1. **Giới thiệu dữ liệu**

* Bộ dữ liệu phim: Gồm 3 thể loại
* Hành động
* Tình cảm
* Hài hước

Mỗi thể loại chứa các thông tin về bộ phim và bình luận của từng phim

* Bộ dữ liệu bình luận: Gồm các trường:
* Cmt: bình luận, cấu trúc là 1 chuỗi text
* Rate: đánh giá dựa trên số sao của người dùng đánh giá
* Genre: thể loại gồm 3 thể loại được đánh số: 1 là hành động, 2 là tình cảm, 3 là hài hước
* Idmovie: mã phim
* Iduser: mã người dùng
* Bộ từ điển: Được chia thành 3 file theo các thể loại có cấu trúc gồm 2 cột:
* Keywords: chứa từ khóa tương ứng về thể loại đó (action.csv, romantic.csv, comedy.csv)
* Value: trọng số của từ khóa đó. Trọng số của các từ khóa được cho khi kết hợp với các trạng từ như sau:

|  |  |
| --- | --- |
| vô cùng | 1 |
| cực kì | 0.95 |
| quá | 0.9 |
| rất | 0.85 |
| khá | 0.8 |
| đầy | 0.7 |
| đủ | 0.6 |
| một chút | 0.2 |
| không | -0.5 |
| chẳng | -1 |

1. **Xử lý dữ liệu:**

Bộ dữ liệu về từ điển sau khi được đánh trọng số xong, ta sẽ tiến hành tạo vector đặc trung cho các từ trong bộ từ điển. Độ dài của vector tương ứng số từ khóa trong bộ từ điển.

A screenshot of a computer

Description automatically generated

Hình 1: Vector của phim hành động

Sau khi có vector của bộ từ điển ta sẽ tính độ tương đồng cosine