Exercises

2.2. *A more-like-this query occurs when the user can click on a particular document in the result list and tell the search engine to find documents that are similarto this one. Describe which low-level components are used to answer this type of query and the sequence in which they are used.*

2.3. *Document filtering is an application that stores a large number of queries or user profiles and compares these profiles to every incoming document on a feed. Documents that are sufficiently similar to the profile are forwarded to that person via email or some other mechanism. Describe the architecture of a filtering engine and how it may differ from a search engine.*

2.2 *More like this* query

According to the search engine ElasticSearch, the required steps to accomplish a query like this one, the following steps are followed:

1. Selects a set of representative terms of these input documents
2. Forms a query with these terms
3. Executes the query
4. Return the results

In order to accomplish these steps, several low-level components are used. The representative terms from the input documents are gotten evaluating the tf-idf and selecting the ones with the highest tf-idf; these formula is used in the **weighting** component from the index creation building-block, then most likely is applied an **inversion** to the obtained terms to create the inverted index for then to be able to perform a disjunctive query of the terms.

Once the query is determined, the **query input** component is used to start the *query process* and proceed to use the **scoring** and, potentially, the **performance optimization** of the *ranking* building block to finally display the results to the user through the **results output** component.

Optionally, certain components of the *evaluation* building block may be used to improve future MLT queries.

2.3

A filtering engine differs from a search engine in the fact that whereas the latter tries to obtain coincidences of any by analyzing the content of the documents in its collection and matching them with an user query to produce a subset of documents that may be relevant to the user's query, a filtering engine does the opposite, it searchs within the users' profiles for coincidences that may make the system assume whether the current document is relevant or not for that specific user; ergo, instead of using the users' input to determine the relevance of a document, now the content of the users' profile will be used to determine the relevance of that document.