Information Retrieval and Web Analytics, 2025

Universitat Pompeu Fabra Barcelona

Project Part 3: Ranking & Filtering

Team: G_004

Alex De La Haya Gutiérrez, 268169

Marc Guiu Armengol, 268920

Nil Tomàs Plans, 268384

Github URL: https://github.com/niiltomas/irwa-search-engine-G_004.git

Tag: IRWA-2025-part-3

Index of Contents

1. You're asked to provide 3 different ways of ranking:

    a. TF-IDF + cosine similarity: Classical scoring, which we have also seen during the practical labs
    b. BM25
    c. Your Score: Here, the task is to create a new score. (Be creative 🎨, think about what factors could make a document more relevant to a query and include them in your formula.)

    Explain how the ranking differs when using TF-IDF and BM25, and think about the pros and cons of using each of them. Regarding your own score, justify the choice of the score (pros and cons). HINT: Look into numerical fields that each record has to build your score.

2. Implement word2vec + cosine ranking score. Return a top-20 list of documents for each of the 5 queries defined in the Part 2 of your project, using search and word2vec + cosine similarity ranking.
    To represent a piece of text using word2vec, we create a single vector that represents the entire text. This vector has the same number of dimensions as the word vectors and is calculated by averaging the vectors of all words in the text.

3. Can you imagine a better representation than word2vec? Justify your answer. (HINT - what about Doc2vec? Sentence2vec? What are the pros and cons?)

Universitat Pompeu Fabra Barcelona

1. Ranking

a. TF-IDF + cosine similarity

For the first ranking strategy, we used the classic TF-IDF formula to build document vectors and cosine similarity to compare them with the query.
The resulting similarity scores are used to rank the documents. To improve diversity in the results, we implemented a functionality to avoid repetitions of product titles, so the top-k results include a wider variety of products. We tested this model using the five queries defined in Part 2 of the project:
Q1: "women track pant"
Q2: "men track pant"
Q3: "men pack"
Q4: "women formal shirt"
Q5: "men slim fit formal shirt"

b. BM25

The second ranking strategy is based on the BM25 model. We also applied the anti-duplication of titles.
Note that in this model the rank values are not normalized because BM25 is an additive weighted sum over query terms, rather than a measure bounded between 0 and 1 as before.

c. Your score: Custom Hybrid Ranking: TF-IDF + Business Factors

For our personalized hybrid ranking, we combined the TF-IDF cosine similarity score with additional product features to better reflect user preferences. The final score is a weighted sum:

$$Score = \alpha \cdot (TF - IDF + cosine\ sim.\ score) + \beta \cdot rating + \gamma \cdot discount + \\ + \delta \cdot (1 - price) + penalty_{stock}$$

Where the weights were configured as: α = 0.7, β = 0.15, γ = 0.10, δ = 0.05. The stock penalty is **-0.2** for products that are out of stock.
Pros: Our model balances textual relevance with business factors, by pushing highly rated, discounted, affordable, and available products higher in the ranking. Also, it is flexible, and the weights can be adjusted according to different priorities.
Cons: The weights are heuristic, and the model does not handle context or synonyms, beyond the TF-IDF cosine similarity.

upf. Universitat Pompeu Fabra Barcelona

2. word2vec + cosine

We have trained word2vec on the whole corpus to obtain embeddings for all words. Then we represent documents and queries by averaging their word embeddings. Finally we compute cosine similarity.

Query	Top-20 list of documents
women track pant	0.958418 — Solid Women Multicolor Track Pants 0.956445 — Solid Men Multicolor Track Pants 0.944847 — Solid Women Black Track Pants 0.943185 — Solid Women Olive Track Pants 0.939513 — Camouflage Women Blue Track Pants 0.935649 — Applique Men Black Track Pants 0.929412 — Solid Women Grey Track Pants 0.926964 — Solid Women White Track Pants 0.92661 — Solid Men Grey Track Pants 0.924791 — Solid Men White Track Pants 0.924061 — Solid Men Black Track Pants 0.92233 — Striped Women Grey Track Pants 0.921872 — Solid Women Blue Track Pants 0.921373 — Printed Women Grey Track Pants 0.921184 — Solid Women Brown Track Pants 0.92033 — Printed Men Grey Track Pants 0.919112 — Striped Women Black Track Pants 0.918927 — Printed Women Black Track Pants 0.918743 — Checkered Women Olive Track Pants 0.918722 — Striped Men Grey Track Pants
men track pant	0.958141 — Solid Men Multicolor Track Pants 0.956712 — Solid Women Multicolor Track Pants 0.946009 — Solid Women Black Track Pants 0.944678 — Solid Women Olive Track Pants 0.941926 — Camouflage Women Blue Track Pants 0.939051 — Applique Men Black Track Pants 0.929505 — Solid Men Grey Track Pants 0.929407 — Solid Women Grey Track Pants 0.927543 — Solid Men Black Track Pants 0.926626 — Solid Men White Track Pants 0.925738 — Solid Women White Track Pants 0.923493 — Striped Women Grey Track Pants 0.922789 — Striped Men Grey Track Pants 0.9226 — Solid Women Blue Track Pants 0.922315 — Solid Women Brown Track Pants 0.922174 — Solid Men Blue Track Pants 0.921639 — Solid Men Brown Track Pants 0.92131 — Checkered Women Olive Track Pants 0.920824 — Striped Women Black Track Pants 0.920498 — Printed Men Grey Track Pants
men pack	0.927697 — Women Mid-Calf/Crew (Pack of 2) 0.926534 — Women, Women Printed Bandana (Pack of 6) 0.913393 — Women Solid Mid-Calf/Crew (Pack of 10) 0.911856 — Men Mid-Calf/Crew (Pack of 3) 0.911217 — Men Solid Mid-Calf/Crew (Pack of 3) 0.911115 — Women Solid Mid-Calf/Crew (Pack of 3)

	0.909179 — Women Solid Mid-Calf/Crew (Pack of 9) 0.907131 — Men Solid Mid-Calf/Crew (Pack of 12) 0.906976 — Women Solid Mid-Calf/Crew (Pack of 12) 0.904024 — Men Solid Mid-Calf/Crew (Pack of 6) 0.903306 — Women Striped Mid-Calf/Crew (Pack of 3) 0.903078 — Women Solid Mid-Calf/Crew (Pack of 6) 0.902685 — Women Peds/Footie/No-Show (Pack of 3) 0.900262 — Women Solid Mid-Calf/Crew (Pack of 5) 0.899166 — Men Solid Mid-Calf/Crew (Pack of 4) 0.898757 — Women Printed Mid-Calf/Crew (Pack of 6) 0.895456 — Men Geometric Print Mid-Calf/Crew (Pack of 2) 0.894968 — Women Mid-Calf/Crew (Pack of 5) 0.894223 — Men Geometric Print Mid-Calf/Crew (Pack of 3) 0.894131 — Women Mid-Calf/Crew (Pack of 6)
women formal shirt	0.956725 — Women Solid Formal Shirt 0.900514 — Women Slim Fit Solid Formal Shirt 0.899778 — Men Slim Fit Solid Formal Shirt 0.895102 — Women Regular Fit Solid Formal Shirt 0.894169 — Men Regular Fit Solid Formal Shirt 0.89122 — Men Slim Fit Printed Formal Shirt 0.89063 — Women Slim Fit Printed Formal Shirt 0.890587 — Women Regular Fit Checkered Formal Shirt 0.889398 — Men Regular Fit Checkered Formal Shirt 0.888496 — Women Slim Fit Checkered Formal Shirt 0.887537 — Men Slim Fit Checkered Formal Shirt 0.885492 — Women Regular Fit Striped Formal Shirt 0.88402 — Men Regular Fit Striped Formal Shirt 0.877506 — Men Tailored Fit Checkered Spread Collar Formal Shirt 0.877341 — Men Slim Fit Solid Spread Collar Casual Shirt 0.876805 — Women Slim Fit Solid Spread Collar Casual Shirt 0.873628 — Men Slim Fit Solid Slim Collar Casual Shirt 0.872974 — Women Slim Fit Solid Slim Collar Casual Shirt 0.872901 — Women Regular Fit Printed Formal Shirt 0.870816 — Men Regular Fit Solid Button Down Collar Formal Shirt
men slim fit formal shirt	0.980363 — Men Slim Fit Solid Formal Shirt 0.979388 — Women Slim Fit Solid Formal Shirt 0.973654 — Men Slim Fit Printed Formal Shirt 0.971379 — Men Slim Fit Checkered Formal Shirt 0.971231 — Women Slim Fit Printed Formal Shirt 0.970786 — Women Slim Fit Checkered Formal Shirt 0.952054 — Men Regular Fit Checkered Formal Shirt 0.951605 — Women Regular Fit Checkered Formal Shirt 0.951304 — Men Regular Fit Solid Formal Shirt 0.951069 — Men Slim Fit Solid Slim Collar Casual Shirt

	0.950442 — Women Regular Fit Solid Formal Shirt 0.949984 — Men Regular Fit Striped Formal Shirt 0.949744 — Women Regular Fit Striped Formal Shirt 0.949188 — Women Slim Fit Solid Slim Collar Casual Shirt 0.942998 — Men Slim Fit Solid Spread Collar Casual Shirt 0.94293 — Women Slim Fit Printed, Checkered Formal Shirt 0.941239 — Women Slim Fit Solid Spread Collar Casual Shirt 0.939426 — Men Slim Fit Solid Spread Collar Formal Shirt 0.938406 — Women Slim Fit Solid Spread Collar Formal Shirt 0.936589 — Men Slim Fit Printed Spread Collar Formal Shirt

3. Part 3.3 — Better Representation Beyond Word2Vec

Word2Vec gives an effective way to capture word-level semantics. Limitations are the loss of context (averaging words ignores word order and structure), resulting in not understanding the sentence's meaning.

Doc2Vec is an alternative. It creates a unique vector for each document. It learns document-level context. On the other hand it requires retraining for each new corpus.

Sentence2Vec is another alternative. Produces embeddings that represent the meaning at the sentence or paragraph level. It handles paraphrases, synonyms and context well. In contrast it is heavier computationally and needs GPU for fast inference.

Nowadays there are much better alternatives such as Contextual Embeddings (BERT, **GPT embeddings**). The vector for a word changes based on its meaning in the sentence. Although it is harder to train and requires larger model sizes, it performs better overall because it handles polysemy, encodes word order and grammar and can be fine tuned.