**Fuzzy String Matching**

1. Giới thiệu về Fuzzy string matching

Fuzzy String Matching là một thư viện cung cấp các kỹ thuật tìm kiếm và so khớp chuỗi với một phần chuỗi đã cho trước hoặc không chính xác. Khi người dùng nhập sai một chuỗi hoặc viết một phần của chuỗi, Fuzzy String Matching giúp so khớp và tìm đúng chuỗi đó.

Thư viện này không chỉ đơn giản so khớp sự tương đồng của hai chuỗi mà còn xác định mức độ gần nhau của hai chuỗi. Nó sử dụng “khoảng cách Levenshtein” để xác định mức độ gần nhau giữa hai chuỗi bằng cách tính toán khoảng cách tối thiểu cần thiết để chuyển đổi chuỗi này thành một chuỗi khác.

2. Khoảng cách Levenshtein

Khoảng cách Levenshtein là một số liệu để đo khoảng cách giữa hai chuỗi. Nó là số lần chỉnh sửa tối thiểu để chuyển đổi chuỗi này thành một chuỗi khác (chèn, xóa hoặc thay thế).

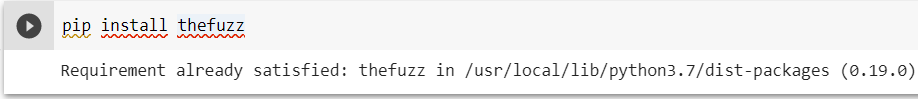
Khoảng cách giữa hai chuỗi a,b có độ dài |a|, |b| tương ứng được xác định bởi:

Trong đó tail(x) là tất cả của một chuỗi x ngoại trừ ký tự đầu tiên của nó.

Ví dụ khoảng cách Levenshtein giữa “hanoi” và “hanam” là 2.

3. Fuzzy String Matching trong python

Để cài đặt thư viện Fuzzy String Matching, chạy lệnh như sau.



a. Tính khoảng cách Levenshtein

Thư viện này có một hàm tính tỉ lệ tương đương với khoảng cách Levenshtein giữa hai chuỗi.

Graphical user interface, text, application

Description automatically generated

Đầu ra cho kết quả là 50 tương đướng với tỉ lệ khoảng cách Levenshtein chia cho độ dài chuỗi lớn nhất trong hai chuỗi.

b. Tính tỉ lệ từng phần

Tỷ lệ từng phần giúp thực hiện đối sánh chuỗi con. Thao tác này lấy chuỗi ngắn nhất và so sánh với tất cả chuỗi con có cùng độ dài của chuỗi còn lại.

Graphical user interface, application

Description automatically generated

Đầu ra cho kết quả 93 là sự đối sánh giữa chuỗi thứ nhất và chuỗi “My name is Ali “là chuỗi con của chuỗi thứ hai.

Cách tính tỉ lệ này có thể rất hữu ích, ví dụ ta cần đối sánh tên của một người giữa hai tập dữ liệu. Trong dữ liệu đầu tiên có tên và họ của người đó, tập dữ liệu thứ hai có tên, tên đệm, họ của người đó. Đầu ra sẽ cho kết quả là 100 do chuỗi đầu tiên sẽ là chuỗi con trong chuỗi thứ hai.

c. Tính tỉ lệ sắp xếp chữ cái

Trong tỉ lệ sắp xếp, các chuỗi được mã hóa và xử lý trước bằng cách chuyển đổi thành chữ thường và loại bỏ dấu chấm câu. Các chuỗi sau đó được sắp xếp theo bảng chữ cái và được nối lại với nhau.

Graphical user interface

Description automatically generated

Đầu ra là 100 do sau khi được xắp xếp lại theo bảng chữ cái thì thứ tự ban đàu của các từ không quan trọng.

d. Tìm chuỗi phù hợp nhất trong một tập hợp chuỗi

Nếu có một danh sách các chuỗi và ta muốn tìm một chuỗi phù hợp nhất từ danh sách với chuỗi đã cho, ta có thể sử dụng “process”:

Text

Description automatically generated

Nếu muốn chỉ tìm chuỗi phù hợp nhất, sử dụng câu lệnh như sau:

Graphical user interface, text

Description automatically generated