Tổng quan về cách tính độ tương đồng(similarity) giữa câu truy vấn và tập dữ liệu trong hệ thống EasyFunc

# Tổng quan

Khi thực hiện tìm kiếm có thể có rất nhiều kết quả trả về tương ứng với câu truy vấn, các kết quả này có mức độ tương đồng khác nhau với câu truy vấn. Do đó, ta cần phải sắp xếp chúng lại theo một giải thuật tương ứng nhằm mang lại những trải nghiệm tốt nhất trong việc tìm kiếm.

Dựa vào mô hình tìm kiếm, có rất nhiều giải thuật tính toán độ tương đồng như TF-IDF, BM25,…Trong hệ thống EasyFunc, chúng ta sẽ hiện thực các giải thuật khác nhau dựa trên mô hình Vector Space Model

# Hiện thực

## Tài liệu(Document)

Tài liệu là thành phần nhỏ nhất trong một hệ thống tìm kiếm. Nó được dùng để lưu trữ thông tin cũng như rút trích thông tin. Việc phân chia thông tin vào trong các tài liệu của hệ thống gặp một số vấn đề sau:

* Một số giải thuật như TF-IDF đánh giá sự tương đồng dựa trên tần suất xuất hiện của một term(Term là đơn vị nhỏ nhất của một câu truy vấn) , vì vậy mỗi tài liệu phải là một file miêu tả của Type(Type: đơn vị của Java bao gồm: Class, Interface và Enum) bao gồm nhiều miêu tả của Method
* Một số giải thuật khác như BNS,Chi Square… lại đánh giá sự tương đồng dựa trên tỷ lệ giữa kết quả đúng và sai(hit and miss) giữa câu truy vấn và tài liệu. Vì vậy, mỗi tài liệu phải là một miêu tả của một Method

Ví dụ: Có đoạn miêu tả Type như sau:

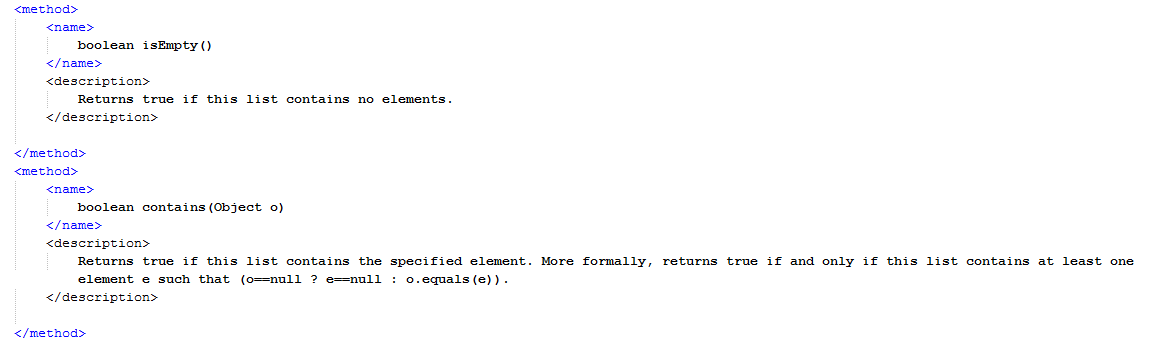


Figure : Document1

Với phương pháp TF-IDF ta cần phải tạo một bảng thống kê tần suất xuất hiện của các term trong tài liệu ví dụ như: *boolean:2; returns:3, element(s):3….* Bằng việc kết hợp tất cả các bảng thống kê của mỗi tài liệu trên toàn bộ cơ sở dữ liệu, chúng ta tính được độ tương đồng

Với phương pháp BNS chúng ta cũng tạo một bảng thống kê dựa trên danh sách các term tuy nhiên chỉ quan tâm tới việc term đấy xuất hiện trong tài liệu hay không. Ví dụ như: *boolean:true, return: true, string: no…* Bằng việc đếm số lần xuất hiện của term trên toàn bộ cơ sở dữ liệu, chúng ta tính toán được độ tương đồng