Valentin Gegoux

IR System efficiency report

Introduction

This report aims to measure the efficiency of a boolean Information Retrieval model based on pre-selected documents relevant to corresponding queries. The measurement is split in 2 different categories of query:

1. Search strictly based on the **words** contained in the query: the user wants a pure logical search of songs that contain the specified words.
2. Search based on **meaning**: the user wants to retrieve the songs that match with a theme he has in mind.

This last part will demonstrate the limits of the boolean model and why it's not ideal for naïve user queries.

1. Precision & Recall measurements based on **words**:

Query	Relevant documents	Documents retrieved	Precision	Recall
city	5	4	4/4 = 1	4/5 = 0.8
life	16	16	16/16 = 1	16/16 = 1
can't stop	2	2	2/2 = 1	2/2 = 1
night	20	19	19/19 = 1	19/20 = 0.95
opera or rock	17	17	17/17 = 1	17/17 = 1
again and me	8	8	8/8 = 1	8/8 = 1
inside	9	9	9/9 = 1	9/9 = 1
dream	5	5	5/5 = 1	5/5 = 1
thing and be	4	4	4/4 = 4	4/4 = 1
hey	11	11	11/11 = 1	11/11 = 1
sing or singing	3	3	3/3 = 1	3/3 = 1

Basing relevance strictly on words, the IR system achieves perfect precision. Indeed, irrelevant documents cannot be retrieved since results are strictly limited to the query content.

Thanks to the sanitization of words and queries, recall is also nearly perfect. In song lyrics, words are often surrounded by characters such as commas, parentheses, apostrophes, exclamation marks, or quotation marks. Sometimes, the letter case varies as well. Without sanitization, many songs would be ignored, and recall would be significantly lower.

Nevertheless, the same word can be written in different ways. For example, verbs can be conjugated, or nouns can appear in singular or plural form. It would be difficult to sanitize this kind of variation without losing accuracy.

2. Precision & Recall measurements based on **meaning**:

Meaning of the request	**Relevant documents**	**Query**	**Documents retrieved**	**Precision**	**Recall**
Songs about love	25	love	36	18/36 = 0.5	18/25 = 0.72
		love or romance	37	18/37 = 0.49	18/25 = 0.72
		love and heart	12	9/12 = 0.75	9/25 = 0.36
		romance or heart	21	13/21 = 0.62	13/25 = 0.52
		you and me	46	22/46 = 0.48	22/25 = 0.88
		kiss	7	4/7 = 0.57	4/25 = 0.16
		love or romance or kiss or heart or baby	50	23/50 = 0.46	23/25 = 0.92
		darling or baby	25	14/25 = 0.56	14/25 = 0.56
Songs about heartache / breakup	7	heart and break	4	1/4 = 0.25	1/7 = 0.14
		down	23	4/23 = 0.17	4/7 = 0.57
		goodbye or cry or tears	8	4/8 = 0.5	4/7 = 0.57

Songs to dance	13	dance	7	6/7 = 0.86	6/13 = 0.46
		everybody or together	6	4/6 = 0.67	4/13 = 0.31
		move or body	13	7/13 = 0.54	7/13 = 0.54
		tonight	23	10/23 = 0.43	10/13 = 0.77
		night	19	8/19 = 0.42	8/13 = 0.61
		show	10	4/10 = 0.4	4/13 = 0.31
		party	3	1/3 = 0.33	1/13 = 0.08

As you can see on this table, results can vary a lot depending on the query. The more we add words related, the higher the recall. Achieving high precision is difficult because even simple queries (e.g., “love”) retrieve many irrelevant documents.

More generally, we often observe low values for both precision and recall, meaning that many relevant documents are not retrieved, while many irrelevant ones are.

We can conclude from these measurements that the boolean model is not well suited for searches based on a concept. Indeed, there are too many words that can illustrate a single meaning. A preferable option would likely be to use another type of system such as a vectorial space model.