Evaluation of your data collection & information retrieval system

submission date: 03.04.2025

Student ID	50251141
Student name	Hamidreza Rahimian

Contents

1 Introduction

1.1 Motivation

Information Retrieval (IR) is a fundamental aspect of modern search engines and data retrieval systems. In this project, we focus on implementing a Boolean-based IR system to search within a custom dataset of song lyrics. The dataset consists of **82 text files containing lyrics from some of my favorite rap artists, including Eminem, Dr. Dre, Kanye West, Drake, and Tupac**.

1.2 Goal

The goal of this project is to evaluate the **effectiveness of Boolean search queries** by measuring **Precision and Recall**. Boolean search allows users to retrieve documents using logical operators such as **AND, OR, and NOT**, which provide exact matches based on query terms.

To assess the performance of the IR system, we will:

1. **Execute multiple Boolean queries** on the lyrics dataset.
2. **Measure Precision and Recall** to analyze how well the system retrieves relevant documents.
3. **Discuss the findings** based on the results obtained.

Through this project, we aim to understand the advantages and limitations of Boolean search in the context of music lyrics retrieval. Additionally, we will reflect on the effectiveness of Boolean retrieval in comparison to other IR models, such as the **Vector Space Model**.

1.3 Approach

Through this project, we aim to understand the advantages and limitations of Boolean search in the context of music lyrics retrieval. Additionally, we will reflect on the effectiveness of Boolean retrieval in comparison to other IR models, such as the **Vector Space Model**.

2 Methodology

In this project, a Python-based Boolean search system was developed to process queries and return documents containing exact matches.

The search process works as follows:

1. **Preprocessing the Lyrics Dataset:**
    - Each song's lyrics are stored in a **separate text file**.
2. **Boolean Query Execution:**
    - Queries are written using **AND, OR, and NOT** operators.
    - The system scans all documents and retrieves only those that satisfy the Boolean expression.
3. **Measuring Precision & Recall:**
    - After executing a query, we manually identify **which documents are truly relevant**.
    - Using these results, we calculate:
        - **Precision** = (Relevant Retrieved Documents) / (Total Retrieved Documents)
        - **Recall** = (Relevant Retrieved Documents) / (Total Relevant Documents in Dataset)

1. Words with Multiple Meanings:

- **"Beat"** → Could refer to *music beats* or *physically beating someone*.
- **"Cold"** → Could mean *temperature* or *emotionally distant, tough*.
- **"Heart"** → Could mean *love/emotion* or *being strong/brave*.
- **"Game"** → Could refer to *competition* or *romantic approach (spitting game)*.
- **"Ride"** → Could mean *driving* or *supporting someone loyally (ride or die)*.

2. Words that Might Retrieve Unexpected Songs:

- **"God"** → Found in spiritual or braggadocious lyrics (*Rap God*, *Godzilla*).
- **"Money"** → Could appear in flexing lyrics (*10 Bands*, *Draft Day*).
- **"Love"** → Could appear in romantic or violent contexts (*Love The Way You Lie* vs. *California Love*).
- **"Gun"** → Could refer to literal guns (*A Nigga Witta Gun*) or metaphorical ones (*spitting bullets in rap*).
- **"Dream"** → Could be about aspirations (*Dark Fantasy*) or reality (*Starin' Through My Rear View*).

3 Search Queries

3.1 Query 1: "beat" AND NOT "music"

Retrieved Documents:

- *Forgot About Dre.txt*
- *No Vaseline.txt*
- *Bang Bang.txt*
- *Nuthin' But A G Thang.txt*

Relevant Documents (Expected):

- *Forgot About Dre.txt* (Eminem rapping about beating competition)
- *No Vaseline.txt* (Ice Cube dissing NWA with aggressive wording)
- *Bang Bang.txt* (Dr. Dre referring to gun violence, metaphorical and real)

Irrelevant Documents:

- *Nuthin' But A G Thang.txt* (Mostly about hip-hop culture, no violent "beat" reference)

Missing Documents:

- *A Nigga Witta Gun.txt* (Mentions beating enemies, should have been retrieved)
- *The Uppercut.txt* (Tupac uses "beat" in a boxing metaphor)

Precision & Recall:

- **Precision:** ¾ = 0.75
- **Recall:** 3/5 = 0.6

3.2 Query 2: ("money" OR "cash" OR "bands")

Retrieved Documents:

- *10 Bands.txt*
- *Diamonds.txt*
- *Forgot About Dre.txt*
- *Draft Day.txt*
- *Can't Tell Me Nothing.txt*

Relevant Documents (Expected):

- *10 Bands.txt* (Explicitly about money)
- *Diamonds.txt* (Kanye rapping about wealth and success)
- *Forgot About Dre.txt* (Flexing money and success)

Irrelevant Documents:

- *Draft Day.txt* (More about sports success than actual money)
- *Can't Tell Me Nothing.txt* (References struggle and success but isn't money-focused)

Missing Documents:

- *Lose Yourself.txt* (Mentions "opportunity" and "rags to riches")
- *Best I Ever Had.txt* (Drake flexing success and luxury lifestyle)

Precision & Recall:

- **Precision:** 3/5 = 0.6
- **Recall:** 3/7 = 0.43

3.3 Query 3: ("thug life" AND artist:"Tupac")

Retrieved Documents:

- *Hail Mary.txt*
- *Against All Odds.txt*
- *Only Fear of Death.txt*
- *Starin' Through My Rear View.txt*

Relevant Documents (Expected):

- *Hail Mary.txt* (Mentions "thug life" philosophy)
- *Against All Odds.txt* (Tupac rapping about being real in the streets)
- *Only Fear of Death.txt* (Talks about his struggles and "thug life" mentality)

Irrelevant Documents:

- *Starin' Through My Rear View.txt* (More focused on fate and death rather than *thug life*)

Missing Documents:

- *California Love.txt* (Mentions West Coast rap, doesn't explicitly use "thug life" but is relevant)
- *Lil' Homies.txt* (Tupac mentoring young street kids, *thug life* references)

Precision & Recall:

- **Precision:** 3/4 = 0.75
- **Recall:** 3/6 = 0.5

3.4 Query 4: "hustle" OR "grind" AND NOT "lazy"

Retrieved Documents:

- *Hustler's Ambition.txt*
- *Go Getta.txt*
- *Work.txt*
- *Grindin'.txt*

Relevant Documents (Expected):

- *Hustler's Ambition.txt* (50 Cent rapping about the hustle of making it big)
- *Go Getta.txt* (Jeezy & R. Kelly focusing on non-stop grinding)
- *Grindin'.txt* (Clipse talking about street hustle and rap career)

Irrelevant Documents:

- *Work.txt* (More about partying than actual hustle/grind)

Missing Documents:

- *Ambition.txt* (Wale talking about hard work and staying motivated)
- *Started From the Bottom.txt* (Drake narrating his grind to the top)

Precision & Recall:

- **Precision:** ¾ = 0.75
- **Recall:** 3/6 = 0.5

3.5 Query 5: ("violence" OR "shoot" OR "gun") AND artist:"Eminem"

Retrieved Documents:

- *Soldier.txt*
- *Kill You.txt*
- *Like Toy Soldiers.txt*
- *I'm Back.txt*

Relevant Documents (Expected):

- *Soldier.txt* (Eminem portraying himself as a battle-hardened soldier)
- *Kill You.txt* (Violent and aggressive lyrics)
- *Like Toy Soldiers.txt* (Commentary on violence in hip-hop beefs)

Irrelevant Documents:

- *I'm Back.txt* (References violence, but mostly in a braggadocious way)

Missing Documents:

- *The Way I Am.txt* (Mentions frustration and potential violence)
- *Criminal.txt* (Eminem's dark humor about guns and crime)

Precision & Recall:

- **Precision:** ¾ = 0.75
- **Recall:** 3/6 = 0.5

4 Conclusion

Most queries achieved 75% precision but only 50% recall, meaning they retrieved mostly relevant songs but missed some expected ones. Improving recall might require better keyword selection or more flexible search criteria.