-grammes de mots

  - uni-gramme : tous les mots
  - bi-gramme : une sous-séquence de 2 éléments
  - etc.

- Différent du groupe de mots d'un point de vue linguistique

- Combien de bi-grammes théoriquement possibles pour *m* mots uniques dans un vocabulaire ?
- Combien de tri-grammes ?
- Jusqu'à quelle valeur de *n* devrait-on aller pour couvrir raisonnablement les besoins d'un utilisateur de moteur de recherche ?

Index de bi-grammes

- Indexer (en plus des mots simples) toutes les paires de termes du texte.

Rien ne sert de	Rien ne ne sert sert de de courir
courir il faut	courir il il faut faut partir
partir à point	partir à à point

Comment éviter d'indexer toutes les paires ?

- On considère donc chaque bi-gramme comme un terme du dictionnaire
- Une requête sur un bi-gramme est immédiate

Index de position

- Idée : dans les listes de documents de l'index, ajouter la position de chaque occurrence de terme dans le document.

terme	fréquence	→	D1	D3	D4

terme	fréquence	→	D1 : pos1, pos2, pos3
			D3 : pos1, pos2
			D4 : pos1, pos2, pos3

Index de position : parcours

« Université Paris Saclay »

- Extraction des entrées du dictionnaires
- Utilisation **récursive** de l'algorithme de fusion, pour les **documents** puis pour les **positions**.
- Mais utiliser une **comparaison incrémentale** au lieu d'une égalité stricte.

université	1252	→	D2 : 546
			D6 : 34, **87**, 145, 243
			D7 : 44, 87, 34
			…
paris	45	→	D2 : 547
			D6 : **88**, 543
			…
saclay	15345	→	D2 : 54, 90
			D6 : **89**
			D4 : 43

Du modèle booléen

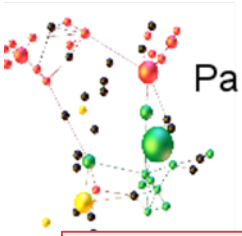
aux modèles à listes de

résultats ordonnés

→ Quels documents retourner et

dans quel ordre ?

Recherche d'Information



Parcours du Web
(robots)

Analyse des
documents

Indexation
(modèle de document)

Collections dynamiques
vs. *statiques*

Modèle de
recherche

Fonction de similarité

Évaluation

Requête

Analyse du
besoin

Modèles de recherche : les trois courants

- Modèles fondés sur la théorie des ensembles
  - ► Modèle **booléen**
- Modèles algébriques
  - ► Modèle **vectoriel**
- Modèles **probabilistes**
  - ► Modélisation de la notion de "pertinence"

- Courants fondés à l'aube de la discipline (années 60, 70)
- Passage à l'échelle : des bases documentaires "jouets" au téraoctet de TREC et au Web

Modèle booléen

- Le premier et le plus simple des modèles
- Basé sur la théorie des ensembles et l'algèbre de Boole
- Les termes de la requête sont soit présents soit absents
  - **Poids binaire** des termes, 0 ou 1
- Un document est soit pertinent soit non pertinent
  - **Pertinence binaire**, et jamais partielle (**modèle *exact***)
- La requête s'exprime avec des **opérateurs logiques**
  - AND, OR, NOT
  - `(cyclisme OR natation) AND NOT dopage`
  - le document est pertinent si et seulement si son contenu respecte la formule logique demandée

Modèle booléen : exemple

Requête Q : **`(cyclisme OR natation) AND NOT dopage`**

Le document contient					Pertinence du document
cyclisme	natation	**cyclisme OR natation**	dopage	**NOT dopage**	
0	0	**0**	0	**1**	**0**
0	0	**0**	1	**0**	**0**
0	1	**1**	0	**1**	**1**
0	1	**1**	1	**0**	**0**
1	0	**1**	0	**1**	**1**
1	0	**1**	1	**0**	**0**
1	1	**1**	0	**1**	**1**
1	1	**1**	1	**0**	**0**

Modèle booléen : avantages et inconvénients

- **Avantages** :
  - Le modèle est transparent et simple à comprendre pour l'utilisateur :
    - Pas de paramètres "cachés"
    - Raison de sélection d'un document claire : il répond à une formule logique
  - Adapté pour les spécialistes (vocabulaire contraint)
- **Inconvénients** :
  - Il est difficile d'exprimer des requêtes longues sous forme booléenne
  - Le critère binaire peu efficace
    - Il est admis que la pondération des termes améliore les résultats
    - cf. modèle booléen étendu
  - Il est impossible d'ordonner les résultats
    - Tous les documents retournés sont sur le même plan
    - L'utilisateur préfère un classement lorsque la liste est grande

Vers des listes ordonnées de résultats

- La plupart des utilisateurs :
  - ont du mal à écrire des requêtes booléennes
  - ne veulent pas parcourir trop de résultats (des milliers, voire des millions++)

↳ On préfère donc des listes ordonnées
  - Du plus utile à l'utilisateur (pertinent) au moins utile
  - Le nombre de résultats n'est plus un problème
  - L'utilisateur en parcourt autant qu'il le souhaite
- *La condition* : avoir un algorithme d'ordonnancement efficace
- Modèle statistique :
  - Aspect **quantitatif** des termes et des documents
  - **Degré** de similarité entre une requête et un document

Modèle vectoriel

Modèle vectoriel

- Mesure de **similarité** : *Plus deux représentations contiennent les mêmes éléments, plus la probabilité qu'elles représentent la même information est élevée.*
- Documents et requête sont représentés par un **vecteur**
  - Les coordonnées du vecteur sont exprimées dans un espace euclidien à *n* dimensions (*n* : nombre de termes)
  - La longueur du vecteur (*i.e.* de sa projection sur chacun des axes/termes) est proportionnelle au poids des termes.
- La pertinence du document correspond au **degré** de similarité entre le vecteur de la requête et celui du document

On ordonne les documents du plus similaire à la requête au moins similaire

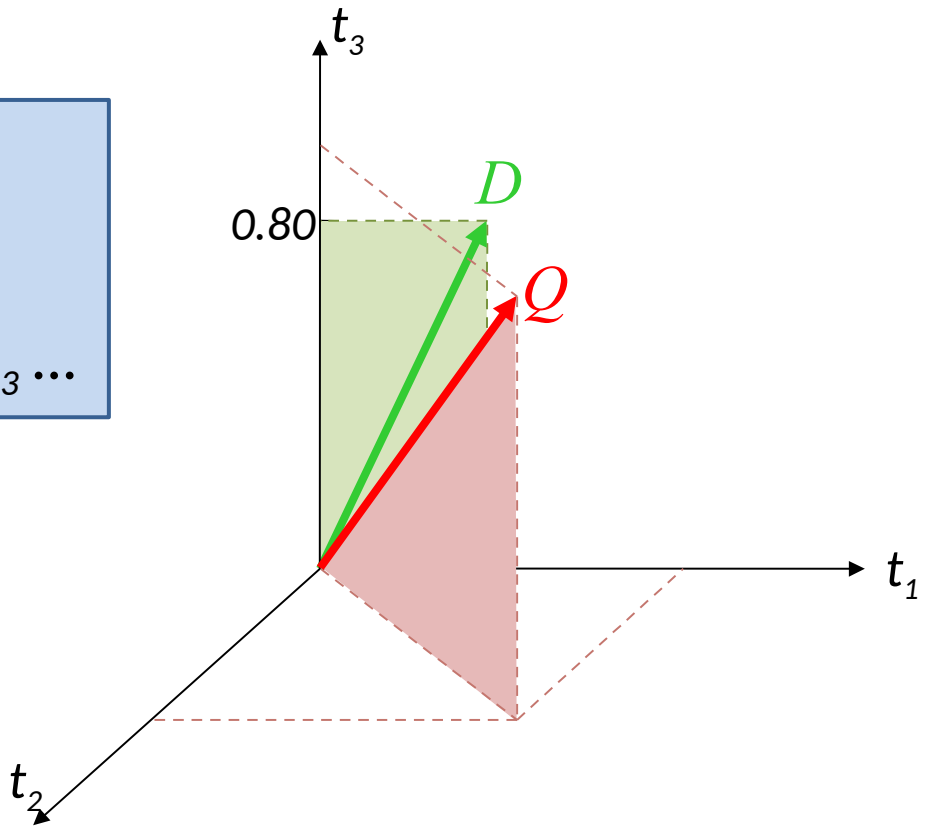
Modèle vectoriel

Requête Q : $t_1\ t_2\ t_3$

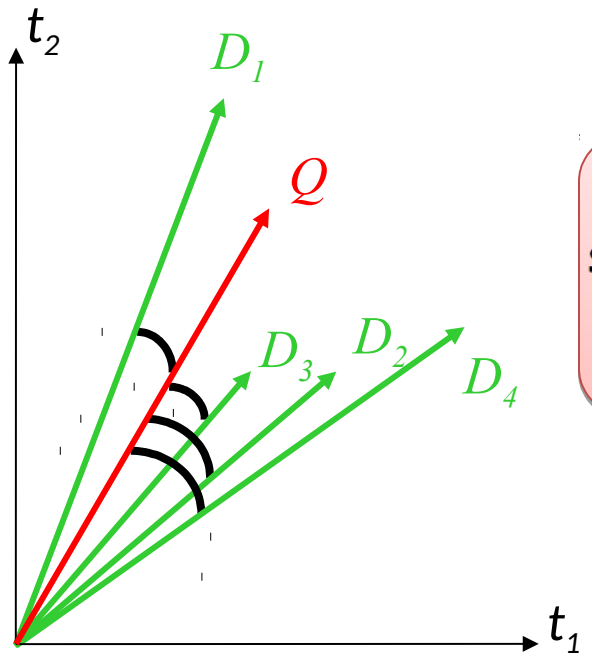
Document D : … t_1 … t_3 …

Poids $w_{D,t1} = 0.45$

Poids $w_{D,t3} = 0.80$



Quelle mesure de similarité ?



Cosinus

$$\text{sim}(\vec{Q},\vec{D}) = \frac{\vec{Q}\cdot\vec{D}}{|\vec{Q}|\times|\vec{D}|} = \frac{\sum_{i=1}^{n} w_{i,Q} \times w_{i,D}}{\sqrt{\sum {w^2}_{i,Q}} \times \sqrt{\sum {w^2}_{i,D}}}$$

(Le produit scalaire avec normalisation de la longueur des vecteurs)

Modèle vectoriel : avantages et inconvénients

- **Avantages** :
  - Le langage de requête est plus simple (liste de mot-clés)
  - Les performances sont meilleures grâce à la pondération des termes
  - Le renvoi de documents à pertinence partielle est possible
  - La fonction d'appariement permet de trier les documents
- **Inconvénients** :
  - Le modèle considère que tous les termes sont indépendants (inconvénient théorique)
  - Le langage de requête est moins expressif
  - L'utilisateur voit moins pourquoi un document lui est renvoyé

Autres modèles

Modèle probabiliste

- Estimation de la <u>probabilité</u> de pertinence d'un document par rapport à une requête
- *Probability Ranking Principle (Robertson 77)*
- R : D est pertinent pour Q
- ¬R : D n'est pas pertinent pour Q

} variables indépendantes, deux ensembles de documents séparés

- Le but : estimer
  - P(R/D) : probabilité que le document D soit contienne de l'information pertinente pour Q
  - P(¬R/D)

$$\text{si } \frac{P(R/\mathrm{D})}{P(\neg R/\mathrm{D})} > 1 \text{ ou si } \log \frac{P(R/\mathrm{D})}{P(\neg R/\mathrm{D})} > 0 \text{ alors D est pertinent}$$

Modèle probabiliste

- Rappel du théorème de **Bayes** :

$$P(A/B) = \frac{P(B/A) \cdot P(A)}{P(B)}$$

- On ne sait pas calculer *P(R/D)*, mais on peut calculer *P(D /R)*

Probabilité d'obtenir *D* en connaissant les pertinents

$$P(R/D) = \frac{P(D/R) \cdot P(R)}{P(D)}$$

Probabilité d'obtenir un document pertinent en piochant au hasard

Probabilité de piocher *D* au hasard

Modèle probabiliste : conclusion

- Deux modèles phares :
  - 2-poisson
  - Okapi
- Autres modèles de type probabiliste :
  - Réseaux bayésiens
  - Modèle de langage

- Conclusion :
  - Problème des probabilités initiales
  - Termes indépendants
  - Résultats comparables à ceux du modèle vectoriel

Autres modèles algébriques

- Modèle vectoriel généralisé
  - Représente les dépendances entre termes
  - Théoriquement intéressant, mais efficacité non démontrée
- *Latent Semantic Indexing*
  - Propose d'étudier les "concepts" plutôt que les termes, car ce sont eux qui relaient les idées d'un texte.
  - Lie les documents entre eux et avec la requête
  - Permet de renvoyer des documents ne contenant aucun mot de la requête
  - Moins de dimensions
- Réseaux de neurones
- …

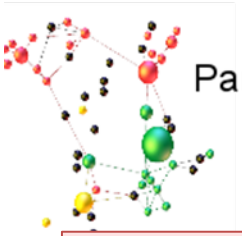
Quelques outils

• lucy/zettair	http://www.seg.rmit.edu.au/zettair/
• cheshire	http://cheshire.lib.berkeley.edu/
• dataparksearch engine	http://www.dataparksearch.org/
• lemur	http://www.lemurproject.org/
• lucene (et solr)	http://jakarta.apache.org/lucene/docs/
• terrier	http://ir.dcs.gla.ac.uk/terrier/
• wumpus	http://www.wumpus-search.org/
• xapian	http://www.xapian.org/

liste et liens sur http://www.emse.fr/~mbeig/IR/tools.html

Évaluation

Recherche d'Information



Parcours du Web
(robots)

Analyse des
documents

Indexation
(modèle de document)

Collections dynamiques
vs. *statiques*

Modèle de
recherche

Fonction de similarité

Évaluation

Requête

Analyse du
besoin

Qu'est-ce qu'un bon moteur de recherche ?

- Il est **rapide** !
  - ↳ Une analyse rapide de la requête
  - ↳ Une recherche rapide dans l'index
  - ↳ Un tri rapide des résultats

- Il est **complet** et **à jour** !
  - Tous les (ou de nombreux) documents de la collection sont traités
  - Les nouveaux documents sont incorporés rapidement aux résultats
  - ↳ Une construction rapide de l'index
  - ↳ (sur le Web) Une découverte permanente, efficace et rapide des nouveaux documents

Qu'est-ce qu'un bon moteur de recherche ?

- Il est **rapide** !
  - ↳ Une analyse rapide de la requête
  - ↳ Une recherche rapide dans l'index
  - ↳ Un tri rapide des résultats

- Il est **complet** et **à jour** !
  - Tous les (ou de nombreux) documents de la collection sont traités
  - Les nouveaux documents sont incorporés rapidement aux résultats
  - ↳ Une construction rapide de l'index
  - ↳ (sur le Web) Une découverte permanente, efficace et rapide des nouveaux documents

Qu'est-ce qu'un bon moteur de recherche ?

- Il est **rapide** !
  - ➥ Une analyse rapide de la requête
  - ➥ Une recherche rapide dans l'index
  - ➥ Un tri rapide des résultats

- Il est **complet** et **à jour** !
  - – Tous les (ou de nombreux) documents de la collection sont traités
  - – Les nouveaux documents sont incorporés rapidement aux résultats
  - ➥ Une construction rapide de l'index
  - ➥ (sur le Web) Une découverte permanente, efficace et rapide des nouveaux documents

Comment mesurer la pertinence ?

- Un moteur sur le Web
  - L'utilisateur clique sur certains liens et pas sur d'autres
  - L'utilisateur retourne sur le moteur
  - L'utilisateur effectue une certaine tâche
- Un site de e-commerce
  - L'utilisateur achète (mais alors, de qui mesure-t-on la satisfaction ?)
  - Il achète vite
  - Une forte proportion de visiteurs achètent
- Un site d'entreprise
  - L'utilisateur gagne en productivité
  - L'accès est sécurisé
  - Etc.

Qu'est-ce qu'une bonne évaluation ?

- Évaluer un système sert à :
  - Savoir s'il remplit la tâche assignée
  - Savoir s'il est meilleur que la concurrence
  - Savoir où on peut l'améliorer
- Il faut donc une évaluation :
  - Reproductible
    - Pour évaluer plusieurs systèmes de la même façon
    - Pour estimer les progrès accomplis
  - Interprétable
    - Pour identifier les zones de progrès possible
  - Rapide
    - Pour pouvoir évaluer chaque modification du système indépendamment
  - Objective

Comment rendre la pertinence objective ?

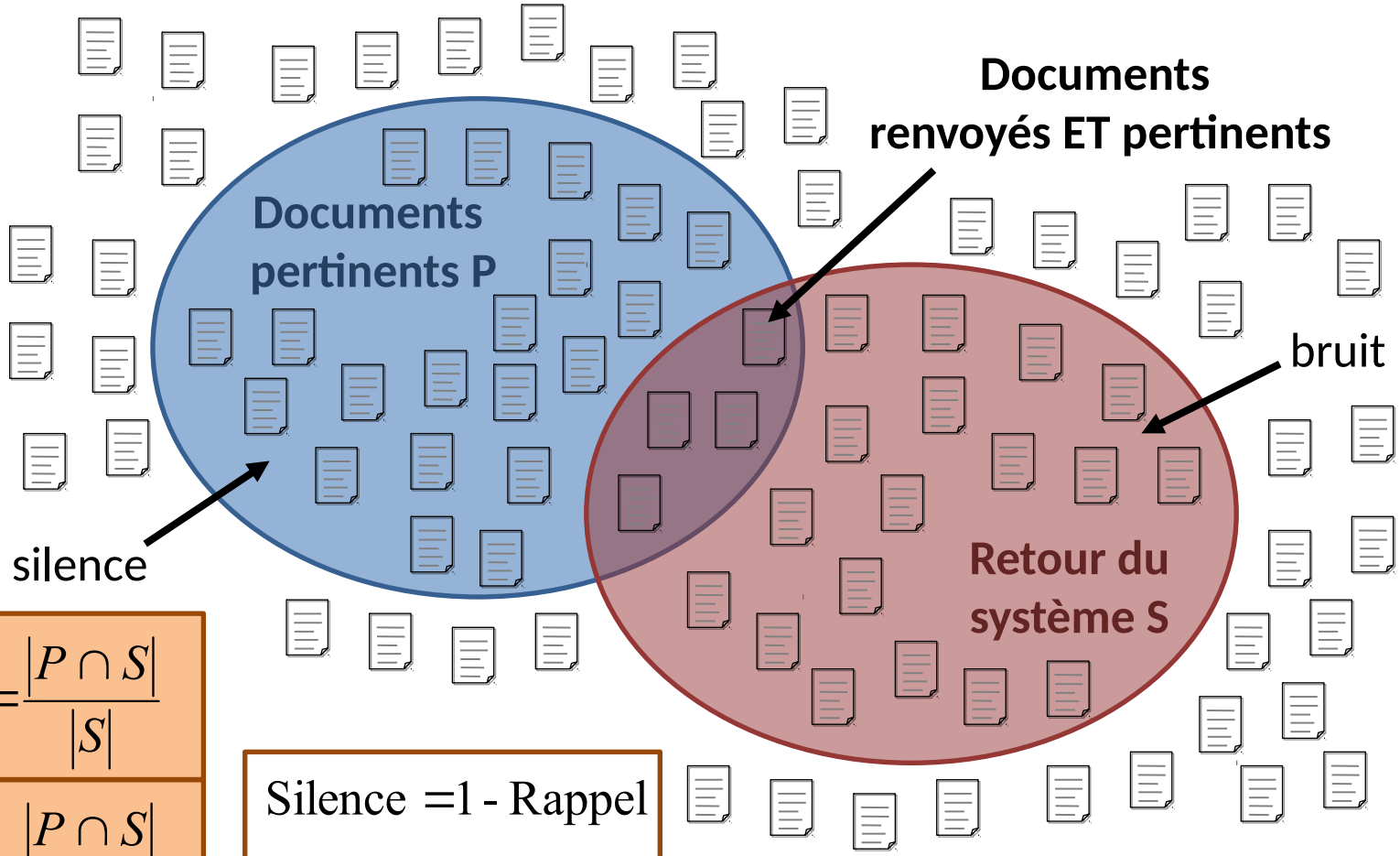
- Rappel :
  - Le besoin de l'utilisateur est d'abord transformé en requête, ce qui comporte déjà une perte d'information.
  - On mesure la pertinence des résultats par rapport au besoin d'information initial, pas par rapport à la requête ! (ex: « java »)
  - Des résultats peuvent être « très pertinents », « pas du tout pertinent », mais aussi « un peu pertinents », « moui » ou « je le savais déjà »
- Pour rendre la pertinence objective :
  - On en simplifie la définition
    - Les documents sont traités indépendamment les uns des autres
    - La pertinence est transformée en notion binaire
  - On utilise des « collections de test »

Collections de test

La collection de test rend les expériences reproductibles

- On met au point un **protocole**
- On juge **manuellement** un nombre significatif d'exemples
  - « Gold standard »
  - Une partie peut également servir d'ensemble de « développement » et/ou d' « apprentissage »
- On calcule un accord inter-annotateurs
  - Pour valider le caractère objectif
- On **compare** les résultats du système aux résultats attendus
- On définit des **mesures** imparfaites mais précises

*Évaluation : **précision et rappel***



$$\text{Précision} = \frac{|P \cap S|}{|S|}$$

$$\text{Rappel} = \frac{|P \cap S|}{|P|}$$

$$\text{Silence} = 1 - \text{Rappel}$$

$$\text{Bruit} = 1 - \text{Précision}$$

Précision et rappel

- Pourquoi pas juste la précision ?
  - La précision évalue la capacité d'un système à renvoyer **SURTOUT** des documents pertinents
  - Renvoyer un seul document pertinent suffit à obtenir 100 % de précision

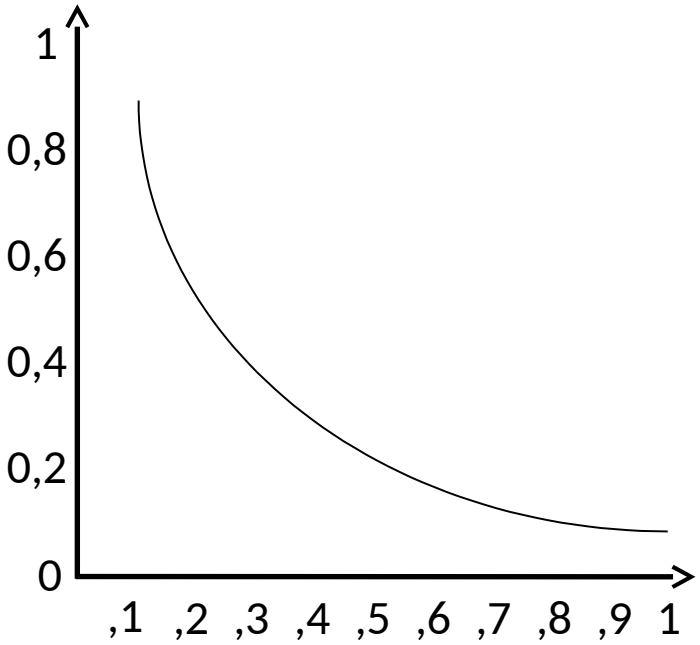
    → *Ce n'est pas compatible avec la satisfaction de l'utilisateur !*

- Pourquoi pas juste le rappel ?
  - Le rappel évalue la capacité d'un système à renvoyer **TOUS** les documents pertinents
  - Renvoyer tous les documents de la collection permet d'obtenir 100 % de rappel

  → *Ce n'est pas compatible avec la satisfaction de l'utilisateur !*

*Courbe **rappel/précision***

- Le rappel augmente bien sûr avec le nombre de réponses
- La précision diminue (en général)
- On utilise la **courbe rappel/précision** pour caractériser les systèmes de recherche d'information



Évaluation : F-mesure

- Pour obtenir une valeur unique entre 0 et 1, on utilise la **F-mesure** (moyenne harmonique)

$$F = \frac{1}{\alpha \frac{1}{p} + (1 - \alpha)\frac{1}{R}} = \frac{(\beta^2 + 1) \times P \times R}{\beta^2 P + R} \qquad \text{avec} \quad \alpha = \frac{1}{\beta^2 + 1}$$

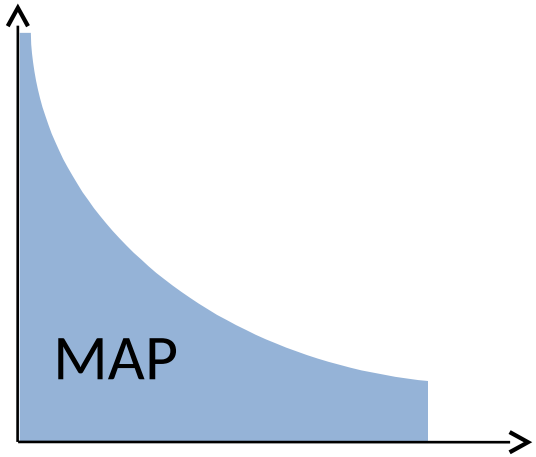
- Pour donner autant d'importance à la précision qu'au rappel, on choisit $\beta = 1$

$$F = \frac{2P.R}{P + R}$$

- $\beta < 1$ favorise la précision, $\beta > 1$ favorise le rappel

Évaluation : autres mesures

- **MAP** (*Mean Average Precision*) : *aire sous la courbe R/P*
- **P@5**, **P@10** : précision après 10 documents retrouvés favorise la haute/très haute précision
- **P@100**, …
- **Taux d'erreur** = (faux positifs + faux négatifs) / pertinents
- et de nombreuses autres…



Références utiles

- Livre « Recherche d'information - Applications, modèles et algorithmes - Data mining, décisionnel et big data » de Massih-Reza Amini et Eric Gaussier (2e édition de 2017)
- Livre « Introduction to Information Retrieval » de Christopher D. Manning, Prabhakar Raghavan et Hinrich Schütze (2009)
  - https://nlp.stanford.edu/IR-book/
- Livre « Modern Information Retrieval: The Concepts and Technology behind Search » de Ricardo Baeza-Yates et Berthier Ribeiro-Neto (2010) (édition de 1999 en ligne)
- Livre « Learning to Rank for Information Retrieval » de Tie-Yan Liu (2011)
- Cours de l'école d'automne EARIA 2016, avec notamment un cours d'introduction à la RI, un cours sur les modèles…
  - http://www.asso-aria.org/index.php?option=com_content&view=article&id=135&Itemid=532