Preprocessing

# Regex là gì?

* Regular Expression: biểu thức chính quy.
* \r\n: xuống dòng và quay về đầu dòng mới.

# Nltk

* Thư viện Natural Language Toolkit
* Import stemmer từ module nltk.stem.porter: có chức năng chuyển các từ về dạng gốc bằng cách loại bỏ các phần phụ định và hậu tố🡪 giúp giảm số lượng biến thể của một từ trong văn bản🡪 giúp việc xử lý và phân loại dễ dàng hơn.

## TfidfVectorizer

* Chuyển đổi một tập hợp các văn bản thành một ma trận TF-IDF(term frequency-inverse document frequency)🡪 dùng để đánh giá tầm quan trọng của một từ trong một văn bản.
* Công thức tính:
* TF: tần số của từ trong văn bản
* IDF(t) = log\_e(Tổng số văn bản / Số văn bản chứa từ t ) 🡪đại diện cho tính quan trọng của từ trong tập hợp các văn bản, IDF của từ giảm khi từ đó xuất hiện nhiều lần trong văn bản. 🡪 giảm độ quan trọng của các từ xuất hiện trong nhiều văn bản như “a”, “the”, … và tăng độ quan trọng của các từ đặc biệt có thể giúp phân biệt giữa các văn bản.
* TF-IDF là tích của TF và IDF. TF-IDF tăng khi một từ xuất hiện nhiều lần trong một văn bản cụ thể, nhưng giảm nếu từ đó xuất hiện trong nhiều văn bản khác nhau.Do đó, các từ có giá trị TF-IDF cao trong một văn bản thường là các từ đặc biệt và quan trọng trong văn bản đó, giúp phân biệt giữa các văn bản trong tập dữ liệu.

## Cosine\_similarity

* Tính toán độ tương đồ cosin giữa hai vector🡪 dùng đề so sánh sự tương đồng giữa các văn bản.
* Trong bối cảnh xử lý văn bản, vector biểu diễn văn bản thường có hàng trăm hoặc hàng nghin chiều, mỗi chiều đại diện cho tần số xuất hiện của một từ trong từ điển. Do đó, các văn bản có nội dung tương tự sẽ có các vector biểu diễn gần nhau trong không gian nhiều chiều.

Cosine-similarity(a,b) =