IIC 3800 Tópicos en CC
NLP

https://github.com/marcelomendoza/IIC3800

- VECTORIZACIÓN DE DOCUMENTOS Y RANKING -

Matriz términos-documentos

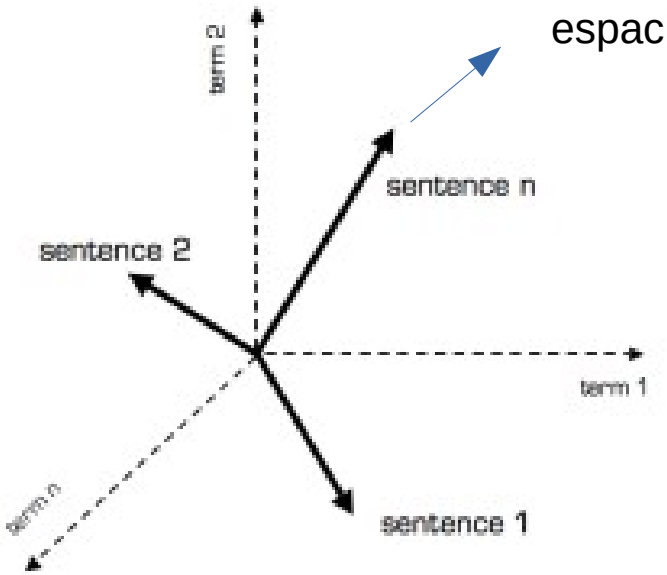
documentos →

	Antonio y Cleopatra	Julio Cesar	La Tempestad	Hamlet	Otelo	Macbeth	...
Antonio	157	73	0	0	0	1	
Brutus	4	157	0	2	0	0	
Cesar	232	227	0	2	1	0	
Calpurnia	0	10	0	0	0	0	
Cleopatra	57	0	0	0	0	0	
...		...					

vectorización

términos

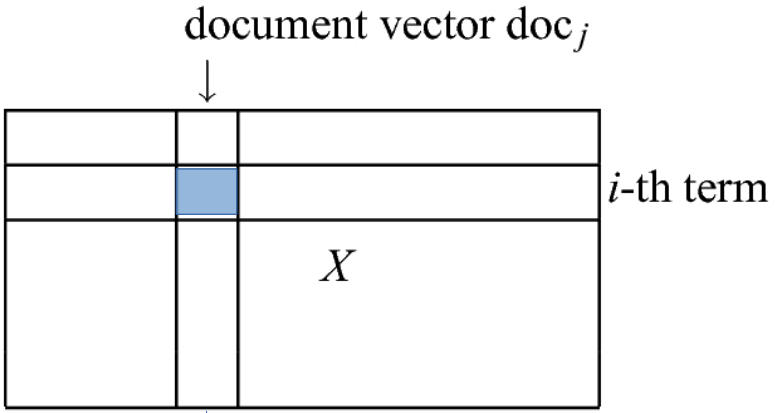
Vector-space model



Cada documento es un vector en el espacio vectorial de términos

Cada término es una dimensión en el espacio vectorial de términos

Vector-space model



$$\vec{d_j} = (w_{1,j}, \ldots, w_{i,j}, \ldots, w_{V,j})$$

Term scoring function:

$$f(t_i, d_j) = w_{i,j}$$

BM25

$f_{i,j}$: # occs. de ti en dj

N : # docs

n_i : # docs donde ti ocurre

$l(d_j)$: # tokens en dj

l_{avg} : largo promedio

$$w_{i,j} = \frac{f_{i,j} \cdot (k_1 + 1)}{k_1 \cdot \left[(1 - b) + b \cdot \frac{l(d_j)}{l_{avg}}\right] + f_{i,j}} \cdot \log\left(\frac{N}{n_i}\right)$$

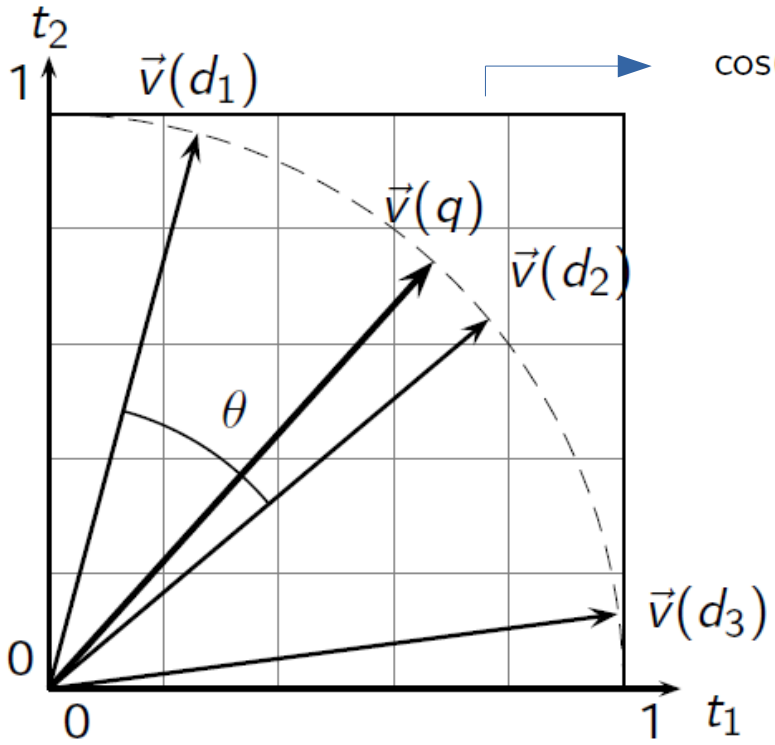
$$b \in [0, 1] \ , \quad k_1 > 0$$

Empírico: $b \approx 0.75$

$k_1 \approx 1.2$

Document ranking

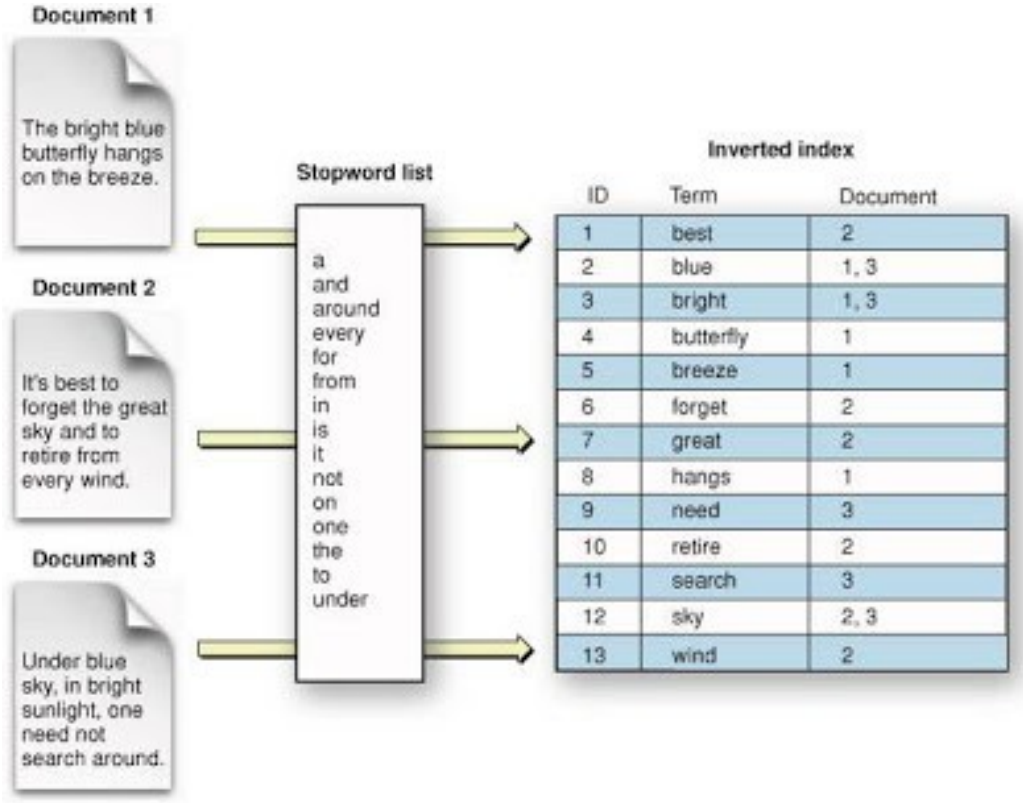
Funciones de proximidad entre vectores:



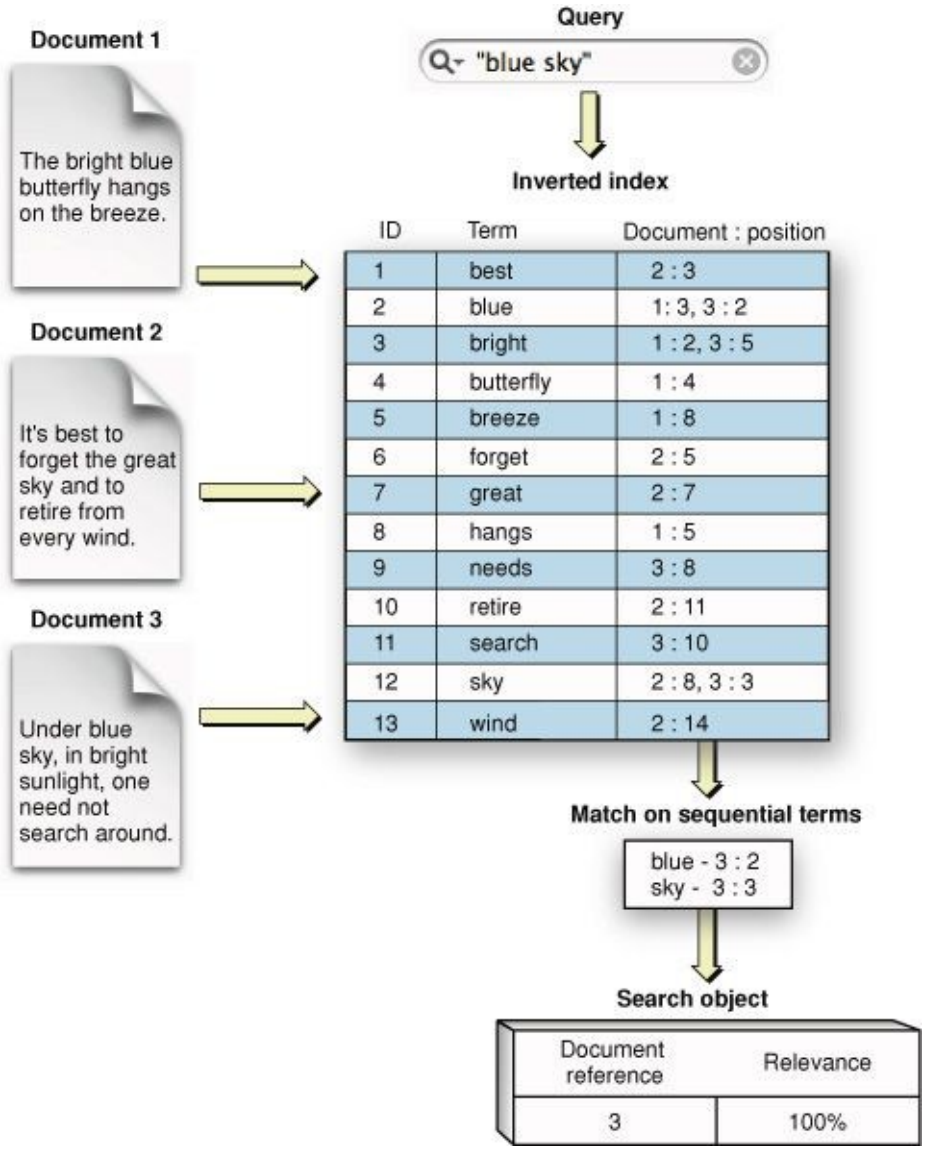
$$\cos(\vec{q},\vec{d}) = \text{SIM}(\vec{q},\vec{d}) = \frac{\vec{q}\cdot\vec{d}}{|\vec{q}||\vec{d}|} = \frac{\sum_{i=1}^{|V|} q_i d_i}{\sqrt{\sum_{i=1}^{|V|} q_i^2}\sqrt{\sum_{i=1}^{|V|} d_i^2}}$$

BM25 está basado en esta idea.

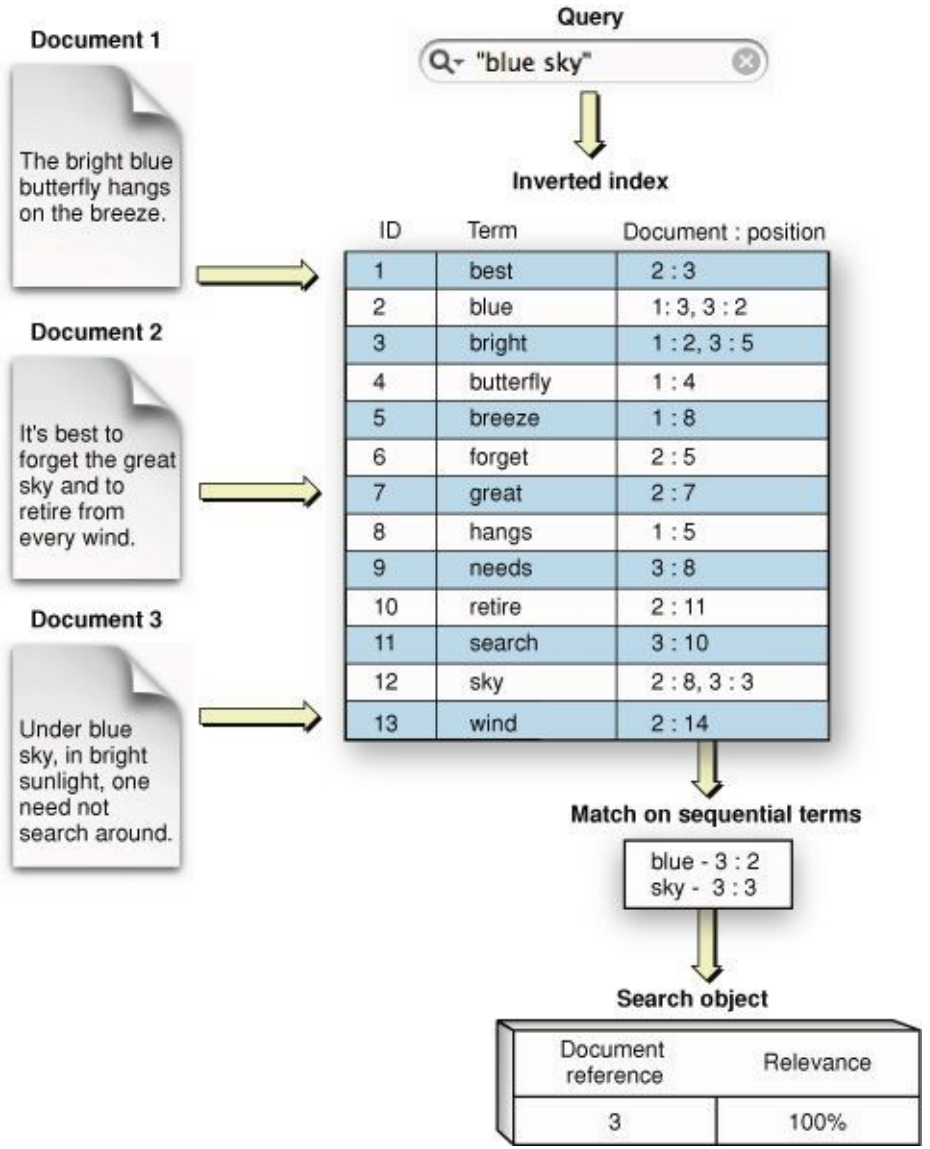
Índice invertido:



Ranking:


Query
"blue sky"
Inverted index
Document 1
The bright blue butterfly hangs on the breeze.
Document 2
It's best to forget the great sky and to retire from every wind.
Document 3
Under blue sky, in bright sunlight, one need not search around.
ID
Term
Document : position
1 best 2 : 3
2 blue 1: 3, 3 : 2
3 bright 1 : 2, 3 : 5
4 butterfly 1 : 4
5 breeze 1 : 8
6 forget 2 : 5
7 great 2 : 7
8 hangs 1 : 5
9 needs 3 : 8
10 retire 2 : 11
11 search 3 : 10
12 sky 2 : 8, 3 : 3
13 wind 2 : 14
Match on sequential terms
blue - 3 : 2
sky - 3 : 3
Search object
Document reference
Relevance
3
100%

Ranking:



https://github.com/marcelomendoza/IIC3800