2.2 Bài toán phân cụm nội dung tin nhắn giống nhau với bộ dữ liệu tweets

2.2.1 Mô tả bài toán

Twitter cung cấp dịch vụ đăng tin nhắn ngắn. Trong thực tế, nhiều tweet rất giống nhau và có thể gom lại được với nhau. Bằng cách nhóm các tweet tương tự lại với nhau, chúng ta có thể tạo ra một bảng trình bày ngắn gọn và có tổ chức hơn cho các tweet thô, điều này sẽ rất hữu ích cho nhiều ứng dụng dựa trên twitter (ví dụ: khám phá sự thật, phân tích xu hướng, xếp hạng tìm kiếm.

Ở đây, các tweet được phân nhóm bằng cách sử dụng thước đo khoảng cách Jaccard và thuật toán K – Mean

Khoảng cách Jaccard đo lường sự không giống nhau giữa hai bộ mẫu (A và B). Nó được định nghĩa là hiệu số của kích thước của hợp và giao của hai tập hợp chia cho kích thước của hợp của các tập hợp.

Lưu ý bạn có thể sử dụng các giải thuật khác để tìm điểm tương đồng giữa nội dung văn bản, Jaccard chỉ có thể phân biệt được 70% nội dung đưa vào.

Dist(A,B) = 1 - |A giao B| / |A union B|

Ví dụ: Hãy xem xét các tweet sau:

Tweet A: Cuộc hành trình dài.

Tweet B: Cuộc sống dài.

|A giao B| = 1 và |A union B| = 5, do đó khoảng cách là 1 – (1/5)

Khoảng cách Jaccard Dist (A,B) giữa tweet A và B có các thuộc tính sau:

1. Thật nhỏ nếu tweet A và B tương tự nhau.

2. Nó là lớn nếu chúng không giống nhau.

3. Nó là 0 nếu chúng giống nhau.

4. Nó là 1 nếu chúng hoàn toàn khác nhau (tức là không có từ trùng lặp).

Thực hiện bài toán:

- Bài toán phân cụm

- Sử dụng kỹ thuật K – Mean, Jaccard

- Tập dữ liệu sử dụng trong thuật toán: Health News in Twitter. Thông tin đầy đủ của tập dữ liệu gồm:

o 3929 dòng dữ liệu

o Id tweet

o Ngày và giờ đăng tweet

o Nội dung tweet

Link tải dataset: <https://archive.ics.uci.edu/ml/datasets/Health+News+in+Twitter>

2.2.2 Tiền xử lý dữ liệu

Các tweet trước khi đưa vào thuật toán sẽ phải tiền xử lý dữ liệu gồm các bước sau:

- Xóa dấu Id và thời gian tweet.

- Các từ bắt đầu bằng @, ví dụ @XuanNam sẽ bị bỏ đi.

- Các biểu tượng thẻ bắt đầu bằng # bị xóa, ví dụ #XuanNam sẽ bị bỏ đi.

- Tất cả Url sẽ bị xóa.

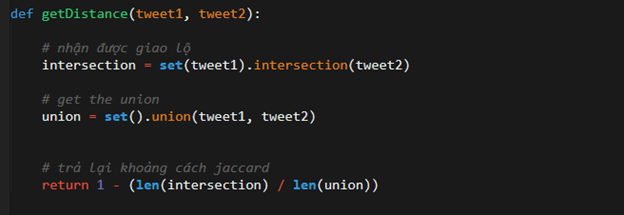
- Chuyển toàn bộ chữ thành chữ thường.

2.2.3 Các bước thực hiện

Bước 1: Đầu tiên xây dựng giải thuật Jaccard.

- Đầu vào sẽ là 2 mảng tweet.

- Đầu ra là độ trùng hợp của 2 mảng tweet (0 -> 1)



Bước 2: Xây dựng thuật toán K – Mean.

- Đầu vào: Số tâm cụm, mảng tweet cần phân cụm, số lần lặp để tìm tâm cụm.

- Đầu ra: mảng tweet đã được đánh dấu tâm cụm tương ứng.

