

# Content Based recommenders for free text Item Descriptions

# Item information as unstructured text data

Movie	Description
The Dark Knight	The Dark Knight is a 2008 superhero film directed, produced, and co-written by <b>Christopher Nolan</b> . Based on the DC Comics character Batman, the film is the second installment of Nolan's <b>The Dark Knight Trilogy</b> and a sequel to 2005's Batman Begins, starring Christian Bale and supported by Michael Caine, Heath Ledger, Gary Oldman, Aaron Eckhart, Maggie Gyllenhaal, and Morgan Freeman. In the film, Bruce Wayne / Batman (Bale), Police Lieutenant James Gordon (Oldman) and District Attorney Harvey Dent (Eckhart) form an alliance to dismantle organized crime in Gotham City, but are menaced by an anarchistic mastermind known as the Joker (Ledger), who seeks to undermine Batman's influence and throw the city into anarchy.
Spiderman: Far from home	Spider-Man: Far From Home is a 2019 American superhero film based on the Marvel Comics character Spider-Man, co-produced by Columbia Pictures and Marvel Studios, and distributed by Sony Pictures Releasing. It is the sequel to Spider-Man: Homecoming (2017) and the 23rd film in the Marvel Cinematic Universe (MCU). The film was directed by Jon Watts, written by Chris McKenna and Erik Sommers, and stars Tom Holland as Peter Parker / Spider-Man, alongside Samuel L. Jackson, Zendaya, Cobie Smulders, Jon Favreau, J. B. Smoove, Jacob Batalon, Martin Starr, Marisa Tomei, and Jake Gyllenhaal. In Spider-Man: Far From Home, Parker is recruited by Nick Fury and Mysterio to face the Elementals while he is on a school trip to Europe.

# Item information as unstructured text data

	Comedy	Adventure	Superhero	Sci-Fi	
	0	1	1	0	
	1	1	1	1	
	1	0	1	0	

Movies  
Matrix

# Term-Frequency - Inverse Document Frequency (*TF – IDF*)

- Term-Frequency – Inverse Document Frequency (TF-IDF)
  - Better than bag of words and takes care of the fact that more common words are often less useful and penalizes their weight
  - TF: Measures, how often a term appears (density in a document)

$$\text{Term Frequency} = \frac{\text{Count of term } i \text{ in a document } j}{\text{Number of terms in document } j}$$

- IDF: Aims to reduce the weight of terms that are very common in the dataset

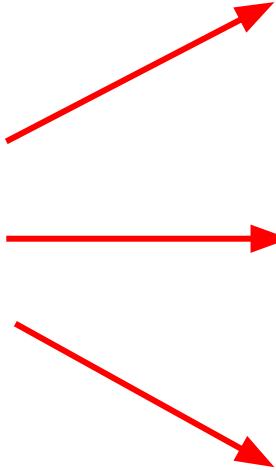
$$\text{Inverse Document Frequency} = \log \frac{\text{Count of documents in entire dataset}}{\text{Count of documents carrying item } i}$$




# Example: TF-IDF Representation



Most frequent	TF-IDF	Most frequent	TF-IDF
film	maui	film	omnidroid
moana	te	incredibles	violet
million	moana	bird	helen
disney	carvalho	movie	parr
day	polynesian	release	polynesian

# Recommending Items



	Christopher Nolan	Marvel Studios	Spiderman	Aliens	.....
	0	0	0	2.3	.....
	3.5	0	0	0	.....
	0	5.6	8.5	1.3	.....

TFIDF Matrix

# Improving the vector space model I

- Vectors are usually long and sparse
- remove stop words
  - They will appear in nearly all documents.
  - e.g. "a", "the", "on", ...
- use stemming
  - Aims to replace variants of words by their common stem
  - e.g. "went"   □   "go", "stemming"   □   "stem", ...
- size cut-offs
  - only use top n most representative words to remove "noise" from data
  - e.g. use top 100 words

# Improving the vector space model II

- Use lexical knowledge, use more elaborate methods for feature selection
  - Remove words that are not relevant in the domain
- Detection of phrases as terms (n-grams)
  - More descriptive for a text than single words
  - e.g. "United Nations"
- Limitations
  - semantic meaning remains unknown
  - example: usage of a word in a negative context
    - "there is nothing on the menu that a vegetarian would like.."
    - The word "vegetarian" will receive a higher weight than desired  
an unintended match with a user interested in vegetarian restaurants



# Limitations of Content Based Filtering

- Keywords alone may not be sufficient to judge quality/relevance of an item
  - up-to-date-ness, usability, aesthetics
  - content may also be limited / too short
  - content may not be automatically extractable (multimedia)
- Ramp-up phase required
  - Some training data is still required
- Overspecialization
  - Algorithms tend to propose "more of the same"
  - Too similar news items