Esercizio 1 Relevance Feedback –

Considerate la query: "*cheap CDs cheap DVDs extremely cheap CDs*"
L'utente esamina i risultati e giudica rilevante
d1: *CDs cheap software cheap CDs*

e non rilevante
d2: *cheap thrills DVDs*

Assumendo di usare come pesi solo i valori di frequenza, e di assumere alfa = 1, beta = 0.75, gamma = 0.25 espandete la query usando la formula di Rocchio

Esercizio 2 LSI
Considerate la matrice L (termini documenti)

0.0 2.0 1.0
0.0 3.0 0.0
2.0 1.0 0.0

E la sua decomposizione SVD

Matrice U

```
-0.545 -0.238 -0.804
-0.762 -0.259  0.593
-0.349  0.936 -0.040
```

Matrice V^T

```
-0.182 -0.973 -0.142
 0.978 -0.165 -0.124
-0.097  0.162 -0.982
```

Matrice Sigma

```
3.830 0.000 0.000
0.000 1.914 0.000
0.000 0.000 0.819
```

Calcolate la matrice L' che si ottiene con una riduzione di rango 2 della matrice sigma, e confrontatela con la matrice L originaria.

Esercizio 3

q	1	1	1	1	0	0	0	0	0
D1	0	1	1	0	0	0	1	1	1
D2	1	1	0	0	0	0	0	1	1
D3	1	0	1	1	1	1	1	1	1
D4	1	0	1	0	0	0	1	0	0
D5	1	0	1	0	0	0	1	0	0
D6	1	0	0	1	0	0	0	1	0
D7	1	0	0	0	0	1	0	0	0
D8	0	0	0	1	0	1	0	1	0
D9	0	1	0	1	1	1	0	1	1

Oridinate i documenti rispetto alla query utilizzando il probabilistic retrieval con il Binary Independence Model, e metodo incrementale per il calcolo del Retrieval Status Value. Ponete |V|=4.

Riassunto delle formule:

$$RSD(Dj) = \sum_{x_{ji}=1, q_i=1} c_i$$

$$c_i = \log \frac{p_i(1-r_i)}{r_i(1-p_i)}$$

Usate le stime di pi e ri con “adjustement” per basse frequenze, semplificato come segue:

$$p_i = \frac{V_i + 0{,}5}{V+1}, \quad r_i = \frac{n_i - V_i + 0{,}5}{N - V + 1}$$

Dove Vi è il numero dei documenti in V che contiene la parola i, ni è il numero dei documenti in N che contiene la parola i , V è usato al posto di |V| ed N è la dimensione della collezione completa (9 nel caso in esame).

Soluzione Esercizio 1

word	query q	d_1	d_2
CDs	2	2	0
cheap	3	2	1
DVDs	1	0	1
extremely	1	0	0
software	0	1	0
thrills	0	0	1

1.0 q+0.75 d1−0.25d2

si ottiene la query di Rocchio:

Q: (3.5 4.25 0.75 1 0.75 0)

Soluzione http://www.bluebit.gr/matrix-calculator/

$U_2\Sigma_2$

```
-2.087 -0.456
-2.918 -0.496
-1.337  1.792
```

Matrice L' $U_2\Sigma_2V_2^T$

```
-0.066  2.106  0.353
0.046  2.921  0.476
1.996  1.005 -0.032
```

la nuova matrice delle co-occorrenze fra termini LL^T è

```
t1 4.564 6.317 1.973
t2 6.317 8.761 3.012
t3 1.973 3.012 4.995

che rispetto alla matrice delle co-occorrenze originaria

t1 5.000 6.000 0.000
t2 6.000 9.000 0.000
t3 2.000 3.000 4.000
```

assegna una probabilità di co-occorrenza anche a t1-t3 e t2-t3

Soluzione

Come spiegato a lezione (si vedano le formule), al primo passo si assume pi=0,5, e, per N>>S, ri≈ni/N
Con queste semplificazioni, ci≈idf(wi)

Dunque, al primo passo:

1) Stima approssimata delle ri: ri=ni/N, N=9 (si noti che in realtà in questo esempio N è molto piccolo, la semplificazione non dovrebbe essere valida..)

Si ha: r1=6/9 r2=3/9 r3=3/9 r4=4/9 r5=2/9 r6=4/9 r7=4/9 r8=6/9 r9=4/9

Al passo 1, pi=0,5 per ogni i. e dunque:

RTV(1)=log9/3 + log9/3=2x0,477=0,954
RTV(2)=log9/6+log9/3=0,107+0,477=0,584
RTV(3)=log=0,107+0,477+log(9/4)=0,584+0,352=0,936
RTV(4)=0,584
RTV(5)=0,584
RTV(6)=0,107+0,352=0,459
RTV(7)=0,107
RTV(8)=0,352
RTV(9)=0,477+0,352=0,829

Poichè abbiamo stabilito |V|=4, vengono presentati: D1, D3,D9,D2

2)Seconda iterazione. Si ri-stimano le pi e ri a partire dai 4 documenti selezionati, usando per pi e ri le formule nel testo dell'esercizio.

Le nuove stime dei pi al passo 2 sono:
p1= $\frac{2+\frac{1}{2}}{4+1}=0,5$ p2= $\frac{3+0,5}{5}=0,7$ p3=p1 p4=p1

e per ri si ha:

r1= $\frac{6-2+\frac{1}{2}}{9-4+1}=\frac{4+0,5}{6}=0,75$, r2= $\frac{3-3+0,5}{6}=0,0883$, r3= $\frac{4-2+0,5}{6}=0,416$,
r4= $\frac{4-2+0,5}{6}=0,416$

Dunque avrò: c1= $\log\frac{p1(1-r1)}{r1(1-p1)}=\log\frac{0,5\cdot 0,25}{0,75\cdot 0,5}=-0,481$, c2= $\log\frac{0,7\cdot 0,91}{0,09\cdot 0,3}=1,37$
c3= $\log\frac{0,5\cdot 0,58}{0,42\cdot 0,5}=\log 1,38=0,14$, c4=c3

Notate come la presenza del primo termine risulta penalizzante: infatti è poco discriminante (elevato IDF) e si distribuisce equamente nei documenti rilevanti e non rilevanti (nello step 1).

Al contrario, il termine più discriminante è c2, che ha un'alta probabilità di comparire fra i rilevanti (in 3 documenti di V su 4).

Ricalcoliamo dunque RTV per i vari documenti:

RTV(1)=c2+c3=1,37+0,14=1,51
RTV(2)=c1+c2=-0,481+1,37=0,889
RTV(3)=c1+c3+c4=-0,481+0,14+0,14=0,201
...
RTV(8)=c4=0,14
RTV(9)=c2+c4=1,37+0,14=1,77

Verrebbero presentati: D9,D1, D2, D3

Come si vede, i documenti restano gli stessi, ma con diverso ranking. Poichè i documenti sono gli stessi, ovviamente ci si arresta, la terza iterazione genererebbe gli stessi valori di RTV.