Vector space models (VSM)

I modelli **vector-space** per l'*information retrieval* rappresentano una classe di tecniche caratterizzate da un modello matematico formale sottostante e da un abbinamento fra *query* e documento basato su caratteristiche individuali

I documenti sono trattati come insiemi di parole, l'importanza delle quali può essere opportunamente ponderata.

La base concettuale dei VSM è la premessa che si può derivare il significato di un documento dalle parole che lo costituiscono.

L'identificazione dei documenti rilevanti per la *query* avviene mediante il confronto della sua rappresentazione con quella di ciascun documento appartenente al *corpus* e può portare anche all'identificazione di documenti che non contengono i termini presenti nella *query*.

Vector space model
(o Term Vector Model)

- VSM rappresenta le *query* e i documenti scritti in linguaggio naturale come vettori in uno spazio multidimensionale
- Le parole rappresentano gli assi orto-normalizzati del sistema di riferimento e le *query* (Q) e i documenti D(n) sono visti come dei vettori
- Il punteggio di rilevanza dei documenti è misurato in termini di COSENI degli angoli formati fra vettori: un coseno pari a 0 significa che il vettore della *query* e quello del documento sono ortogonali: il documento non corrisponde minimamente alla *query*, perché non hanno parole in comune
- La versione classica di VSM (Salton, Wong&Yang, 1975) include un parametro che tiene conto della frequenza del temine nel documento (peso "locale") e un parametro che tiene conto del potere discriminante del termine all'interno del *corpus* (peso globale), mediante il già ricordato indice noto come tf-idf

Principali caratteristiche di VSM

- VSM non cerca di ridurre le dimensioni dello spazio, trattando ogni termine indipendentemente
- VSM è molto flessibile, poiché consente di ponderare ogni termine individualmente, così l'importanza di ciascun termine all'interno di un documento, o di un documento all'interno del *corpus* può essere valutata opportunamente
- Inoltre, si possono utilizzare misure di similarità differenti per confrontare *query* e documenti, così da enfatizzare o sottovalutare specifiche caratteristiche del *corpus*.

La scelta della misura di prossimità

- Se si utilizza come misura il **prodotto interno**, si calcola semplicemente la distanza euclidea fra i vettori.
- Se si sceglie, invece, il **coseno dell'angolo** formato fra i vettori si decide di neutralizzare l'effetto della diversa lunghezza dei documenti

  In moltii casi, la direzione dei vettori è l'indicazione più importante per l'identificazione di similarità semantiche fra i documenti (o query e documenti) rispetto alla loro distanza nello spazio parole-documenti

I punti critici di VSM sono i seguenti:

1. I *documenti lunghi* sono mal rappresentati perché hanno prodotti scalari piccoli e una dimensione grande

2. I documenti con *contesti simili* ma con vocabolario diverso formano angoli con coseni piccoli ("*False negative match*")

3. La ricerca di *keywords* scritte in modo non appropriato danno risultati poveri ("*False positive match*")

4. Limiti semantici

Latent Semantic Indexing

Il modello di IR noto come Latent Semantic Indexing si basa sull'operazione di algebra matriciale, denominata **Decomposizione in Valori Singolari** (Eckart&Young, 1936) con l'obiettivo di ridurre la dimensionalità dello spazio parole-documenti, cercando, inoltre, di risolvere almeno in parte i problemi di polisemia e sinonimia che affliggono la maggior parte delle tecniche di IR.

LSI rappresenta esplicitamente parole e documenti in uno spazio multidimensionale, consentendo di far emergere le relazioni semantiche "latenti" fra parole e documenti

Principali caratteristiche di LSI

- LSI si basa sull'idea che le parole che costituiscono un documento suggeriscano il contenuto semantico del documento stesso
- Il modello LSI vede le parole contenute in un documento come degli indicatori poco affidabili dei concetti contenuti nel documento
- LSI assume che la variabilità che caratterizza la scelta delle parole oscuri almeno parzialmente la struttura semantica del documento
- Attraverso la riduzione dimensionale dello spazio parole-documenti le relazioni semantiche sottostanti che esistono fra i documenti vengono ad essere rivelate e grand parte del "rumore" (differenze nell'utilizzo delle parole, termini poco discriminanti, etc.), invece, eliminato
- LSI analizza "statisticamente" le diverse strutture di utilizzo delle parole, guardando al *corpus* nella sua interezza e collocando i documenti con un utilizzo simile delle parole vicini nello spazio di dimensioni ridotte e consentendo anche a documenti semanticamente correlati di essere vicini, anche se non hanno parole in comune

La struttura dei dati in LSI

- Nel modello LSI, nella sua versione-base, parole e documenti sono rappresentati in una matrice **T** parole-documenti, le cui celle contengono la frequenza con cui una parola appare in un documento
- In genere **T** è una matrice molto sparsa
- Anche in LSI sono state proposte forme differenti di èonderazione degli elementi (locali) e (globali)