UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Escola Politècnica Superior d'Enginyeria
de Vilanova i la Geltrú

RECUPERACIÓ DE LA INFORMACIÓ

EXERCICIS DEL TEMA 2: Models de recuperació de la informació

Abans de començar a resoldre els exercicis, comprova si saps respondre aquestes preguntes:

1. El model booleà no ordena els documents de resposta mentre que el model vectorial sí que els ordena. Cert o fals?
2. En el model booleà, podem veure els documents com a vectors? Les operacions AND, OR i BUTNOT, es poden descriure com a operacions sobre vectors?
3. Suposa que et donen la freqüència de cada terme en un document. Quina altra informació et cal per calcular la seva representació amb pesos *tf-idf*?
4. Amaga els apunts de l'assignatura. Escriu la fórmula de la mesura cosinus per calcular la similitud entre dos documents. Assegura't de que entens bé cadascun dels símbols que conté. Ara comprova si l'has escrita correctament.
5. Fes el mateix amb la fórmula de càlcul dels pesos *tf-idf*.

Exercici 1

Considereu els documents següents:

D_1: Shipment of gold damaged in a fire

D_2: Delivery of silver arrived in a silver truck

D_3: Shipment of gold arrived in a truck

i el conjunt de termes següent:

$$T = \{\text{fire, gold, silver, truck}\}$$

Calculeu, usant el model booleà, quins documents satisfan la consulta

```
(fire OR gold) AND (truck OR NOT silver)
```

i justifiqueu la vostra resposta. Feu el mateix amb la consulta

```
(fire OR NOT silver) AND (NOT truck OR NOT fire)
```

Discutiu si és possible reescriure aquestes consultes utilitzant només els operadors `AND`, `OR` i `BUTNOT` de forma lògicament equivalent. Això vol dir que han de ser equivalents per qualsevol col·lecció de documents, no només per aquesta.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Escola Politècnica Superior d'Enginyeria
de Vilanova i la Geltrú

Exercici 2

Considereu la següent col·lecció formada per 5 documents:

`Doc1`: we wish efficiency in the implementation for a particular application

`Doc2`: the classification methods are an application of Li's ideas

`Doc3`: the classification has not followed any implementation pattern

`Doc4`: we have to take care of the implementation time and implementation efficiency

`Doc5`: the efficiency is in terms of implementation methods and application methods

Suposeu que tota paraula de 6 o més lletres és un terme i que els termes s'ordenen per ordre d'aparició.

1. Doneu la representació de cada document en el model booleà.
2. Doneu la representació dels documents `Doc1` i `Doc5` en el model vectorial usant pesos *tf-idf*. Calculeu el coeficient de similitud entre aquests dos documents usant com a mesura el cosinus.

doc1 = [1/1 . log(5/3), 1/1 .log(5/4), 1/1*log(5/1),1/1*log(5/3),0,0,0
doc5 = [1/2.log(5/3),1/2*log(5/4),0,1/2*log(5/3),0,2/2*log(5/2),0,0]

(Resposta de l'apartat 2: 0.162*)*

sum(dos1, doc5) = (doc1* doc5)/(|doc1|*|doc5|) = 0.054/0.332 = 0.162 = 16% de similitit

Exercici 3

Hem indexat una col·lecció de documents que contenen els termes de la taula següent; la segona columna indica el percentatge de documents en els que cada terme apareix.

Terme	% docs
computer	10%
software	10%
bugs	5%
code	2%
developer	2%
programmers	2%

idf(t) = $\log_{10}$(N/dft) dft = N * (10/100)
log10(N/((10/100)*N)

Donada la consulta Q="computer software programmers", calculeu la similitud entre Q i els documents següents, usant pesos *tf-idf* i la mesura cosinus. Determineu la seva ordenació relativa:

- D1 = "programmers build computer software"
- D2 = "most software has bugs, but good software has less bugs than bad software"
- D3 = "some bugs can be found only by executing the software, not by examining the source code"

Q = [1, 1, 0, 0, 0, 1]
D1 [1/1*log10(100/10), 1/1*log10(100/10), 0 , 0, 0, 1/1*log10(100/20)]
D2 [0, 3/3*log10(100/10), 2/3*log10(100/2), 0, 0, 0]
D3 [0, 1/1*log10(100/10), 1/1*log10(100/5), 1/1*log10(100/2), 0, 0]

||Q|| = 1.732 ||D1|| = 2.21 ||D2|| = 1.324 ||D3|| = 2.36

sim(D1, Q) = D1 * Q / ||D1|| * ||Q|| = 0.963
sim(D2, Q) = ... = 0.436
sim(D2, Q) = ... = 0.224

Exercici 4

Suposem que els termes A, B, C i D apareixen, respectivament, en 10 000, 8 000, 5 000 i 3 000 documents, respectivament, en una col·lecció de 100 000.

1. Considereu la consulta booleana `(A and B) or (C and D)`. De quants documents pot constar la resposta en el cas pitjor?
2. I per la consulta `(A and B) or (A and D)`? Penseu-ho bé.
3. Calculeu la similitud entre els documents "A B B A C C" i "D A D B B C C" usant pesos *tf-idf* i la mesura cosinus.

(Respostes: 1) 11.000 2) 10.000 3) 0.736)

Exercici 5

Hem indexat una col·lecció d'un milió de documents que inclouen els termes següents:

Terme	# docs
computing	300 000
networks	200 000
computer	100 000
files	100 000
system	100 000
client	80 000
programs	80 000
transfer	50 000
agents	40 000
p2p	20 000
applications	10 000

1. Calculeu la similitud entre els documents D1 i D2 següents, usant pesos *tf-idf* i la mesura cosinus:

   D1 = "p2p programs help users sharing files, applications, other programs, etc. in computer networks"

   D2 = "p2p networks contain programs, applications, and also files"

2. Suposeu que estem usant la mesura cosinus i pesos *tf-idf* per calcular la similitud entre documents. Doneu un document format d'exactament *dos* termes diferents que aconsegueixi una similitud màxima amb el document següent:

   "p2p networks contain programs, applications, and also files"

Calculeu aquesta similitud i justifiqueu que és màxima d'entre tots els documents que constin de dos termes.

(Resposta de l'apartat 1: 0.925*)*

Exercici 6

Considereu la col·lecció següent formada per quatre documents:

`Doc1`: Shared Computer Resources

`Doc2`: Computer Services

`Doc3`: Digital Shared Components

`Doc4`: Computer Resources Shared Components

Suposant que cada paraula és un terme

1. Escriviu la representació del document `Doc3` usant el model booleà.
2. En el model booleà, quins documents es recuperarien a partir de la consulta "Computer BUTNOT Components"?
3. Calculeu el valor de `idf` pels termes "Computer" i "Components".
4. Calculeu la representació de la col·lecció de documents usant el model vectorial amb pesos *tf-idf*.
5. Calculeu la similitud entre la consulta "Computer Components" (amb pesos binaris) i el document Doc4 (amb pesos *tf-idf*), usant la mesura cosinus.

(Resposta de l'apartat 5: 0.6535*)*

Exercici 7

Donats un document d i una consulta q en el model vectorial vist a classe, suposeu que la similitud entre la consulta i el document és $0,08$. Si intercanviem els continguts entre el document i la consulta, és a dir, tots els termes de q van a d i tots els termes de d van a q, quina és ara la similitud entre q i d? Raoneu la resposta.

1. menor de $0,08$
2. igual: $0,08$
3. major que $0,08$
4. depèn de l'esquema de pesos

Exercici 8

Per què l'*idf* d'un terme és sempre finit?

Exercici 9

Quin és l'*idf* d'un terme que apareix en tots els documents? Compareu-ho amb l'eliminació de paraules funcionals.

Exercici 10

El pes *tf-idf* d'un terme en un document, pot ser més gran que 1?

Exercici 11

En el model vectorial amb l'esquema de pesos *tf-idf* i usant com a mesura de similitud el cosinus, suposeu que afegim alguns documents a la col·lecció. Els pesos dels termes dels documents ja indexats, es veuen afectats per aquest canvi? Raoneu la resposta.

1. no
2. sí, afecta les tf dels termes que apareguin en els altres documents
3. sí, afecta les *idf* dels termes que apareguin en els altres documents
4. sí, afecta les tf i les *idf* dels termes que apareguin en els altres documents