Université Mohamed Boudiaf de M'sila

Faculté des mathématiques et de l'informatique

Département d'informatique

Module : RechInf

SÉRIE N°02

Exercice 01

Soit la collection des documents décrite ci-après :

D1 : {L'internet est un univers de services, ou chacun peut produire, utiliser, et donner de l'information ou en trouver, et participer librement à cette grande planète de l'information.}

D2 : {Internet est le plus grand réseau au monde avec des centaines de millions d'ordinateurs et de réseaux connectés. Les réseaux varient en taille.}

D3 : {Internet est le réseau informatique mondial qui rend accessibles au public des services variés comme le courrier électronique, la messagerie instantanée et le World Wide Web (le web), en utilisant le protocole de communication IP (Internet Protocol).}

D4 : {Internet (INTERconnected NETworks) : interconnexion des réseaux, réseau des réseaux, le réseau international (INTERnational NETwork).}

1. ***Appliquer une opération de prétraitement (preprocessing) sur D1, D2, D3, D4***
2. ***Définir le vocabulaire de cette collection.***
3. ***Représenter les différents documents de la collection en utilisant :***
   a) Une représentation binaire.
   b) Une représentation à 03 valeurs.
   c) Une représentation basée sur la fréquence simple TF.
   d) Une représentation TF-IDF

Exercice 02

Un espace de documents (vocabulaire) est défini par les termes suivants :

{hardware, software, information, users, stemming, retrieval, relevance, computer, system, processor}

Soient les documents suivants:

D1: {hardware, computer, processor}

D2 : {software, information, users, system}

D3: {stemming, retrieval, relevance, system}

D4: {hardware, software, information, users}

D5: {hardware, users, computer}

D6: {Information, users, retrieval, relevance}

D7: {stemming, relevance, system}

1. ***Représenter chaque document en utilisant le modèle binaire.***

Soient les requêtes suivantes :

R1: « hardware AND software ».

R2: « hardware OR software ».

R3: « Information AND retrieval».

R4: «(hardware AND software)OR(information AND retrieval)OR(NOT(computer AND processor)) »

2. ***Donner les documents retournés par le système de recherché pour chaque requête et selon chacun des modèles de recherche suivants :***
    a) Modèle booléen.
    b) Modèle vectoriel basé sur le calcul de similarités (cosine, Euclidien)

Exercice 03

---

1. Donner une liste contenant 10 stopwords pour le Français
2. Donner une liste contenant 10 stopwords pour l'Anglais
3. Donner une liste contenant 10 stopwords pour l'Arabe
4. Donner une définition claire et exacte pour le stemming
5. Donner des exemples de « stem » pour les langues : Français, Anglais, Arabe
6. Citer les méthodes les plus utilisées pour le stemming

Bon Courage

Resp.Module : Said KADRI

Université Mohamed Boudiaf de M'sila

Faculté des mathématiques et de l'informatique

Département d'informatique

Module : RechInf

SÉRIE N°02

Exercice 01

D1, D2, D3, D4 sont des documents d'une collection C

1. Application d'une opération de prétraitement sur D1, D2, D3, D4

La phase de prétraitement comporte les taches suivantes :

- Elimination de la ponctuation
- Elimination des stopwords
- Elimination des chiffres et des abréviations, ainsi que les noms propres
- Normalisation des caractères (Ecrire tous les caractères en minuscule ou en majuscule)
- Appliquer la technique de stemming (chercher les stems des mots)
- Appliquer une opération de lémmatisation (lemmatization) (convertir le pluriel en singulier et les verbes à l'infinitif)
- Segmenter les textes en mots et représenter chaque texte par un vecteur (représentation sac de mots)

On note le suivant :

S : pour désigner le stemming ; L : pour désigner la lémmatisation

Après avoir accompli ces taches, on obtient :

D1 : {internet être$^{(L)}$ univers service$^{(L)}$ pouvoir$^{(L)}$ produire utiliser donner information trouver participer libre$^{(S)}$ grand planète information.}

D2 : {Internet être$^{(L)}$ grand$^{(S)}$ réseau$^{(L)}$ monde centaine$^{(L)}$ million$^{(L)}$ ordinateur$^{(L)}$ réseau connecter$^{(L)}$ réseau$^{(L)}$ Varier$^{(L)}$ taille.}

D3 : {Internet être$^{(L)}$ réseau$^{(L)}$ information$^{(S)}$ monde$^{(S)}$ rendre$^{(L)}$ accéder$^{(S)}$ public service$^{(L)}$ varier$^{(L)}$ courrier électronique message$^{(S)}$ instant$^{(S)}$ world wide web web utiliser protocol$^{(S)}$ communication Internet Protocol$^{(S)}$}

D4 : {Internet connecter$^{(S+L)}$ network connecter$^{(S+L)}$ réseau$^{(L)}$ réseau réseau réseau international international network}

**** Représentation des différents documents par des vecteurs de mots (Rep. Sac de Mots):***

D1= (internet, être, univers, service, pouvoir, produire, utiliser, donner, information, trouver, Participer, libre, grand, planète, information)

D2 = (Internet, être, grand, réseau, monde, centaine, million, ordinateur, réseau, connecter réseau, varier, taille.}

D3 = (Internet, être, réseau, informatique, monde, rendre, accéder, public, service, varier, courrier, électronique, message, instant, world, wide, web, web, utiliser, protocol, communication, Internet, Protocol)

D4 = (Internet, connecter, network, connecter, réseau, réseau, réseau, réseau, international, International, Network)

2. Le vocabulaire associé à la collection C

On collecte les mots à partir des différents documents sans réptition

V = (internet[1], être[2], univers[3], service[4], pouvoir[5], produire[6], utiliser[7], donner[8], information[9], trouver[10], Participer[11], libre[12], grand[13], planète[14], réseau[15], monde[16], centaine[17], Million[18], ordinateur[19], connecter[20], varier[21], taille[22], rendre[23], accéder[24], public[25], courrier[26], Electronique[27], message[28], instant[29], world[30], wide[31], web[32], protocol[33], communication[34], network[35], international[36])

|V| = 36

3. Représentation des différents documents de la collection C

On note par : V1, V2, V3, V4 les vecteurs associés aux documents D1, D2, D3, D4

a) Une représentation binaire.

Indice	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**	**32**	**33**	**34**	**35**	**36**
V1 =	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
V2 =	1	1	0	0	0	0	0	0	0	0	0	1	0	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
V3 =	1	1	0	1	0	0	1	0	1	0	0	0	0	0	1	1	0	0	0	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	0	0
V4 =	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	

b) ***Une représentation à 03 valeurs.*** (0 : terme inéxistant ; 1 : existe une seule fois ; 2 : existe plus qu'une fois)

Indice	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**	**32**	**33**	**34**	**35**	**36**
V1 =	1	1	1	1	1	1	1	1	2	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
V2 =	1	1	0	0	0	0	0	0	0	0	0	1	0	2	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
V3 =	2	1	0	1	0	0	1	0	1	0	0	0	0	0	1	1	0	0	0	0	1	0	1	1	1	1	1	1	1	1	2	2	1	0	0	
V4 =	1	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	2	2	

c) Une représentation basée sur la fréquence simple TF.

Indice 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36

V1 = (1, 1, 1, 1, 1, 1, 1, 1, 2, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)

V2 = (1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 3, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)

V3 = (2, 1, 0, 1, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 1, 0, 0)

V4 = (1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 4, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 2)

d) Une représentation TF-IDF

On calcule TF-IDF pour chaque terme par la formule :

TF-IDF(i, j) = TF(i, j)*Log(N/df(i))

Avec: i: un terme

J: le document contenant le terme

N : la taille de la collection

Df(i) : nombre de documents contenant le terme i

Prenons l'exemple du document D4

La réprésentation TF est :

V4 = (1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 4, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 2)

==> la réprésentation TF-IDF sera comme suit :

V4 = (1*Log(4/4), 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 4*Log(4/3), 0, 0, 0, 0, 2*Log(4/2), 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2*Log(4/1), 2*Log(4/1))

==> V4 = (0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,49975, 0, 0, 0, 0, 0,602, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1,204, 1,204)

Exercice 02

On a le vocabulaire suivant :

(hardware, software, information, users, stemming, retrieval, relevance, computer, system, processor)

|V| = 10

D1: {hardware, computer, processor}

D2 : {software, information, users, system}

D3: {stemming, retrieval, relevance, system}

D4: {hardware, software, information, users}

D5: {hardware, users, computer}

D6: {Information, users, retrieval, relevance}

D7: {stemming, relevance, system}

1. Représentation des documents :

V1 = (1, 0, 0, 0, 0, 0, 0, 0, 1, 1)

V2 = (0, 1, 1, 1, 0, 0, 0, 0, 1, 0)

V3 = (0, 0, 0, 0, 1, 1, 1, 0, 1, 0)

V4 = (1, 1, 1, 1, 0, 0, 0, 0, 0, 0)

V5 = (1, 0, 0, 1, 0, 0, 0, 1, 0, 0)

V6 = (0, 0, 1, 1, 0, 1, 1, 0, 0, 0)

V7 = (0, 0, 0, 0, 1, 0, 1, 0, 1, 0)

Soient les requêtes suivantes :

R1: « hardware AND software ».

R2: « hardware OR software ».

R3: « Information AND retrieval».

R4: «(hardware AND software)OR(information AND retrieval)OR(NOT(computer AND processor)) »

2. Les documents retournés par le système de recherche pour chacune des requêtes précédentes selon les deux cas suivants :

a) Modèle booléen.

A chaque fois, on cherche les documents qui remplissent la condition booléenne de la requête :

Pour R1 : on trouve L'ensemble {D4}

==> le système retourne le document {D4}

Pour R2 : les documents retournés par le système est : {D1, D2, D4, D5}

Pour R3 : {D6}

Pour R4 :

R4 = R41∪R42∪R43

R41 = R1 = {D4} ; R42 = R3 = {D6}

R43 = NOT(ComputerANDProcessor): désigne le nombre de documents qui ne contiennent pas les deux termes Computer et Processor à la fois ==> {D2, D3, D4, D5, D6, D7}

Ou bien :

C – {nb.docs contenant les deux termes à la fois}= C – {D1} = {D2, D3, D4, D5, D6, D7}

Ou bien :

$\overline{Computer\ AND\ Processor} = \overline{Computer}\ OR\ \overline{Processor} =$

{nb.docs qui ne contiennent pas le terme Computer}∪{nb.docs qui ne contiennent pas le terme Processorr}∪{nb.docs qui ne contiennent pas les deux terme Computer et Processor à la fois}

{D2, D3, D4, D5, D6, D7}∪{D2, D3, D4, D5, D6, D7}∪{D2, D3, D4, D5, D6, D7}

= {D2, D3, D4, D5, D6, D7}= C – {D1}

Alors : R4 = R1∪R3∪R43 = {D4}∪{D6}∪{D2, D3, D4, D5, D6, D7}=

{D2, D3, D4, D5, D6, D7}

b) Modèle vectoriel basé sur le calcul de similarités (cosine)

- R1: « hardware software ».==> VR1 = (1, 1, 0, 0, 0, 0, 0, 0, 0, 0)
- R2: « hardware software ».==> VR2 = (1, 1, 0, 0, 0, 0, 0, 0, 0, 0)
- R3: « Information retrieval». ==> VR3 = (0, 0, 1, 0, 0, 1, 0, 0, 0, 0)
- R4: «(hardware software information retrieval computer processor) » ==> VR4 = (1, 1, 1, 0, 0, 1, 0, 1, 0, 1)

Pour trouver les documents retournés par le système dans le cas de l'application de R1, on calcule les similarités comme suit :

$$\text{Sim(VR1, V1)} = \frac{\overrightarrow{VR1}.\overrightarrow{V1}}{||VR1||.||V1||} = \frac{(1,1,0,0,0,0,0,0,0,0).(1,0,0,0,0,0,0,0,1,1)}{\sqrt{2}.\sqrt{3}} = \frac{1}{\sqrt{6}} = 0{,}408$$

$$\text{Sim(VR1, V2)} = \frac{\overrightarrow{VR1}.\overrightarrow{V2}}{||VR1||.||V2||} = \frac{(1,1,0,0,0,0,0,0,0,0).(0,1,1,1,0,0,0,0,1,0)}{\sqrt{2}.\sqrt{4}} = \frac{1}{2\sqrt{2}} = 0{,}353$$

$$\text{Sim(VR1, V3)} = \frac{\overrightarrow{VR1}.\overrightarrow{V3}}{||VR1||.||V3||} = \frac{(1,1,0,0,0,0,0,0,0,0).(0,0,0,0,1,1,1,0,1,0)}{\sqrt{2}.\sqrt{4}} = \frac{0}{2\sqrt{2}} = 0{,}0$$

$$\text{Sim(VR1, V4)} = \frac{\overrightarrow{VR1}.\overrightarrow{V4}}{||VR1||.||V4||} = \frac{(1,1,0,0,0,0,0,0,0,0).(1,1,1,1,0,0,0,0,0,0)}{\sqrt{2}.\sqrt{4}} = \frac{2}{2\sqrt{2}} = 0{,}707$$

$$\text{Sim(VR1, V5)} = \frac{\overrightarrow{VR1}.\overrightarrow{V5}}{||VR1||.||V5||} = \frac{(1,1,0,0,0,0,0,0,0,0).(1,0,0,1,0,0,0,1,0,0)}{\sqrt{2}.\sqrt{3}} = \frac{1}{\sqrt{6}} = 0{,}408$$

$$\text{Sim(VR1, V6)} = \frac{\overrightarrow{VR1}.\overrightarrow{V6}}{||VR1||.||V6||} = \frac{(1,1,0,0,0,0,0,0,0,0).(0,0,1,1,0,1,1,0,0,0)}{\sqrt{2}.\sqrt{4}} = \frac{0}{2\sqrt{2}} = 0{,}0$$

$$\text{Sim(VR1, V7)} = \frac{\overrightarrow{VR1}.\overrightarrow{V7}}{||VR1||.||V7||} = \frac{(1,1,0,0,0,0,0,0,0,0).(0,0,0,0,1,0,1,0,1,0)}{\sqrt{2}.\sqrt{3}} = \frac{0}{\sqrt{6}} = 0{,}0$$

On ordonne la liste des documents selon l'ordre décroissant de la similarité

{D4, D1, D5, D2}

- On refait la même chose pour les requêtes : R2, R3, R4 ==> on calcule :

b. Pour R2 :

Sim(VR2, V1), Sim(VR2, V2), Sim(VR2, V3), Sim(VR2, V4), Sim(VR2, V5), Sim(VR2, V6), Sim(VR2, V7)

c. Pour R3 :

Sim(VR3, V1), Sim(VR3, V2), Sim(VR3, V3), Sim(VR3, V4), Sim(VR3, V5), Sim(VR3, V6), Sim(VR3, V7)

d. Pour R4 :

Sim(VR4, V1), Sim(VR4, V2), Sim(VR4, V3), Sim(VR4, V4), Sim(VR4, V5), Sim(VR4, V6), Sim(VR4, V7)

- On refait la même chose pour la similarité euclidienne en utilisant la formule ci-après :

$$Sim(T1, T2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Avec : $T1(x_1, y_1)$, $T2(x_2, y_2)$ deux textes (documents)

Exercice 03

---

1. Donner une liste contenante 10 stopwords pour le Français

   Fr : {le, la, les, de, du, des, pour, au, ou, sur, devant, en face, en, par, avec, …}
2. Donner une liste contenante 10 stopwords pour l'Anglais

   En : {The, of, for, on, from, up, down, in, with, until, since, …}
3. Donner une liste contenante 10 stopwords pour l'Arabe

   {من، إلى، على، فوق، تحت، أمام، وراء، من أجل، بينما، مادام، لأن، كأن، ...}
4. Donner une définition claire et exacte pour le stemming

   Stemming : Une technique linguistique qui permet de réduire le mot à sa forme de base (la racine/le radical) ou à une forme plus réduite (Stem)
5. Donner des exemples de « stem » pour les langues : Français, Anglais, Arabe

   Exemples de stems :

   Fr : pour les mots : Informationnelles, Informationel, Information, Informatique, informeront, informel ➔ Inform/Informer

   En : Usefullness, Usefull, Users, Using, Used ➔ Use

   Ar: يعلمونهن، يعلمونهم، يعلمنهن، معلمات، معلمون، معلمة، معلم، عالم، علوم/ عوالم/ معالم، معلم

   ➔ معلم، يعلم، علم
6. Citer les méthodes les plus utilisées pour le stemming
   - ***Les méthodes morphologiques : sont basées sur les règles de grammaire de chaque langue***
   - ***Les méthodes statistiques : sont basées sur des calculs des probabilités des stems les plus proches à partir d'une liste de stems candidats.***
   - ***Les méthodes mixtes : on combine les deux approches précédentes (morphologiques et statistiques)***