Révision
Indexation et Modèles de Recherche d'Information

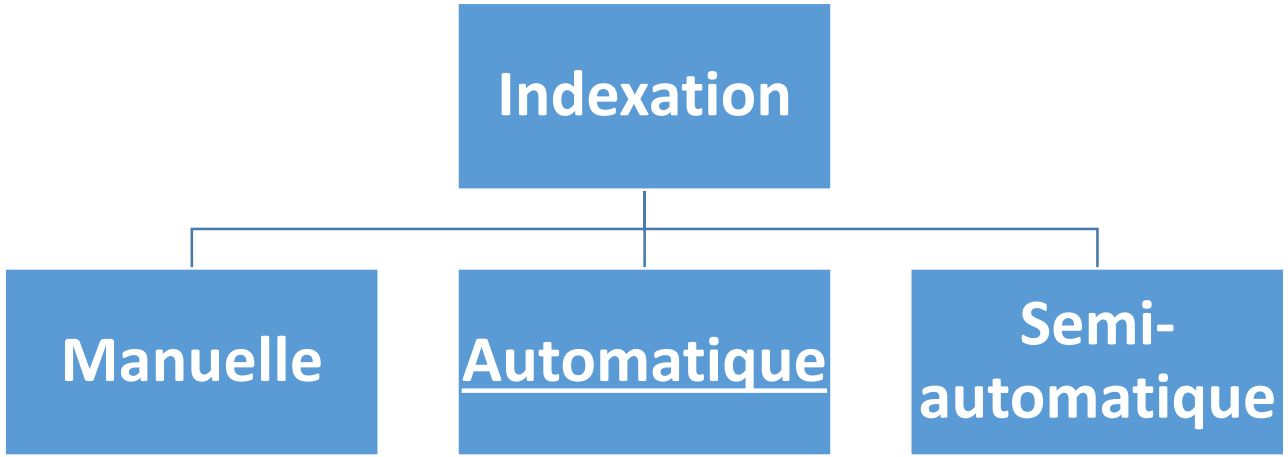
L3 ISIL

Par: Dr. OUKID Lamia

2019-2020

Indexation: Rappel

- C’est quoi ?
- Processus de représentation des documents et de la requête par un ensemble de descripteurs (mots-clefs, termes)



Exemple d'un processus d'indexation automatique: Rappel

Extraction des termes « *Tokenization* »

- terme = suite de caractères séparés par (blanc ou signe de ponctuation, caractères spéciaux,…)
- Dépend du langage

Elimination des mots vides

Exemples:

Anglais : the, or, a, you, I, us, ...etc.

Français : le , la de , des, je, tu, ...etc.

Normalisation

Techniques:

Lemmatisation: : pris, prend, prisse : **prendre**

Racinisation: économie, économiquement : **économ**

Troncature à x caractères: économiquement : écomoni (x=7)

Pondération des termes

Techniques:

Méthode des fréquences

Formule TF

Formule TF-IDF

Méthode des fréquence d'occurrences: Rappel

- **Principe :** Un mot est important si sa fréquence d'apparition dépasse **un seuil défini**
  1. Calculer la fréquence d'apparition de chaque terme dans le document
  2. Définir un seuil minimal
  3. Garder uniquement les termes dont la fréquence est supérieure ou égale au seuil

- **Exemple: seuil=2**
  - D: « langage Java basé langage C++ Java langage puissant»
  - **Index D:** « langage (3); Java (2) »

Formule Tf-Idf: Rappel

- TF *« Term Frequency »*

$$Tf_{t,d} = \frac{n_{t,d}}{N_d}$$

  - Où $n_{t,d}$ est la fréquence d'apparition du terme *t* dans le document *d* et N_d est le nombre total des termes dans *d*

- IDF *« Inverse Document Frequency »:*

$$Idf_t = \log \frac{D}{\{d_j : t_i \in d_j\}}$$

  - Où *D* est le nombre total de documents dans la collection et $\{d_j : c_i \in d_j\}$ représente le nombre de documents où le terme *t* apparait.

$$Tf - Idf = Tf_{t,d} \times Idf_t$$

Exercice

- **Soit la collection de documents suivants :**
- D1 : « Les prochaines élections présidentielles aux Etats-Unis sont prévues pour novembre prochain ».
- D2 : « Les New-Yorkais ont élu massivement le démocrate Bill de Blasio maire de leur ville lors des élections municipales ».
- D3 : « La course à la présidentielle aux Etats-Unis pour succéder à Barack Obama à la maison blanche est lancée ».
- Construire l'index de cette collection après :
  - Elimination des mots vides.
  - Normalisation par troncature à 8 caractères.
  - Pondération des termes en utilisant la formule Tf.

- Soit la requête suivante : Q= présidentielle ^ ⅂ municipale ˅ élection

- En se basant sur l'index calculé précédemment:
- Donner les documents retournés par la requête Q en se basant sur le modèle booléen classique.
- Quel est le document le plus pertinent pour Q en utilisant le modèle flou.

Solution exercice:

- Elimination des mots vides :

- D1 : « prochaines élections présidentielles Etats-Unis prévues novembre prochain ».
- D2 : « New-Yorkais élu massivement démocrate Bill Blasio maire ville élections municipales ».
- D3 : « course présidentielle Etats-Unis succéder Barack Obama maison blanche lancée ».

- Normalisation par troncature à 8 caractères :

- D1 : « prochain élection présiden Etats-Un prévues novembre prochain ».
- D2 : « New-York élu massivem démocrat Bill Blasio maire ville élection municipa ».
- D3 : « course présiden Etats-Un succéder Barack Obama maison blanche lancée ».

- Pondération des termes en utilisant la formule TF :

- Index D1 : « prochain (0.28) ; élection (0.14) ; présiden (0.14) ; Etats-Un (0.14) ; prévues (0.14) ; novembre(0.14) ».
- Index D2 : « New-York(0.1) ; élu(0.1) ; massivem(0.1) ; démocrat(0.1) ; Bill(0.1) ; Blasio(0.1) ; maire(0.1) ; ville(0.1) ; élection(0.1) ; municipa(0.1) ».
- Index D3 : « course(0.11) ; présiden(0.11) ; Etats-Un(0.11) ; succéder(0.11) ; Barack(0.11) ; Obama(0.11) ; maison(0.11) ; blanche(0.11) ; lancée(0.11) ».

Solution exercice (suite):

- Modèle booléen

- Q= présidentielle ^ ¬ municipale ˅ élection.

- Après prétraitement Q devient : présiden ^ ¬ municipa ˅ élection.

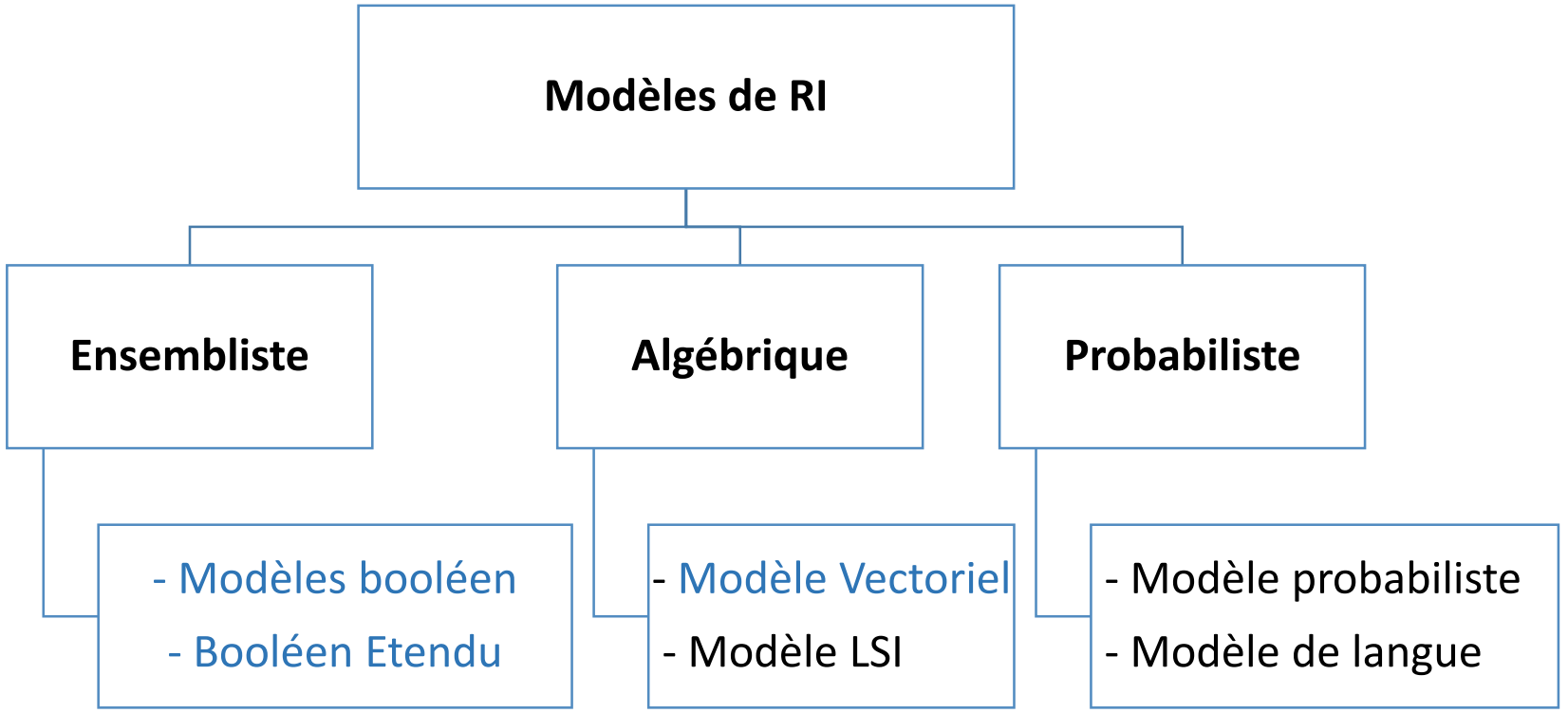
- Résultats retournés par Q en se basant sur le modèle booléen classique (1pts):
- Rsv(D1 , Q) = 1 ^ ¬0 ˅ 1= 1
- RSV(D2, Q) = 0 ^ ¬1 ˅ 1= 1
- RSV(D3, Q) =1 ^ ¬0 ˅ 0= 1

- Les documents retournés : D1, D2 et D3.

Solution exercice (suite):

- **Modèle flou:**
- Rsv(D1 , Q) :
- Rsv(D1, ٦ municipa) =1- RSV(D1, municipa)= 1-0=1
- Rsv(D1, présiden ^ ٦ municipa)= Min(Rsv(D1, présiden) ; Rsv(D1, ٦ municipa)= Min(0.14 ;1)= 0.14
- Rsv(D1 , Q)= Max (Rsv(D1, présiden ^ ٦ municipa) ; Rsv(D1, élection))=Max(0.14 ;0.14)=**0.14.**
-
- Rsv(D2 , Q) :
- Rsv(D2, ٦ municipa) =1- RSV(D1, municipa)= 1-0.1=0.9
- Rsv(D2, présiden ^ ٦ municipa)= Min(Rsv(D1, présiden) ; Rsv(D1, ٦ municipa)= Min(0 ;0.9)= 0
- Rsv(D2 , Q)= Max (Rsv(D1, présiden ^ ٦ municipa) ; Rsv(D1, élection))=Max(0 ;0.1)= **0.1.**
-
- Rsv(D3 , Q) :
- Rsv(D3, ٦ municipa) =1- RSV(D1, municipa)= 1-0 =1
- Rsv(D3, présiden ^ ٦ municipa)= Min(Rsv(D1, présiden) ; Rsv(D1, ٦ municipa)= Min(1 ;0.11)= 0.11
- Rsv(D3 , Q)= Max (Rsv(D1, présiden ^ ٦ municipa) ; Rsv(D1, élection))=Max(0.11 ;0)=**0.11**.
-
- Le document le plus pertinent est **D1.**

Modèles de Recherche d'Information (RI): Rappel



Modèle Booléen classique: Rappel

- **Un document di** est représenté par un **ensemble de termes ti**

- **Une requête q** est un **ensemble de termes tj** reliés par les opérateurs booléens : « **AND** », « **OR** » et « **NOT** »

- **Appariement exact:** présence ou l'absence des termes **tj de la requête q** dans les documents **di**
  - RSV (q,di) =1 ou 0

- **Exemple:**
- Q= langage AND (java OR NOT php)
- D1: « langage java base données »

- RSV(D1, Q)= 1 et (1 ou 1)=1

Modèle flou: Rappel

- **Un document di** est représenté par un **ensemble de termes ti** tel que**:**
  - **A chaque terme ti est associé un poids Wti € [0, 1]**

- **Une requête q** est un **ensemble de termes tj** reliés par les opérateurs booléens : « **AND** », « **OR** » et « **NOT** »

- **Appariement approché:**
  - interpréter la conjonction **« AND »** par le **« Min »**
  - Interpréter la disjonction **« OR »** par le **« Max »**

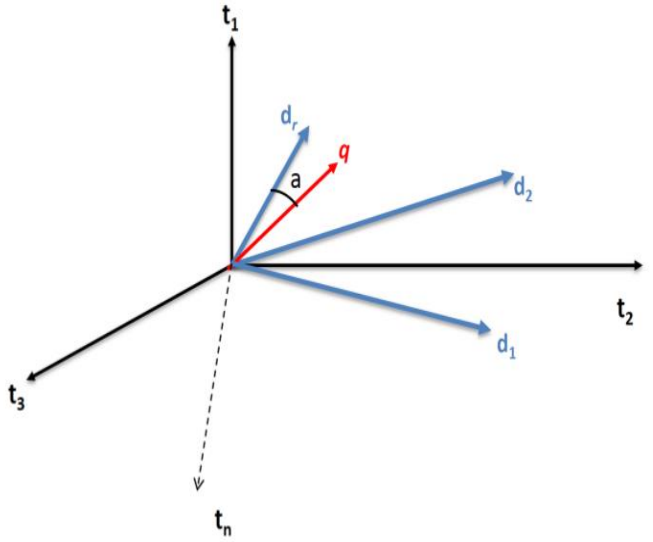
$$Rsv(d, t_1 \wedge t_2) = \min(Rsv(d, t_1), Rsv(d, t_2))$$
$$Rsv(d, t_1 \vee t_2) = \max(Rsv(d, t_1), Rsv(d, t_2))$$
$$Rsv(d, \neg t_1) = 1 - Rsv(d, t_1)$$

Modèle vectoriel: Rappel

- Basé sur un espace vectoriel **R** défini par l'ensemble des termes : **R < t1 , t2 , …, tn >**
- **Représentation d'un document:**

  d < wt1 , wt2 , …, wtn >
- **Représentation de la requête:**

  q < wqt1 , wqt2 , …, wqtn >
  - **wti** et **wqti** sont les poids du terme **ti** dans le document **d** et dans la requête **q**
  - **n** représente le nombre de termes dans l'espace.
- **Représentation sous forme de matrice terme X document**



- **Appariement approché:**
  - Calcul de similarité par différentes mesures: produit interne, mesure du cosinus, coef de Dice, mesure de Jaccard

Exercice 3, série 2:

- Soit le corpus de documents suivant:
- Document 1 : « Le professeur parle de la recherche d'information textuelle »
- Document 2 : « La recherche d'information est un domaine de recherche qui s'intéresse à des nombreux problèmes »
- Document 3 : « Le modèle vectoriel de recherche d'information est un modèle algébrique »

- Construire l'index de ces trois documents après **élimination des mots vides**, en utilisant la méthode de **pondération TF**.
- Construire la **matrice terme×document** de ce corpus.
- Quels sont les résultats retournés pour les requêtes suivantes :
  - Q1 : recherche documentaire
  - Q2 : recherche d'information
  - Q3 : recherche d'information textuelle
  - Q4 : domaine du modèle vectoriel
- Remarque : utiliser les mesures : produit interne, Cosinus, coef de Dice et Jaccard.

Série d’exercices 2; Exercice 3:

- **Elimination des mots vides :**
- Document 1 : « Le professeur parle de la recherche d'information textuelle »
- Document 2 : « La recherche d'information est un domaine de recherche qui s'intéresse à des nombreux problèmes »
- Document 3 : « Le modèle vectoriel de recherche d'information est un modèle algébrique »

- **Pondération par la formule Tf :**
- Index Document 1 : « professeur (0.2) ; parle (0.2) ; recherche (0.2) ; information (0.2) ; textuelle (0.2) »
- Index Document 2 : « recherche (0.28) ; information (0.14) ; domaine (0.14) ; intéresse(0.14) ; nombreux(0.14) ; problèmes (0.14) »
- Index Document 3 : « modèle(0.33) ; vectoriel(0.17) ; recherche(0.17) ; information(0.17) ; algébrique(0.17) »

Série d'exercices 2; Exercice 3:

- **Matrice terme X Document :**

	d1	d2	d3
professeur	0.2	0	0
Parle	0.2	0	0
recherche	0.2	**0.28**	**0.17**
information	0.2	0.14	0.17
textuelle	0.2	0	0
domaine	0	0.14	0
intéresse	0	0.14	0
nombreux	0	0.14	0
problèmes	0	0.14	0
modèle	0	0	0.33
vectoriel	0	0	0.17
algébrique	0	0	0.17

Index Document 1 : « professeur (0.2) ; parle (0.2) ; recherche (0.2) ; information (0.2) ; textuelle (0.2) »
Index Document 2 : « recherche (0.28) ; information (0.14) ; domaine (0.14) ; intéresse(0.14) ; nombreux(0.14) ; problèmes (0.14) »
Index Document 3 : « modèle(0.33) ; vectoriel(0.17) ; recherche(0.17) ; information(0.17) ; algébrique(0.17) »

Mesures de similarité

Inner product	$\|X \cap Y\|$	$\sum x_i * y_i$
Coef. de Dice	$\frac{2*\|X \cap Y\|}{\|X\|+\|Y\|}$	$\frac{2*\sum x_i * y_i}{\sum x_i^2 + \sum y_j^2}$
Mesure du cosinus	$\frac{\|X \cap Y\|}{\sqrt{\|X\|} * \sqrt{\|Y\|}}$	$\frac{\sum x_i * y_i}{\sqrt{\sum x_i^2 * \sum y_j^2}}$
Mesure du Jaccard	$\frac{\|X \cap Y\|}{\|X\|+\|Y\|-\|X \cap Y\|}$	$\frac{\sum x_i * y_i}{\sum x_i^2 + \sum y_j^2 - \sum x_i * y_i}$

Série d'exercices 2; Exercice 3:

- **Calcul de similarité RSV pour la requête Q1:**

- **Produit Interne** $Rsv(d_j, q_k) = \sum_{i=1}^{n} w_{ij} \times w_{q_{ik}}$

- w_{ij} est le poids du terme i dans le document d_j
- $w_{q_{ik}}$ est le poids du terme i dans la requête q_k

- **Q1 : recherche documentaire**
- **Index Q1: recherche (0.5); documentaire (0.5)**
- RSV (d1, Q1)= 0.2X0.5= **0.1**
- RSV (d2, Q1)= 0.28x0.5= **0.14**
- RSV (d3, Q1)= 0.17x0.5= **0.085**

- Résultats retournés pour Q1 :

Classement	Document	RSV(di, Q1)
1	D2	0.14
2	D1	0.1
3	D3	0.085

Série d'exercices 2; Exercice 3:

- **Calcul de similarité RSV pour la requête Q2:**
- **Mesure du Cosinus:**
- **Q2 : recherche d'information**
- **Index Q2: recherche (0.5); information (0.5)**

- RSV (d1, Q2)= $\frac{0.2x0.5+0.2x0.5}{\sqrt{(0.2^2+0.2^2+0.2^2+0.2^2+0.2^2)x(0.5^2+0.5^2)}} = 0.625$

- RSV (d2, Q2)= $\frac{0.28x0.5+0.14x0.5}{\sqrt{(0.28^2+0.14^2+0.14^2+0.14^2+0.14^2+0.14^2)x(0.5^2+0.5^2)}}$ = 0.71

- RSV (d3, Q2)= $\frac{0.17x0.5+0.17x0.5}{\sqrt{(0.17^2+0.17^2+0.33^2+0.17^2+0.17^2)x(0.5^2+0.5^2)}}$ = 0.51

- Résultats retournés pour Q2 :

Classement	Document	RSV(di, Q2)
1	D2	0.71
2	D1	0.625
3	D3	0.51

Série d'exercices 2; Exercice 3:

- **Calcul de similarité RSV pour la requête Q3:**
- **Coef de Dice:**
- **Q3 : recherche d'information textuelle**
- **Index Q3: recherche(0.33); information(0.33); textuelle(0.33)**

- RSV (d1, Q3)= $\frac{2x(0.2x0.33+0.2x0.33+0.2x0.33)}{(0.2^2+0.2^2+0.2^2+0.2^2+0.2^2)+(0.33^2+0.33^2+0.33^2)}$=0.75

- RSV (d2, Q3)= $\frac{2x(0.28x0.33+0.14x0.33)}{(0.28^2+0.14^2+0.14^2+0.14^2+0.14^2+0.14^2)+(0.33^2+0.33^2+0.33^2)}$=0.55
-
- RSV (d3, Q3)= $\frac{2x(0.17x0.33+0.17x0.33)}{(0.17^2+0.17^2+0.33^2+0.17^2+0.17^2)+(0.33^2+0.33^2+0.33^2)}$= 0.41

- Résultats retournés pour Q3:

Classement	Document	RSV(di, Q3)
1	D1	0.75
2	D2	0.55
3	D3	0.41

Série d'exercices 2; Exercice 3:

- **Calcul de similarité RSV pour la requête Q4:**
- **Mesure de Jaccard:**
- **Q4 : domaine du modèle vectoriel**
- **Index Q4: domaine(0.33); modèle(0.33); vectoriel(0.33)**

- RSV (d1, Q4)= 0
- RSV(d2,Q4)=$\frac{0.14x0.33}{(0.28^2+0.14^2+0.14^2+0.14^2+0.14^2+0.14^2)+(0.33^2+0.33^2+0.33^2)-(0.14x0.33)}$=0.1

- RSV (d3, Q4)=$\frac{0.33x0.33+0.17x0.33}{(0.17^2+0.17^2+0.33^2+0.17^2+0.17^2)+(0.33^2+0.33^2+0.33^2)-(0.33x0.33+0.17x0.33)}$
- RSV (d3, Q4)=0.18
- Résultats retournés pour Q4:

Classement	Document	RSV(di, Q3)
1	D3	0.18
2	D2	0.1