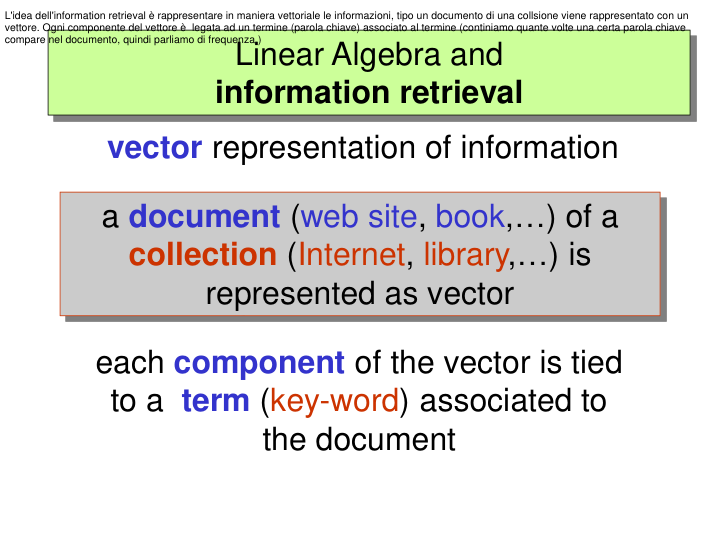
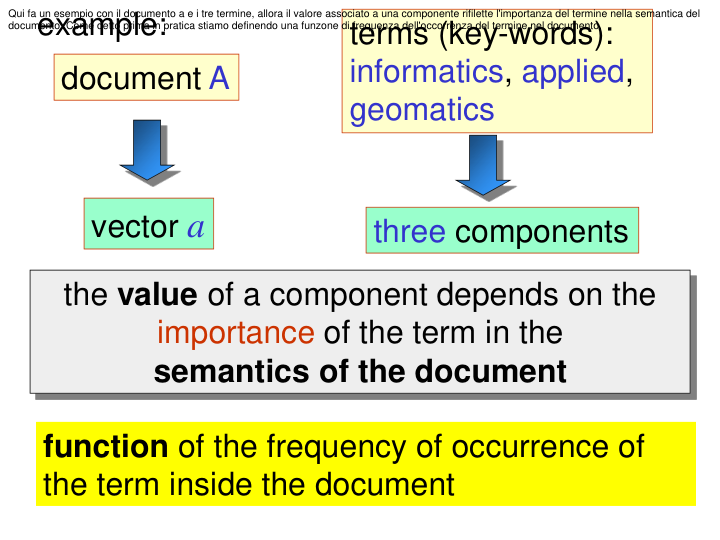
Lezione 4

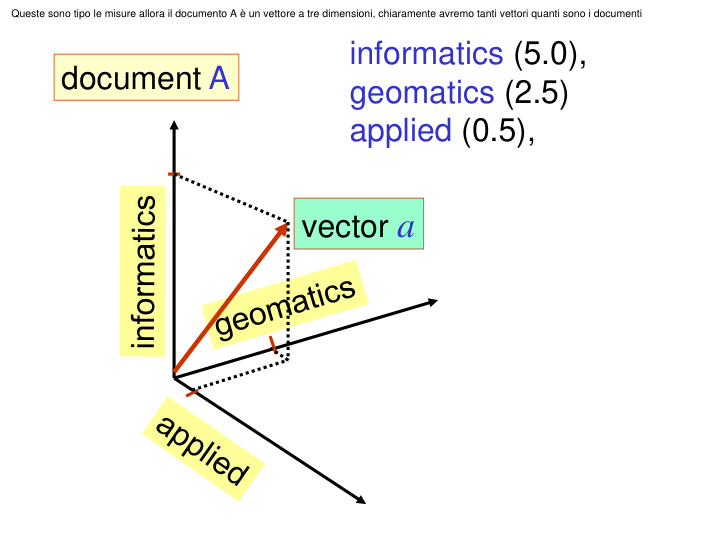
L'idea dell'information retrieval è rappresentare in maniera vettoriale le informazioni, tipo un documento di una collsione viene rappresentato con un vettore. Ogni componente del vettore è legata ad un termine (parola chiave) associato al termine (continiamo quante volte una certa parola chiave compare nel documento, quindi parliamo di frequenza.)



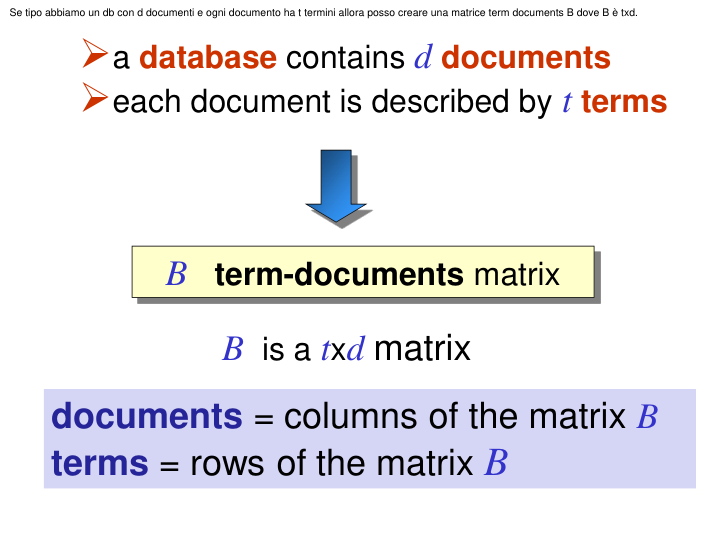
Qui fa un esempio con il documento a e i tre termine, allora il valore associato a una componente rifilette l'importanza del termine nella semantica del documento. Come detto prima in pratica stiamo definendo una funzone di frequenza dell'occorrenza del termine nel documento.



Queste sono tipo le misure allora il documento A è un vettore a tre dimensioni, chiaramente avremo tanti vettori quanti sono i documenti

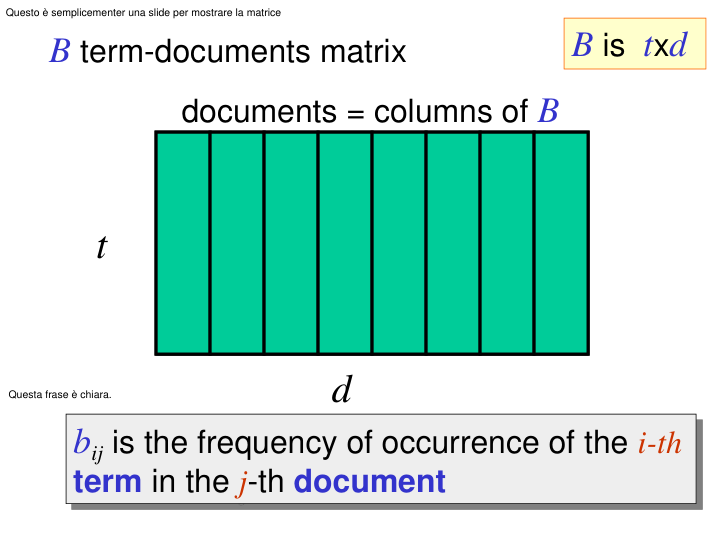


Se tipo abbiamo un db con d documenti e ogni documento ha t termini allora posso creare una matrice term documents B dove B è txd.

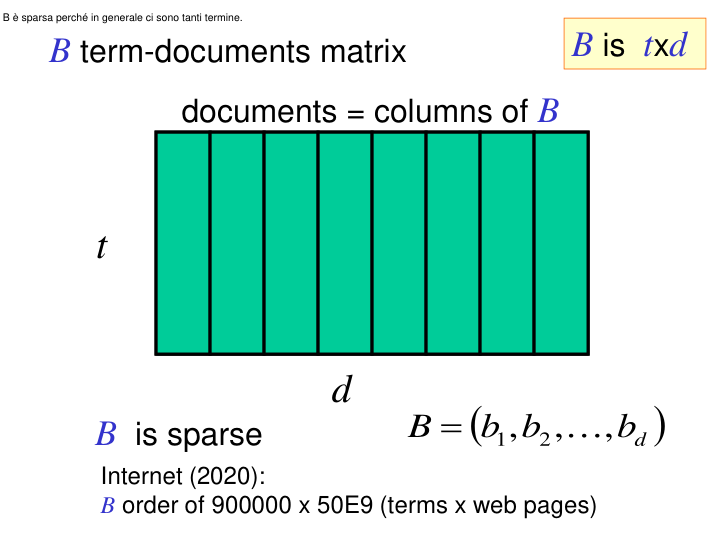


Questo è semplicementer una slide per mostrare la matrice

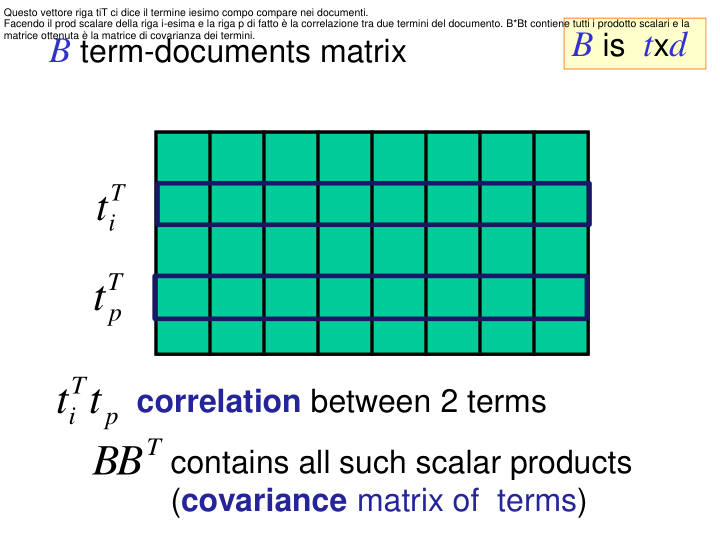
Questa frase è chiara.



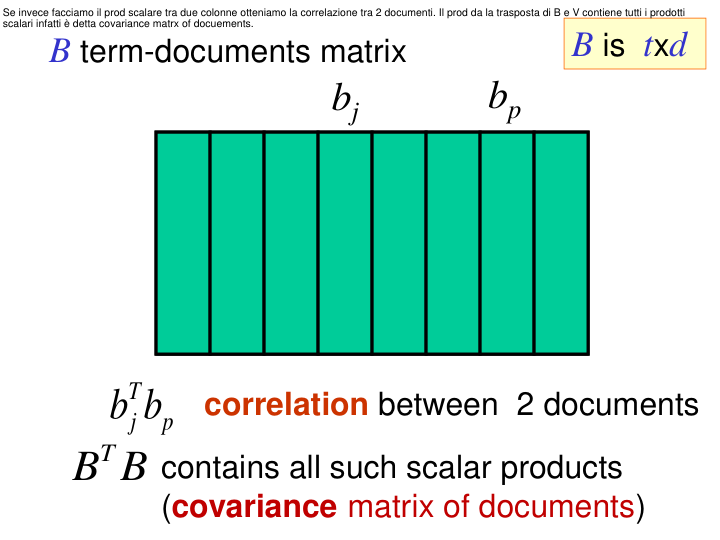
B è sparsa perché in generale ci sono tanti termine.



Questo vettore riga tiT ci dice il termine iesimo compo compare nei documenti.  
Facendo il prod scalare della riga i-esima e la riga p di fatto è la correlazione tra due termini del documento. B\*Bt contiene tutti i prodotto scalari e la matrice ottenuta è la matrice di covarianza dei termini.

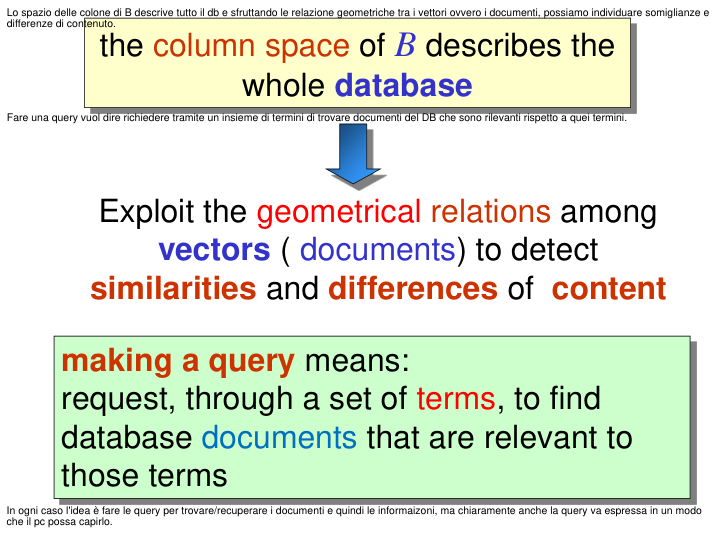


Se invece facciamo il prod scalare tra due colonne otteniamo la correlazione tra 2 documenti. Il prod da la trasposta di B e V contiene tutti i prodotti scalari infatti è detta covariance matrx of docuements.



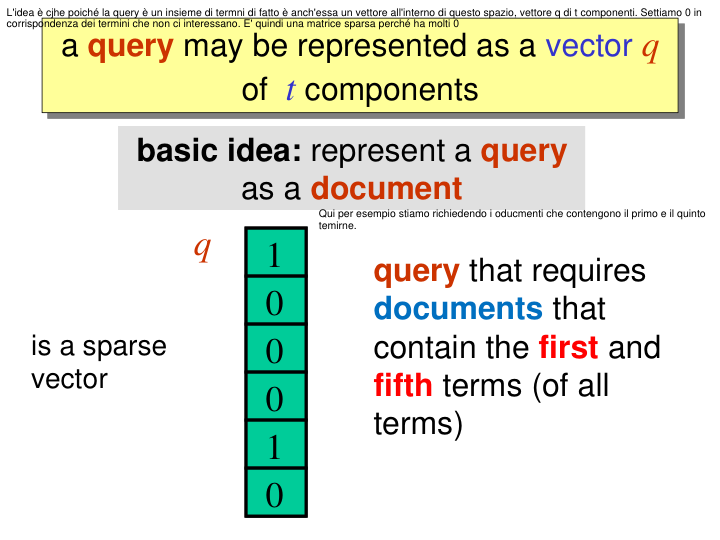
Lo spazio delle colone di B descrive tutto il db e sfruttando le relazione geometriche tra i vettori ovvero i documenti, possiamo individuare somiglianze e differenze di contenuto.  
  
  
  
  
  
  
  
Fare una query vuol dire richiedere tramite un insieme di termini di trovare documenti del DB che sono rilevanti rispetto a quei termini.

In ogni caso l'idea è fare le query per trovare/recuperare i documenti e quindi le informaizoni, ma chiaramente anche la query va espressa in un modo che il pc possa capirlo.

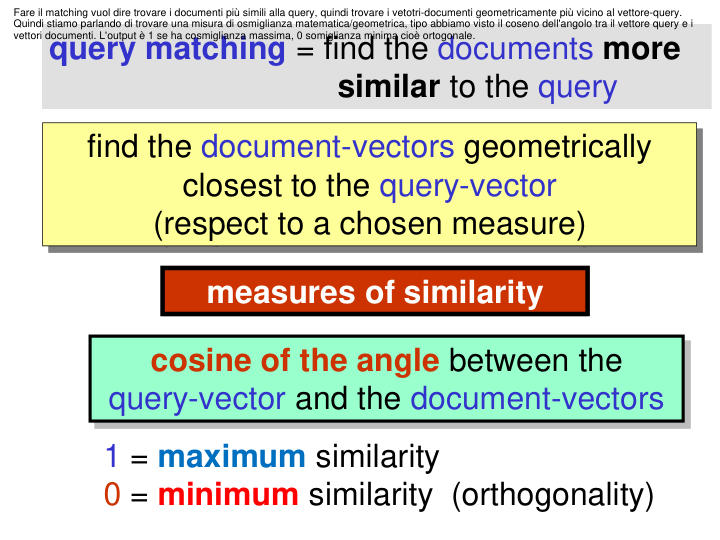


L'idea è cjhe poiché la query è un insieme di termni di fatto è anch'essa un vettore all'interno di questo spazio, vettore q di t componenti. Settiamo 0 in corrispondenza dei termini che non ci interessano. E' quindi una matrice sparsa perché ha molti 0

Qui per esempio stiamo richiedendo i oducmenti che contengono il primo e il quinto temirne.

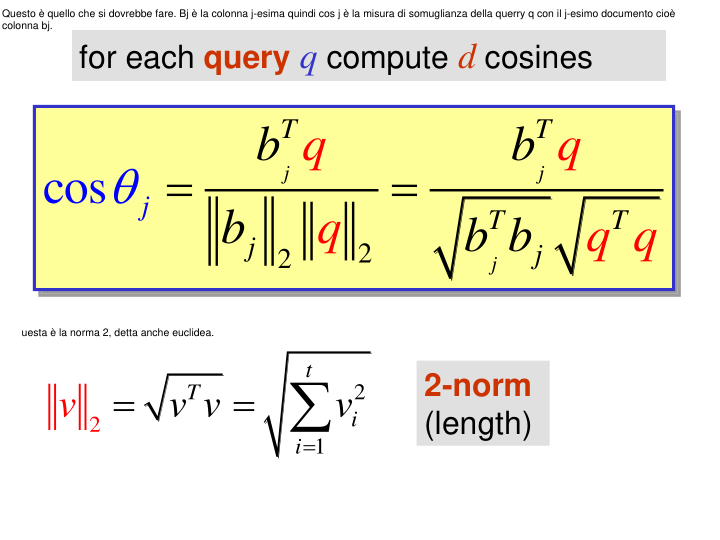


Fare il matching vuol dire trovare i documenti più simili alla query, quindi trovare i vetotri-documenti geometricamente più vicino al vettore-query. Quindi stiamo parlando di trovare una misura di osmiglianza matematica/geometrica, tipo abbiamo visto il coseno dell'angolo tra il vettore query e i vettori documenti. L'output è 1 se ha cosmiglianza massima, 0 somiglianza minima cioè ortogonale.

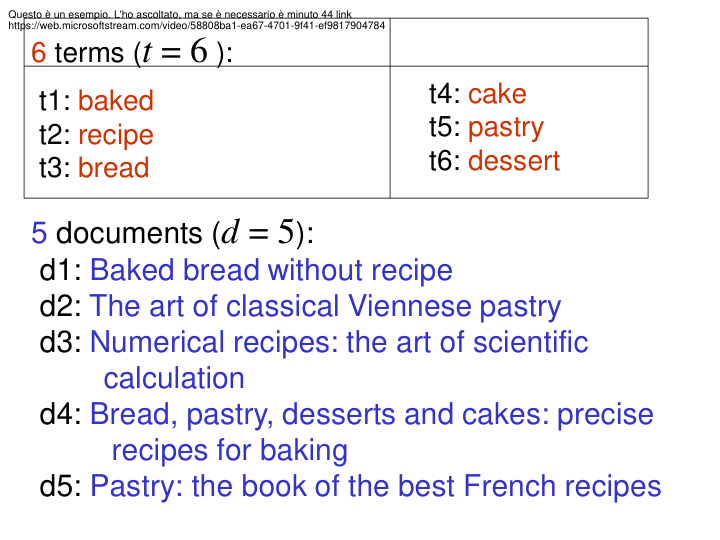


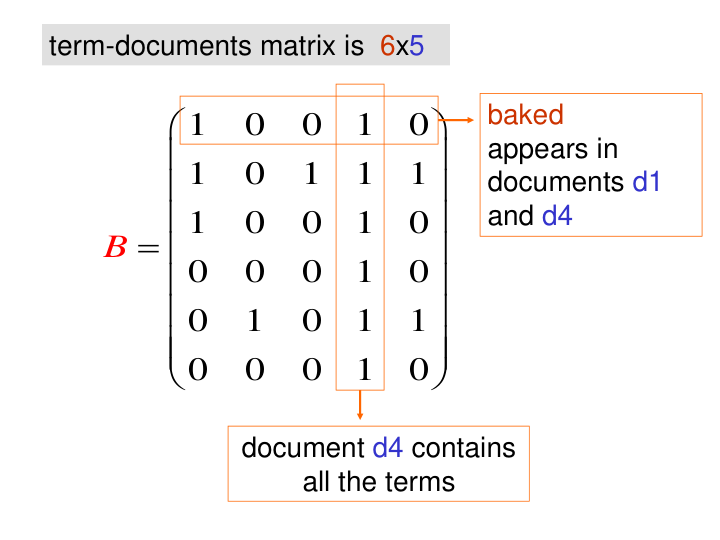
Questo è quello che si dovrebbe fare. Bj è la colonna j-esima quindi cos j è la misura di somuglianza della querry q con il j-esimo documento cioè colonna bj.

uesta è la norma 2, detta anche euclidea.

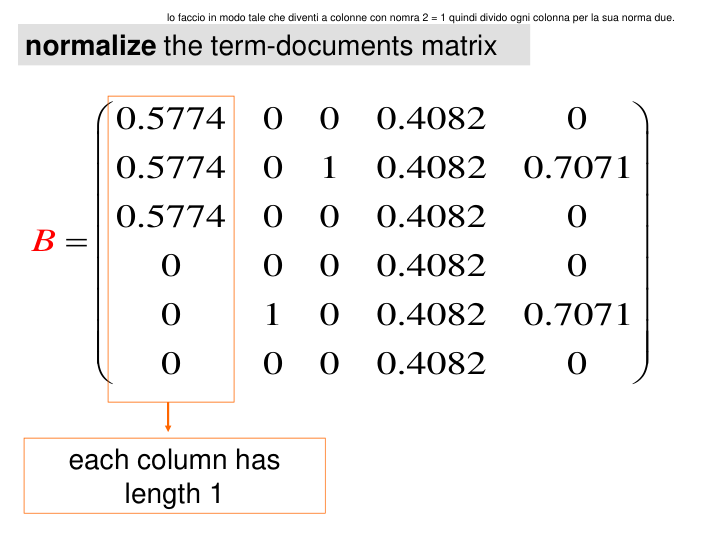


Questo è un esempio. L'ho ascoltato, ma se è necessario è minuto 44 link https://web.microsoftstream.com/video/58808ba1-ea67-4701-9f41-ef9817904784

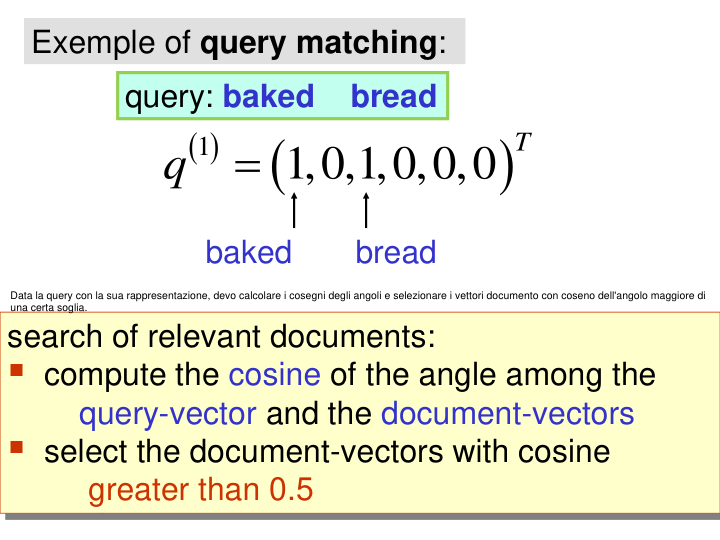




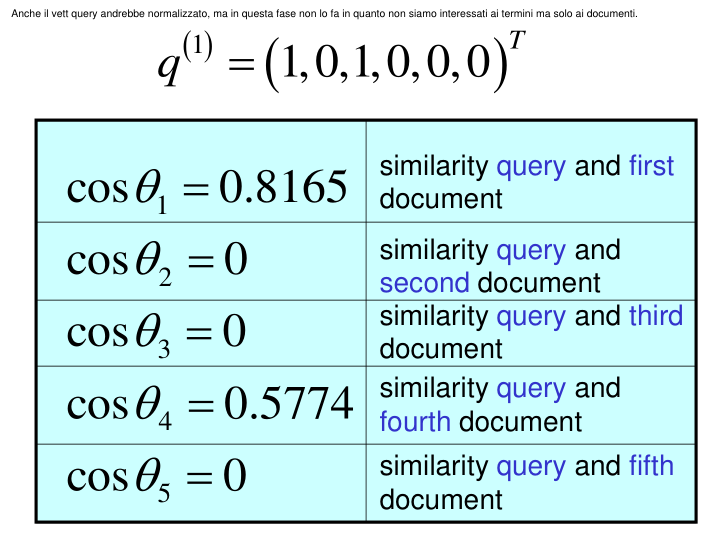
lo faccio in modo tale che diventi a colonne con nomra 2 = 1 quindi divido ogni colonna per la sua norma due.



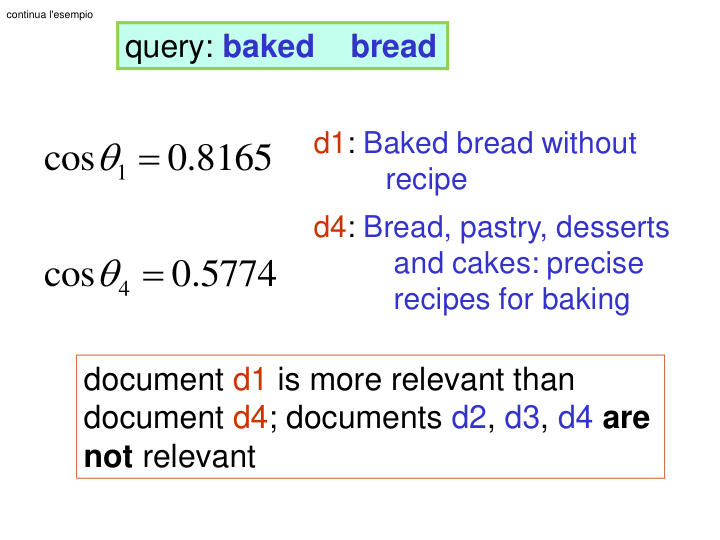
Data la query con la sua rappresentazione, devo calcolare i cosegni degli angoli e selezionare i vettori documento con coseno dell'angolo maggiore di una certa soglia.

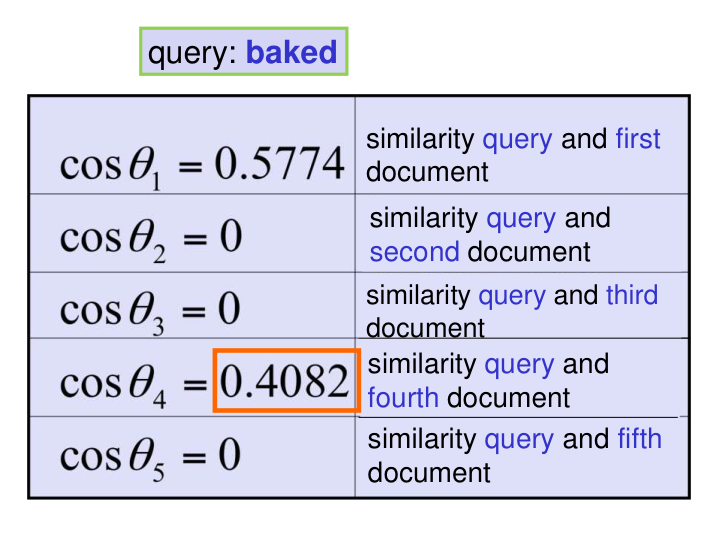


Anche il vett query andrebbe normalizzato, ma in questa fase non lo fa in quanto non siamo interessati ai termini ma solo ai documenti.

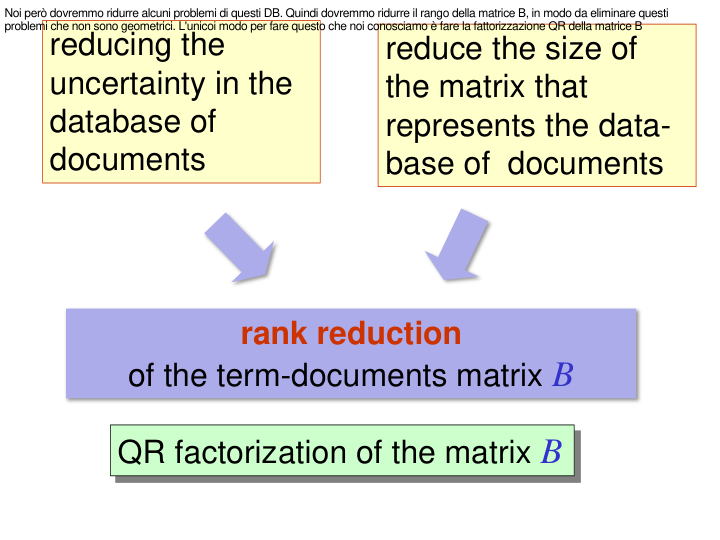


continua l'esempio



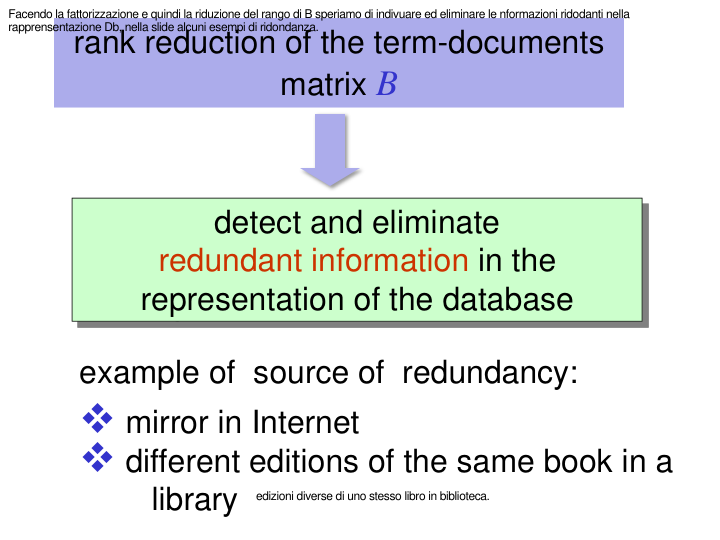


Noi però dovremmo ridurre alcuni problemi di questi DB. Quindi dovremmo ridurre il rango della matrice B, in modo da eliminare questi problemi che non sono geometrici. L'unicoi modo per fare questo che noi conosciamo è fare la fattorizzazione QR della matrice B

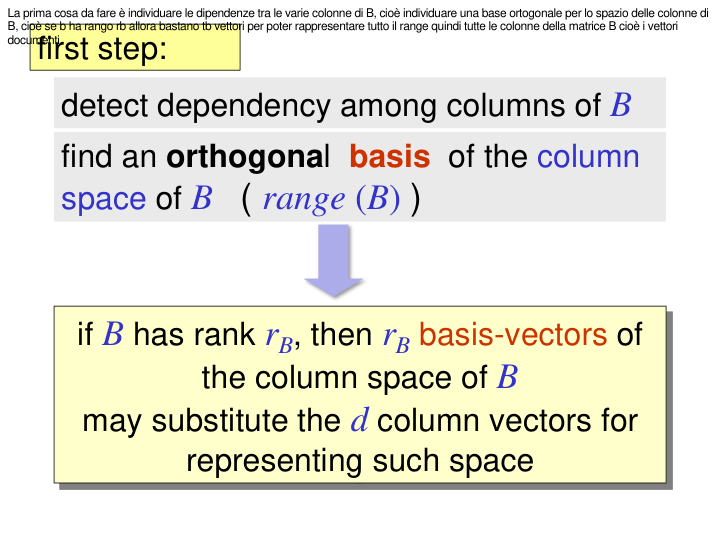


Facendo la fattorizzazione e quindi la riduzione del rango di B speriamo di indivuare ed eliminare le nformazioni ridodanti nella rapprensentazione Db, nella slide alcuni esempi di ridondanza.

edizioni diverse di uno stesso libro in biblioteca.

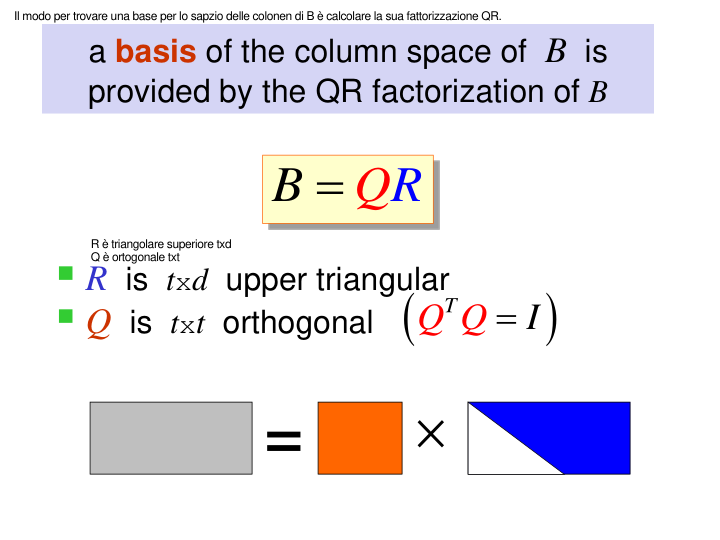


La prima cosa da fare è individuare le dipendenze tra le varie colonne di B, cioè individuare una base ortogonale per lo spazio delle colonne di B, cioè se b ha rango rb allora bastano tb vettori per poter rappresentare tutto il range quindi tutte le colonne della matrice B cioè i vettori documenti.



Il modo per trovare una base per lo sapzio delle colonen di B è calcolare la sua fattorizzazione QR.

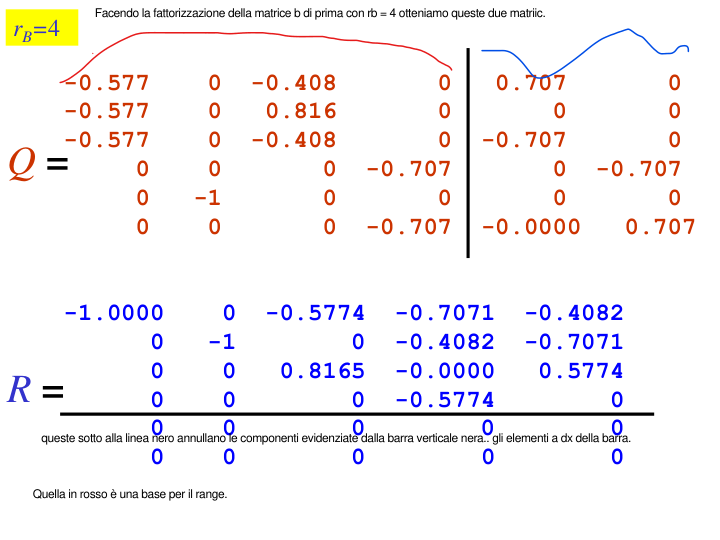
R è triangolare superiore txd  
Q è ortogonale txt



Facendo la fattorizzazione della matrice b di prima con rb = 4 otteniamo queste due matriic.

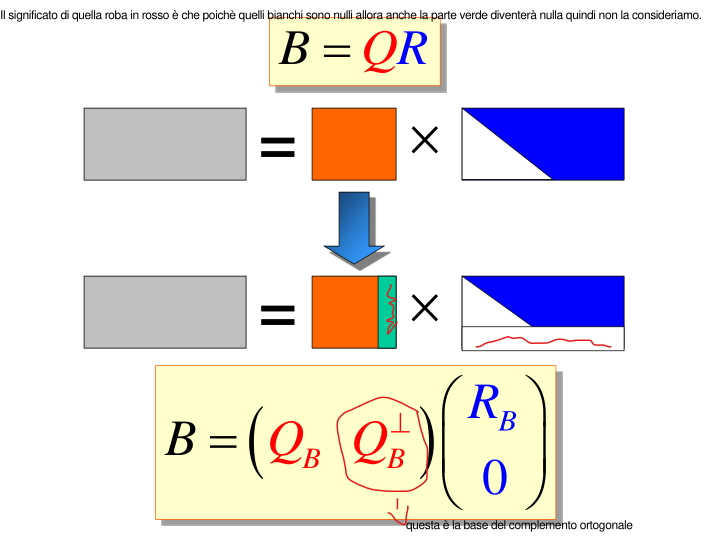
queste sotto alla linea nero annullano le componenti evidenziate dalla barra verticale nera.. gli elementi a dx della barra.

Quella in rosso è una base per il range.

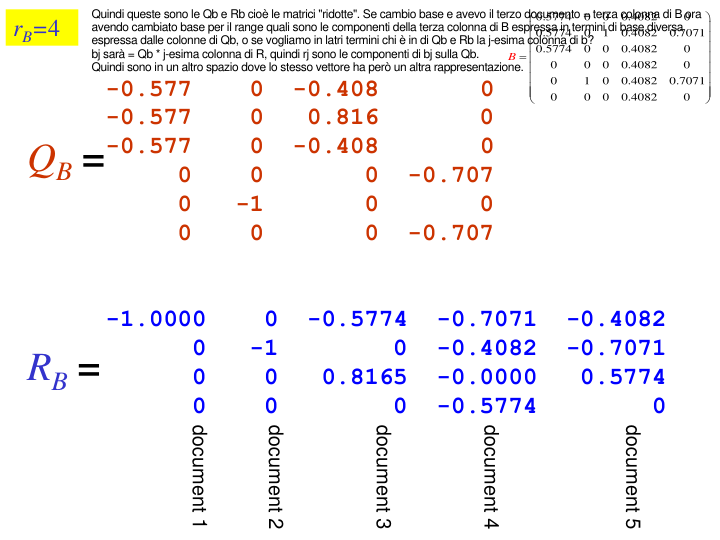


Il significato di quella roba in rosso è che poichè quelli bianchi sono nulli allora anche la parte verde diventerà nulla quindi non la consideriamo.

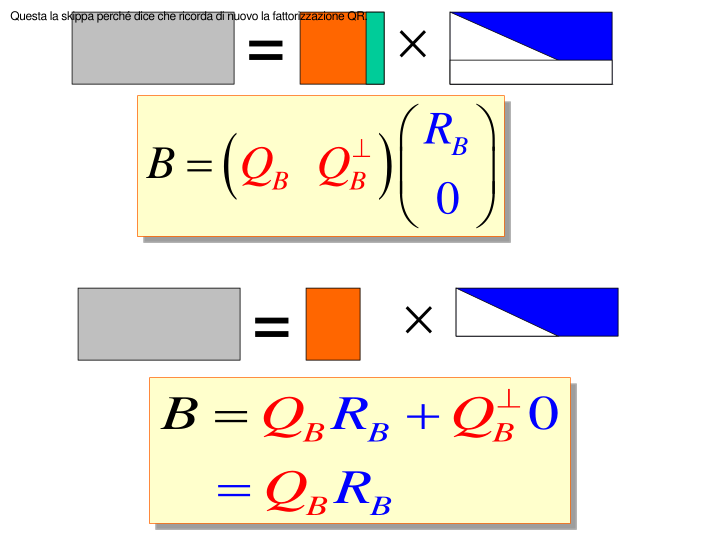
questa è la base del complemento ortogonale



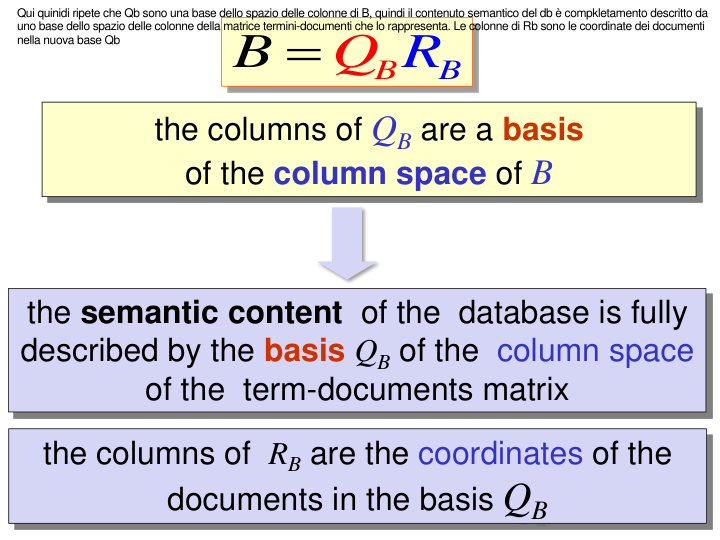
Quindi queste sono le Qb e Rb cioè le matrici "ridotte". Se cambio base e avevo il terzo documento = terza colonna di B ora avendo cambiato base per il range quali sono le componenti della terza colonna di B espressa in termini di base diversa espressa dalle colonne di Qb, o se vogliamo in latri termini chi è in di Qb e Rb la j-esima colonna di b?  
bj sarà = Qb \* j-esima colonna di R, quindi rj sono le componenti di bj sulla Qb.  
Quindi sono in un altro spazio dove lo stesso vettore ha però un altra rappresentazione.



Questa la skippa perché dice che ricorda di nuovo la fattorizzazione QR.



Qui quinidi ripete che Qb sono una base dello spazio delle colonne di B, quindi il contenuto semantico del db è compkletamento descritto da uno base dello spazio delle colonne della matrice termini-documenti che lo rappresenta. Le colonne di Rb sono le coordinate dei documenti nella nuova base Qb

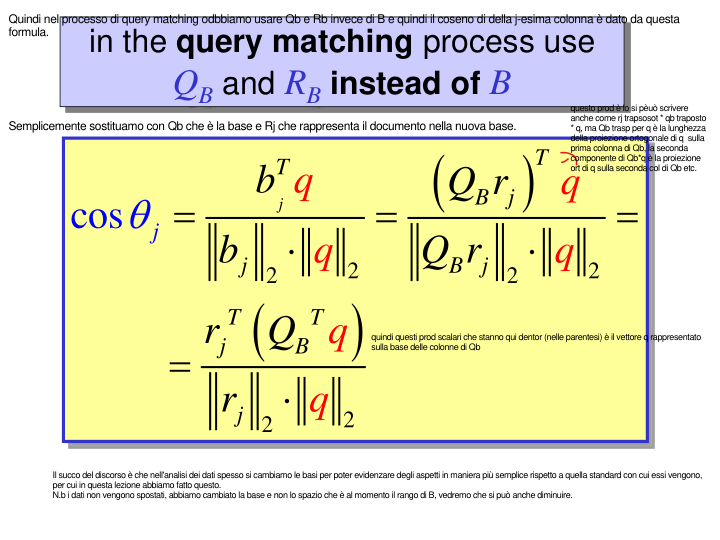


Quindi nel processo di query matching odbbiamo usare Qb e Rb invece di B e quindi il coseno di della j-esima colonna è dato da questa formula.   
  
  
  
  
  
  
Semplicemente sostituamo con Qb che è la base e Rj che rappresenta il documento nella nuova base.

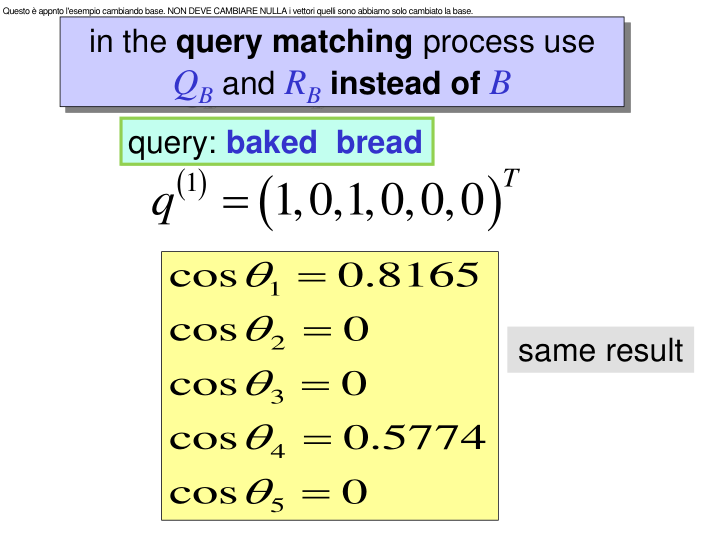
questo prod è lo si pèuò scrivere anche come rj trapsosot \* qb traposto \* q, ma Qb trasp per q è la lunghezza della proiezione ortogonale di q sulla prima colonna di Qb, la seconda componente di Qb\*q è la proiezione ort di q sulla seconda col di Qb etc.

quindi questi prod scalari che stanno qui dentor (nelle parentesi) è il vettore q rappresentato sulla base delle colonne di Qb

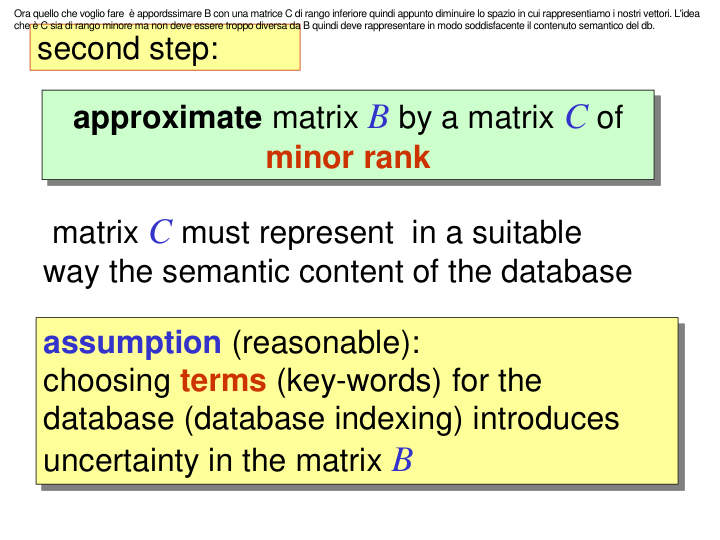
Il succo del discorso è che nell'analisi dei dati spesso si cambiamo le basi per poter evidenzare degli aspetti in maniera più semplice rispetto a quella standard con cui essi vengono, per cui in questa lezione abbiamo fatto questo.   
N.b i dati non vengono spostati, abbiamo cambiato la base e non lo spazio che è al momento il rango di B, vedremo che si può anche diminuire.



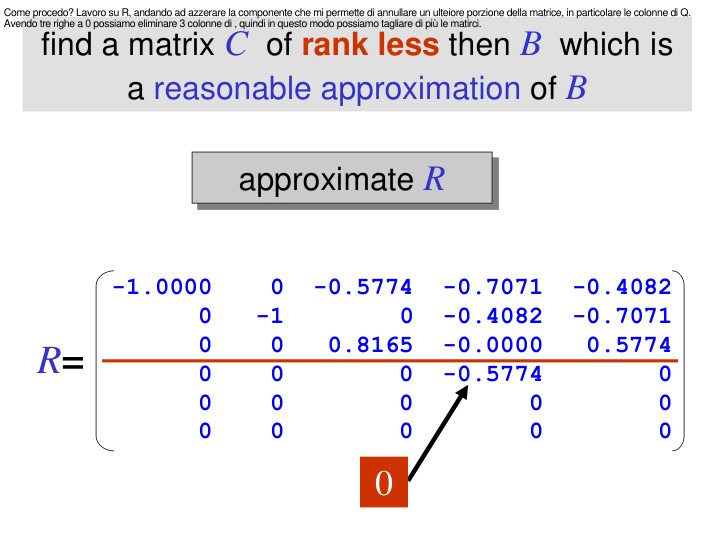
Questo è appnto l'esempio cambiando base. NON DEVE CAMBIARE NULLA i vettori quelli sono abbiamo solo cambiato la base.



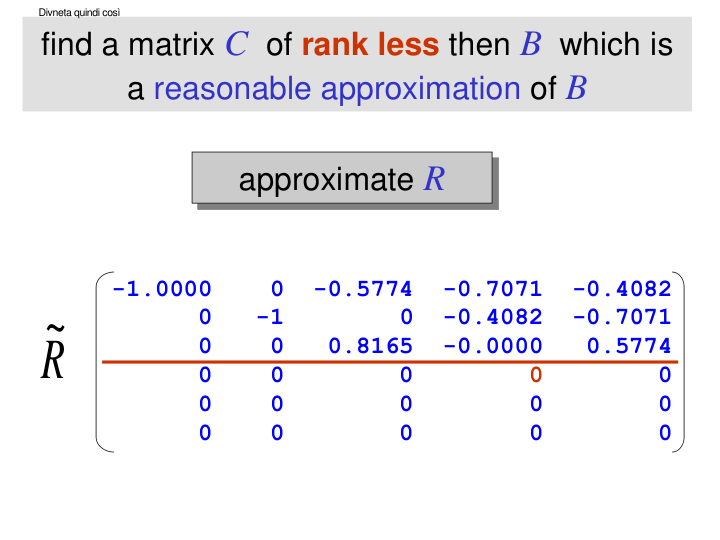
Ora quello che voglio fare è appordssimare B con una matrice C di rango inferiore quindi appunto diminuire lo spazio in cui rappresentiamo i nostri vettori. L'idea che è C sia di rango minore ma non deve essere troppo diversa da B quindi deve rappresentare in modo soddisfacente il contenuto semantico del db.



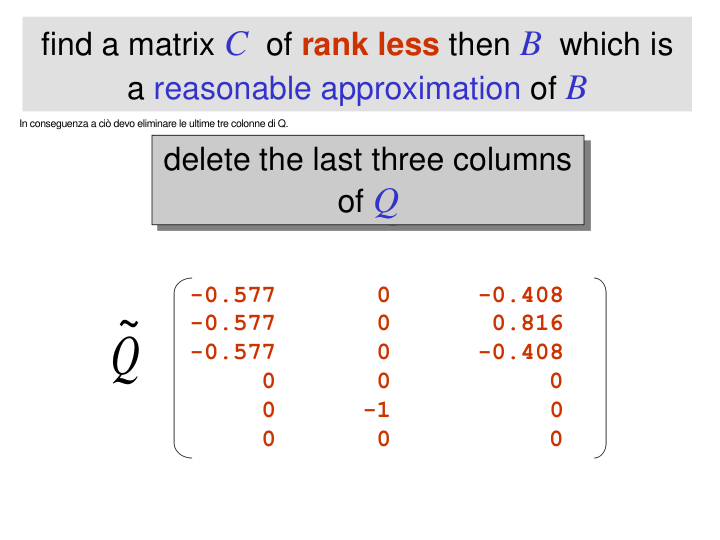
Come procedo? Lavoro su R, andando ad azzerare la componente che mi permette di annullare un ulteiore porzione della matrice, in particolare le colonne di Q.  
Avendo tre righe a 0 possiamo eliminare 3 colonne di , quindi in questo modo possiamo tagliare di più le matirci.



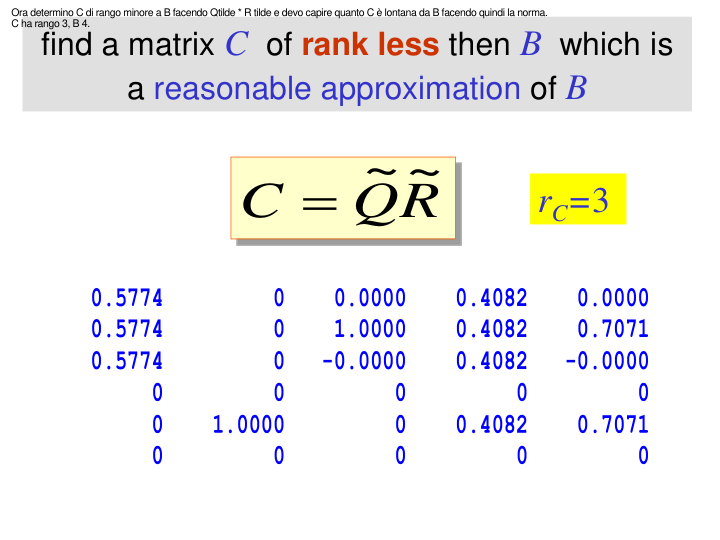
Divneta quindi così



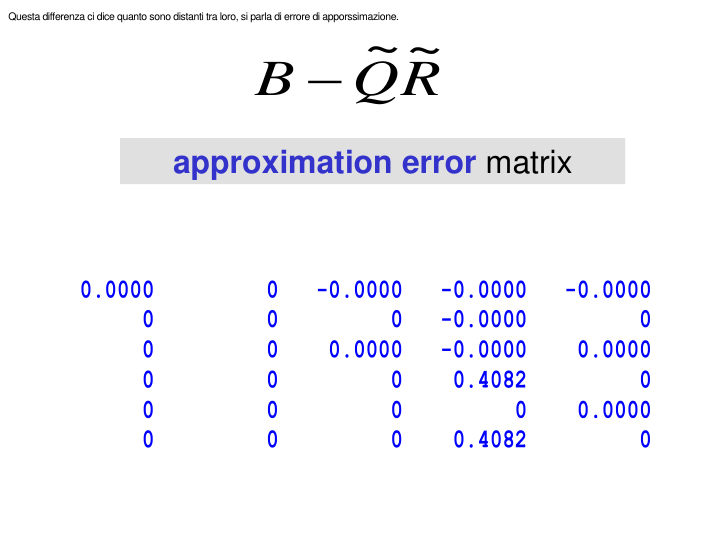
In conseguenza a ciò devo eliminare le ultime tre colonne di Q.



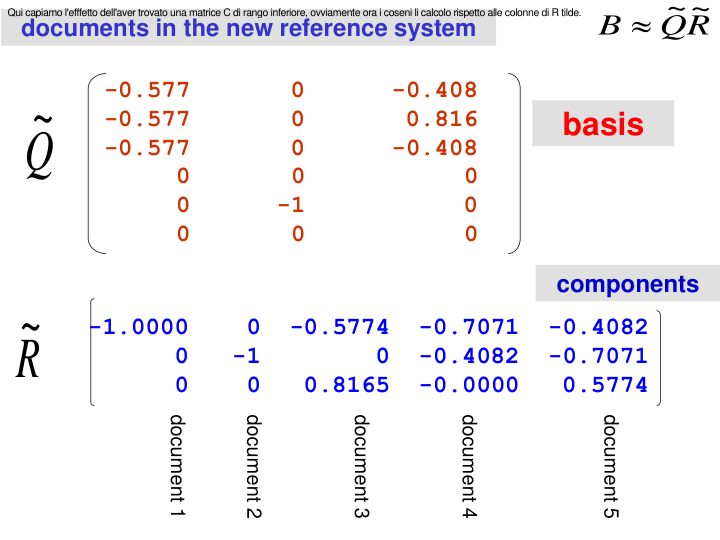
Ora determino C di rango minore a B facendo Qtilde \* R tilde e devo capire quanto C è lontana da B facendo quindi la norma.  
C ha rango 3, B 4.



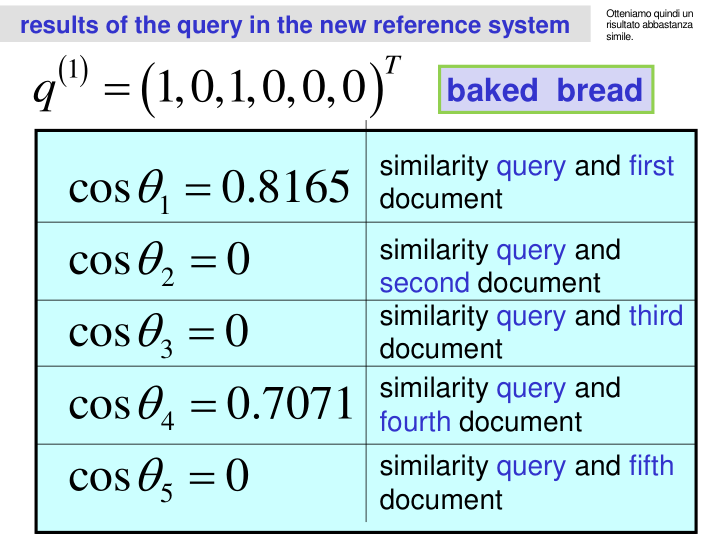
Questa differenza ci dice quanto sono distanti tra loro, si parla di errore di apporssimazione.



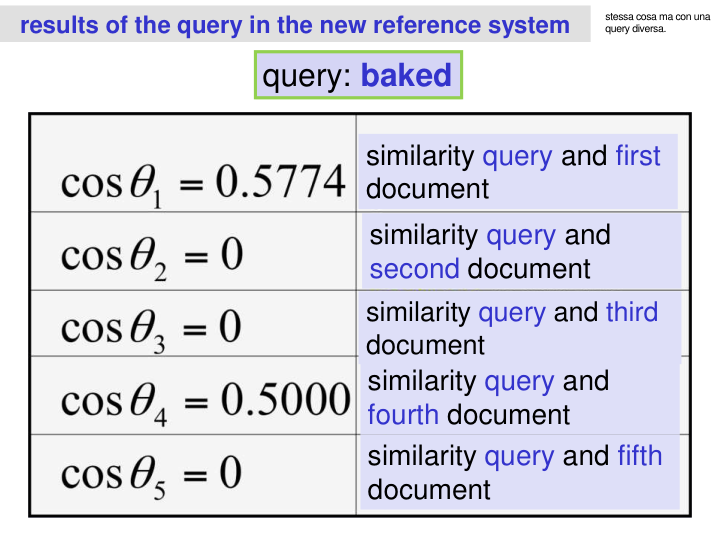
Qui capiamo l'efffetto dell'aver trovato una matrice C di rango inferiore, ovviamente ora i coseni li calcolo rispetto alle colonne di R tilde.



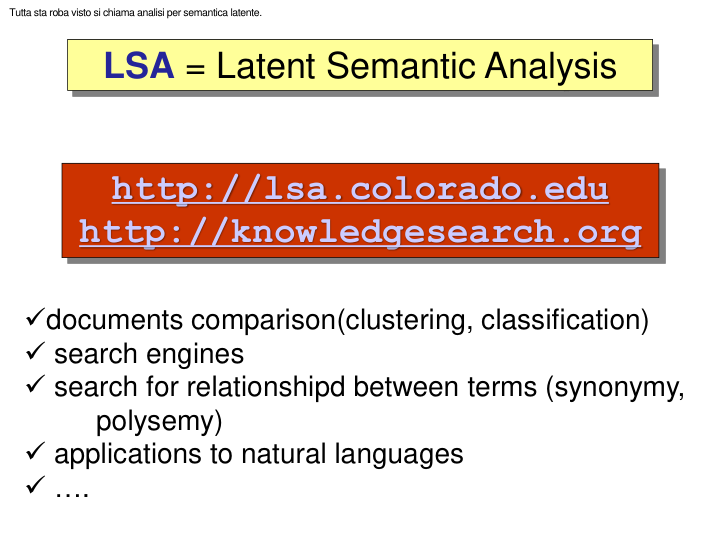
Otteniamo quindi un risultato abbastanza simile.



stessa cosa ma con una query diversa.



Tutta sta roba visto si chiama analisi per semantica latente.



Invece di fare la fattorizazizone QR possiamo usare la fattorizzazione SVD della matrice temrini documenti che consente di agire anche sullo spazio delle righe della matrice.

