

Modelo TF-IDF

Antonio Pita Lozano

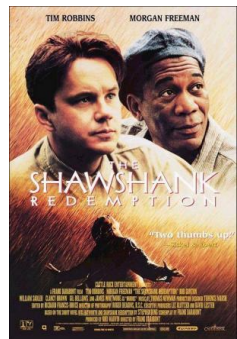
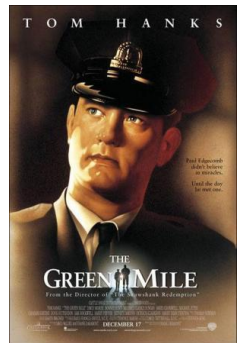
El **modelo TF-IDF** o *term frequency – inverse document frequency* es un modelo de text mining que permite asignar un valor o peso a los términos de un texto en función a su importancia en dicho texto y en el conjunto de la colección de documentos.

Este modelo es ampliamente utilizado para representar los documentos como vectores en un espacio dimensional permitiendo el uso de técnicas algebraicas sobre documentos.

$$\text{tf-idf}(\text{término}, \text{documento}) = \text{tf}(\text{término}, \text{documento}) * \text{idf}(\text{término})$$

TF (término, documento) = frecuencia del término en el documento

$$\text{IDF}(\text{término}) = \log\left(\frac{\text{número total de documentos}}{\text{\# documentos que contienen el término}}\right)$$



Visión algebraica

→ $(1,0,2,0,1, \dots, 1)$

enterprise infrastructure
technology operations
information
scorecards
analyze text mining
metrics objectives
applications manage
connection technical
solution stakeholder

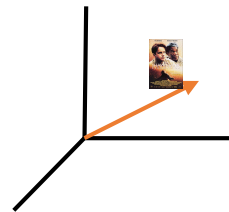
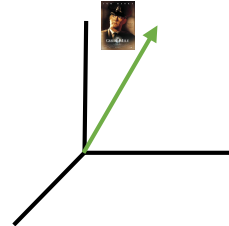
Dimensión: 745.763

→ $(2,0,0,1,1, \dots, 0)$

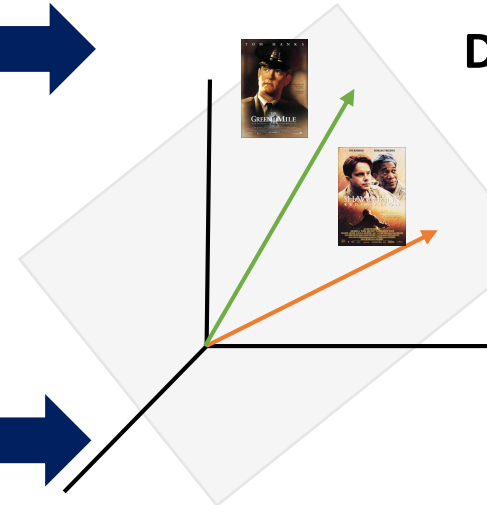
$$\text{TF-IDF}_{(n,d)} = \text{TF}_{(n,d)} \times \text{IDF}_{(n)}$$

Peso de un término (n) en un documento (d)	Frecuencia de aparición de un término (n) en un documento (d)	Factor IDF de un término (n)
--	---	------------------------------

Visión geométrica





Visión geométrica



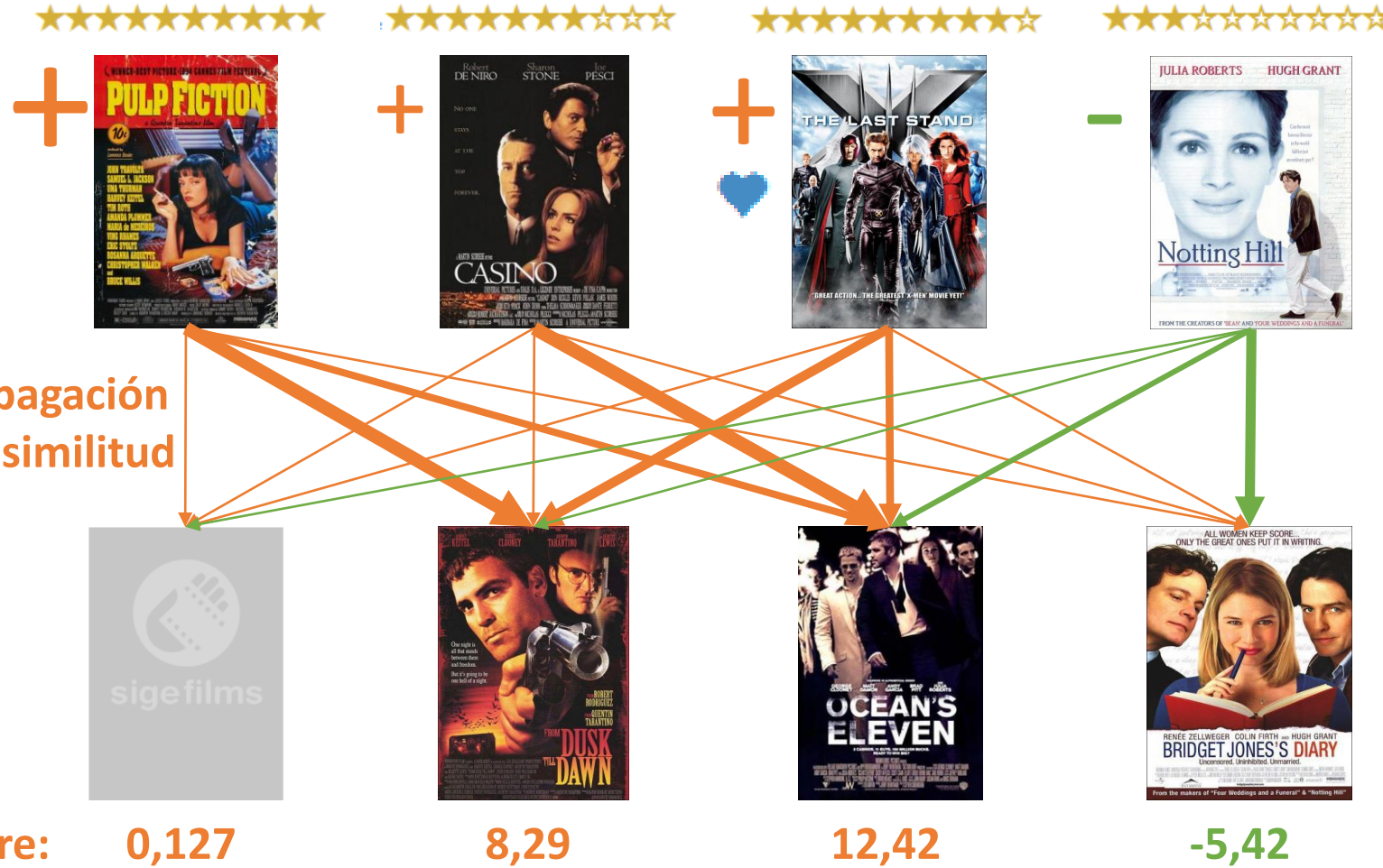
Dimensión: 2

Similitud



Sim(, ) = $\cos(\theta) = 0,373$

Recomendaciones Content-Based



Recomendaciones

- 1 Ocean's Eleven
- 2 Dusk Dawn

Modelo TF-IDF

Antonio Pita Lozano