Esercizio 1

D1: You say goodbye, I say hello
D2: You say stop, I say go
D3: Hello, hello, you say goodbye
D4: I say high, you say low

Costruire l'inverted index di questi documenti
Scegliete le stop words

Per la Query: "you goodbye" descrivete il processo di ricerca dell'inverted index e l'ordinamento secondo il modello vettoriale

Esercizio 2

QUERY: **nuovi farmaci epilessia --> nuov farmac epilie**

	testo	Manual rank	Norm. Word frequency rank	Rank booleano
D1	La terapia **farmaco**logica dell' **epilessia** può essere divisa, dal punto di vista storico, in due fasi, ovvero quella dei "vecchi" (dal fenobarbital - 1912, all'acido valproico - 1964), e quella dei "**nuovi**" **farmaci** (dal vigabatrim degli anni '80 ai più recenti). Questi ultimi, tuttavia vanno ulteriormente suddivisi in "**nuovi**" e "**nuovi**ssimi", in quanto negli ultimi anni sono apparsi sul mercato diversi **farmaci** con caratteristiche di meccanismo d'azione ed efficacia molto interessanti.			
D2	Dall'inizio degli anni '90 sono			

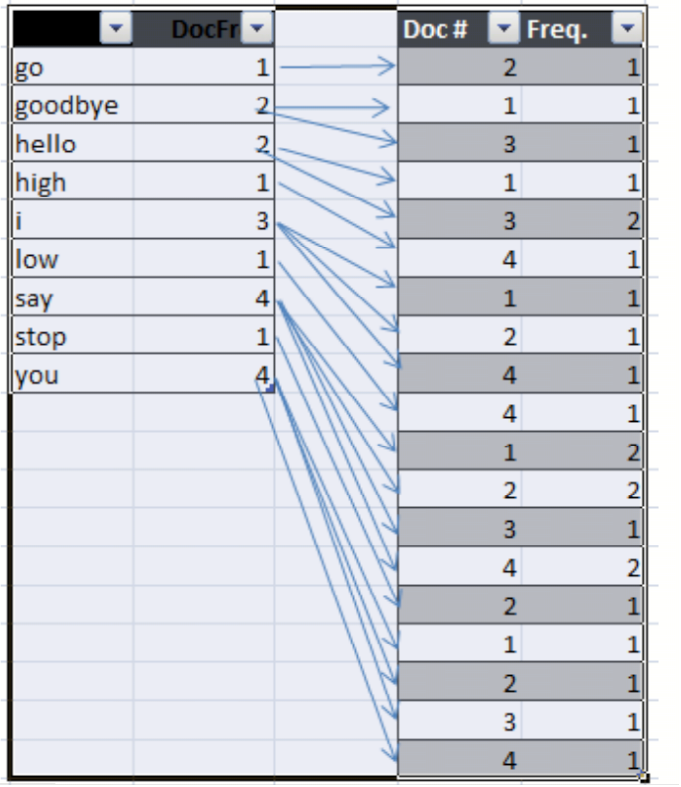
	stati commercializzati quasi una decina di **nuovi farmaci** anti**epilett**ici, dopo quasi 20 anni di silenzio in questo campo. I **nuovi farmaci** hanno dimostrato globalmente efficacia almeno pari a quella dei vecchi **farmaci**, ma sicuramente una minore incidenza di effetti collaterali e quindi maggiore tollerabilità da parte del paziente. Esistono poi alcune forme particolari di **epilessia** come gli Spasmi Infantili e la Sindrome di Lennox-Gastaut in cui con i **nuovi farmaci** si ottengono risultati sicuramente migliori che in passato.			
D3	Si può curare **l'epilessia**? Certamente sì, anche se solo nel 60-70% dei casi. Si usano infatti **farmaci** che controllano e bloccano la tendenza delle cellule cerebrali a produrre scariche **epiletti**che.			
D4	I classici **farmaci** anti**epilett**ici sono valproato e carbamazepina (che sono spesso i farmaci di prima scelta), fenitoina e fenobarbital. Da pochi anni è disponibile una serie di **farmaci** di **nuova** generazione (felbamato, gabapentin, lamotrigina, levetiracetam, oxcarbazepina, tiagabina, topiramato, vigabatrin) usati per indicazioni particolari o per aumentare l'efficacia della terapia quando la monoterapia			

	con un anti**epiletti**co classico non riesce a sopprimere le crisi **epiletti**che.			
D5	I nuovi progressi nella ricerca aprono scenari interessanti per tutte le persone con **epilessia** - ha precisato il prof. Emilio Perucca, Professore Ordinario di Farmacologia, Facoltà di Medicina e Chirurgia			
D6	All'Istituto di ricerche farmacologiche Mario Negri di Milano, è attualmente in studio una **nuova** terapia genica contro **l'epilessia**.			
D7	**L'epilessia** ha una prevalenza di quasi il 10 per 1.000: in Italia si contano circa 500.000 soggetti epilettici ed ogni anno si verificano circa 25.000 **nuovi** casi.			
D8	L'incidenza è di 50 **nuovi** casi per ogni 100.000 persone ogni anno. La terapia **farmaco**logica dell' **epilessia** viene sempre iniziata con un solo **farmaco**..			
D9	**Epilessia** sotto controllo. Molti malati non prendono i **farmaci** nel modo giusto, cn il rischio di avere così **nuovi** attacchi del disturbo.			
D10	Ricercatori dell'IWK Health Centre di Halifax, in Canada, hanno valutato l'incidenza di **epilessia** intrattabile in seguito ad interruzione del trattamento con farmaci antiepilettici in bambini			

	epilettici della **Nuova** Scozia.			

1) confrontate rank manuale, rank basato su normalized word frequency, e rank booleano. Elencate le vostre idee per fare di meglio, sulla base della vostra interpretazione delle esigenze di chi fa la query.
2) costruite l'inverted index di questi documenti per il seguente vocabolario: *bloccare, caso, cellula,crisi,cura,efficacia, epilessia,farmaco,farmacologia,epilessia, epilettico,malato,malattia,studio,terapia,trattamento,risultato*
3) *costruite un "trie" del vocabolario*

Soluzioni
Es. 1



	DocFr
go	1
goodbye	2
hello	2
high	1
i	3
low	1
say	4
stop	1
you	4

Doc #	Freq.
2	1
1	1
3	1
1	1
3	2
4	1
1	1
2	1
4	1
4	1
1	2
2	2
3	1
4	2
2	1
1	1
2	1
3	1
4	1

tf	**D1**	**D2**	**D3**	**D4**
you	1	1	1	1
goodbye	1	0	1	0

Esercizio 2

Bloccare	1	D3,1
Caso	3	D3,1-D7,1-D8-1
Cellula	1	D3-1
Crisi	1	D4-1
Cura	1	D3-1
Efficacia	3	D1,1-D2,1-D4,1
Epilessia	9	D1,1-D2,1-D3,1-D5,1-D6,1-D7,1-D8,1-D9,1-D10,1
Epilettico	5	D2,1-D3,1-D4,3-D7,1-D10,3
Farmaco	6	D1,2-D2,4-D3,1-D4,3-D8,1-D9,1
Farmacologia	3	D1,1-D5,1-D8,1
Malato	1	D9,1
Malattia	0	
Studio	1	D6,1
Terapia	4	D1,1-D4,1-D6,1-D8,1-D10,1
Trattamento	1	D10,1
Risultato	1	D2,1

Vettori dei documenti e della query, e calcolo del rank usando $w_{ij} = \frac{freq(w_i \ in \ D_j)}{\max_k(freq(w_k \ in \ D_j))}$ (senza idf)

	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	Query
Bloccare			1								
Caso			1				1	1			
Cellula			1								
Crisi				1							
Cura			1								
Efficacia	1	1		1							
Epilessia	1	1	1		1	1	1	1	1	1	
Epilettico		**0,33**	**0,33**	**1**			**0,33**		**0,33**	**1**	**1**
Farmaco	**0,5**	**1**		**0,75**				**1**	**1**		**1**
Farmacologia	1				1			1			
Malato									1		
Malattia											
Studio						1					
Terapia	1			1		1		1		1	
Trattamento										1	
Risultato		1									
$\|D_i\|^2$	4,25	4,1	5,1	4,56	2	3	2,1	5	3,1	4	2
|Di|	2,06	2,02	2,258	2,13	1,41	1,73	1,44	2,23	1,76	2	1,41
$D_i \bullet Q$	**0,5**	**1,33**	**0,33**	**1,75**	**0**	**0**	**0,33**	**1**	**1,33**	**1**	
$\|D_i\| \cdot \|Q\|$	**2,91**	**4,33**	**4,83**	**4,56**	**3,02**	**3,70**	**3,10**	**4,78**	**3,76**	**4,28**	
COS-SIM	**0,17**	**0,30**	**0,068**	**0,38**	**0**	**0**	**0,10**	**0,20**	**0,35**	**0,23**	

(Nota: potrebbe esserci qualche errore di calcolo..)

Formula: $\frac{w_{ij} \cdot w_{iq}}{\sqrt{\Sigma w_{ij}^2} \cdot \sqrt{\Sigma w_{iq}^2}}$