**Polyfuzzy**

PolyFuzz thực hiện so sánh chuỗi gần đúng (fuzzy string matching), nhóm chuỗi, và chứa các chức năng đánh giá mở rộng. PolyFuzz có nghĩa là mang nhiều kỹ thuật đối sánh chuỗi gần đúng cùng với nhau trong 1 framework.

Hiện tại, phương pháp bao gồm 1 loạt độ đo khoảng cách, tần suất xuất hiện của kí tự TF-IDF, kĩ thuật word embedding như FastText và Glove, và transformer embedding.

**Cài đặt:**

-Cài đặt qua pip:

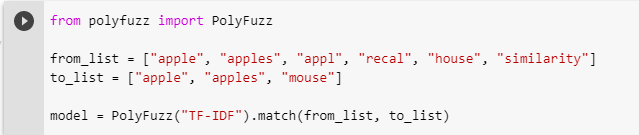


**Demo:**

Khởi tạo 2 chuỗi from\_list và to\_list.

Sử dụng TF-IDF để tạo tần suất xuất hiện kí tự để so sánh độ tương đồng giữa 2 xâu.

Sau đó tính độ tương đồng giữa 2 xâu bằng việc tính độ tương đồng cos giữa các véc tơ biểu diễn.



Kết quả đối sánh có thể truy cập thông qua hàm get\_matches()

Ảnh có chứa bàn

Mô tả được tạo tự động

**Fit/Transfrom**

Hàm match cho phép bạn nhanh chóng trích xuất các chuỗi tương tự. Tuy nhiên, sau khi chọn mô hình phù hợp để sử dụng, bạn muốn sử dụng PolyFuzz để đối sánh với những chuỗi mới. Để làm vậy, chúng ta có thể sử dụng 1 số hàm tương tự như fit, transform và fit\_transform.

Chúng ta có một danh sách các từ mà chúng ta biết là chính xác được gọi là train\_words. Chúng ta muốn bất cứ từ mới nào đến để ánh xạ đến 1 trong các từ trong train\_words. Nói cách khác, chúng ta sử dụng fit trên tập train\_words và sử dụng transform trên bất kỳ từ mới nào.

Ảnh có chứa văn bản

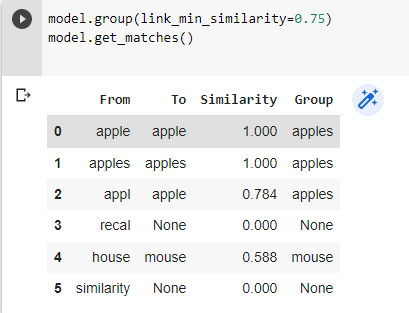
Mô tả được tạo tự động

Trong ví dụ trên, chúng ta sử dụng hàm fit trên tập train\_words để tính biểu diễn TF-IDF của những từ đó để lưu lại và sử dụng lại trong hàm transform. Việc này giúp tăng tốc hàm transform một chút vì tất cả biểu diễn TF-IDF đã được lưu lại khi sử dụng hàm fit.

**Group Matches**

Chúng ta có thể nhóm các chuỗi trong cột **To** vì có thể gây ra sự chồng chéo trong các chuỗi trong to\_list của chúng ta. Để làm vậy, chúng ta tính độ tương đồng giữa các chuỗi trong to\_list và sử dụng liên kết đơn để nhóm các chuỗi với độ tương đồng cao

Khi chúng ta trích xuất kết quả đối sánh mới, chúng ta có thể thấy cột bổ sung Group:



**Một số mô hình được thực hiện trong PolyFuzz:**

1. TF-IDF
2. Edit Distance with RapidFuzz
3. FastText and Glove
4. Transformers
5. SentenceTransformers
6. Gensim 7. Spacy