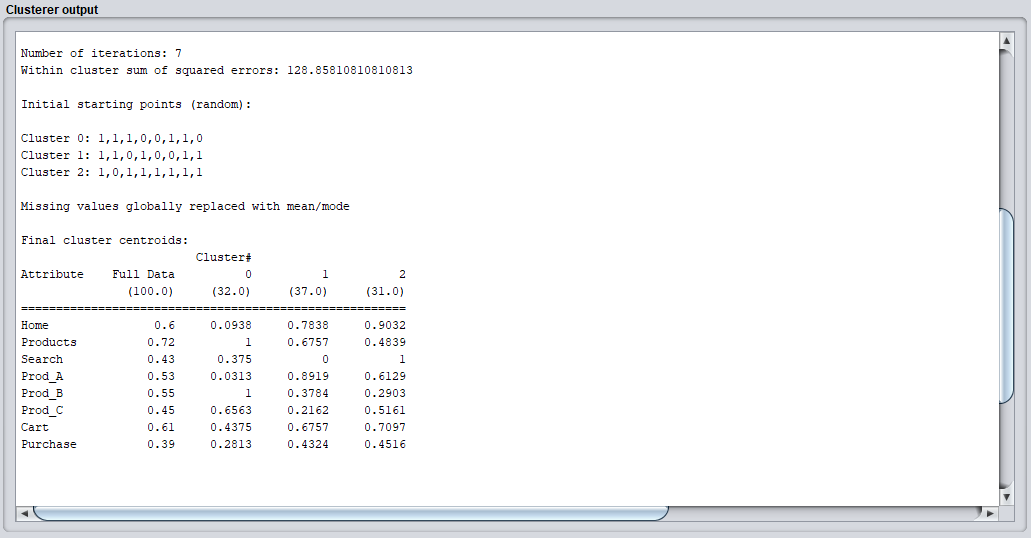
LAB 04: CLUSTERING

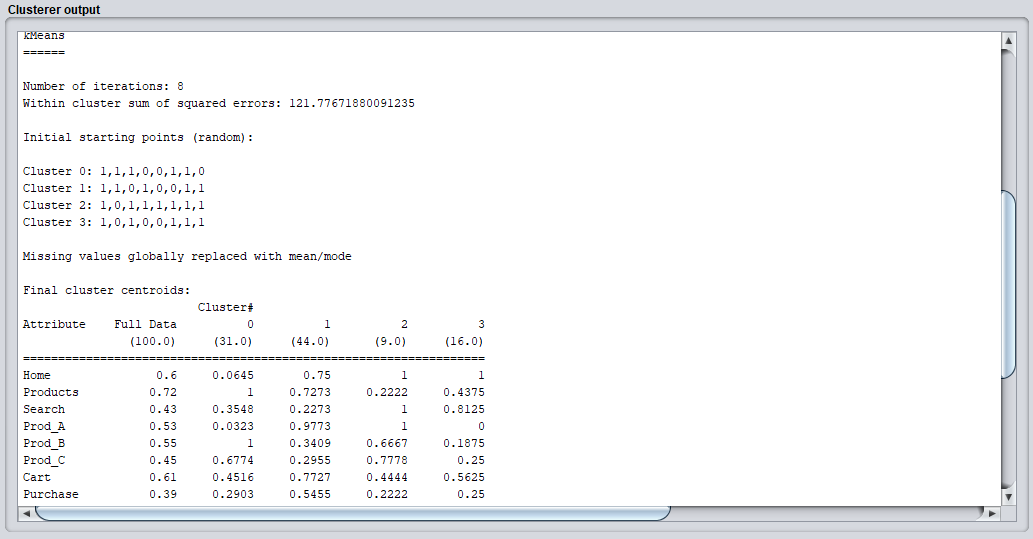
**Nội dung thực hiện báo cáo với ứng dụng WEKA**

1. Lập bảng với các giá trị k từ 3 – 8

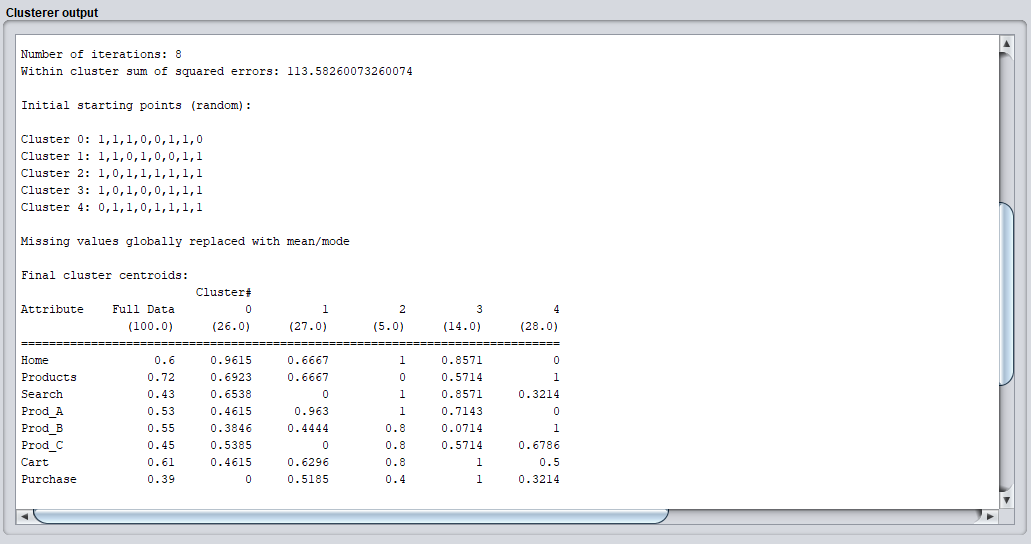
|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| k | SSE | Cluster centroids | | | | | | | | |
|  | Home | Products | Search | Prod\_A | Prod\_B | Prod\_C | Cart | Purchase |
| 3 | 128.85810810810813 | 1 | 0.0938 | 1 | 0.375 | 0.0313 | 1 | 0.6563 | 0.4375 | 0.2813 |
| 2 | 0.7838 | 0.6757 | 0 | 0.8919 | 0.3784 | 0.2162 | 0.6757 | 0.4324 |
| 3 | 0.9032 | 0.4839 | 1 | 0.6129 | 0.2903 | 0.5161 | 0.7097 | 0.4516 |
| 4 | 121.77671880091235 | 1 | 0.0645 | 1 | 0.3548 | 0.0323 | 1 | 0.6774 | 0.4516 | 0.2903 |
| 2 | 0.75 | 0.7273 | 0.2273 | 0.9773 | 0.3409 | 0.2955 | 0.7727 | 0.5455 |
| 3 | 1 | 0.2222 | 1 | 1 | 0.6667 | 0.7778 | 0.4444 | 0.2222 |
| 4 | 1 | 0.4375 | 0.8125 | 0 | 0.1875 | 0.25 | 0.5625 | 0.25 |
| 5 | 113.58260073260074 | 1 | 0.9615 | 0.6923 | 0.6538 | 0.4615 | 0.3846 | 0.5385 | 0.4615 | 0 |
| 2 | 0.6667 | 0.6667 | 0 | 0.963 | 0.4444 | 0 | 0.6296 | 0.5185 |
| 3 | 1 | 0 | 1 | 1 | 0.8 | 0.8 | 0.8 | 0.4 |
| 4 | 0.8571 | 0.5714 | 0.8571 | 0.7143 | 0.0714 | 0.5714 | 1 | 1 |
| 5 | 0 | 1 | 0.3214 | 0 | 1 | 0.6786 | 0.5 | 0.3214 |
| 6 | 109.36117952928299 | 1 | 0.931 | 0.7241 | 0.7241 | 0.5862 | 0.3448 | 0.6552 | 0.5172 | 0.1379 |
| 2 | 0.9583 | 0.875 | 0.083 | 0.9167 | 0.3333 | 0.0833 | 0.7083 | 0.5833 |
| 3 | 1 | 0 | 1 | 1 | 1 | 0.75 | 0.75 | 0.5 |
| 4 | 1 | 0.1667 | 1 | 0.1667 | 0 | 0.3333 | 1 | 0.8333 |
| 5 | 0 | 1 | 0.3214 | 0 | 1 | 0.6786 | 0.5 | 0.3214 |
| 6 | 0 | 0.1111 | 0.1111 | 1 | 0.5556 | 0 | 0.6667 | 0.5556 |
| 7 | 93.79009103641458 | 1 | 0.9048 | 0.9048 | 0.7143 | 0.7619 | 0.381 | 0.9048 | 0.6667 | 0.1905 |
| 2 | 1 | 0.9375 | 0.125 | 0.875 | 0.4375 | 0.125 | 1 | 0.8125 |
| 3 | 1 | 0 | 1 | 1 | 1 | 0.75 | 0.75 | 0.5 |
| 4 | 1 | 0.2857 | 1 | 0.1429 | 0 | 0.2857 | 1 | 0.7142 |
| 5 | 0 | 1 | 0.3214 | 0 | 1 | 0.6786 | 0.5 | 0.3214 |
| 6 | 0 | 0.1429 | 0.1429 | 1 | 0.2857 | 0 | 1 | 0.8571 |
| 7 | 0.8235 | 0.4118 | 0.2941 | 0.6471 | 0.3529 | 0 | 0 | 0 |
| 8 | 88.93190836940838 | 1 | 0.8889 | 1 | 0.6667 | 0.7778 | 0.3889 | 0.9444 | 0.6667 | 0.2222 |
| 2 | 1 | 0.9375 | 0.125 | 0.875 | 0.4375 | 0.125 | 1 | 0.8125 |
| 3 | 1 | 0 | 1 | 1 | 0.6667 | 0.8333 | 0.6667 | 0.3333 |
| 4 | 1 | 0.5 | 1 | 0 | 0 | 1 | 1 | 1 |
| 5 | 0 | 1 | 0.3214 | 0 | 1 | 0.6786 | 0.5 | 0.3214 |
| 6 | 0 | 0.1429 | 0.1429 | 1 | 0.2857 | 0 | 1 | 0.8571 |
| 7 | 0.75 | 0.5 | 0 | 0.8333 | 0.8333 | 0 | 0 | 0 |
| 8 | 1 | 0.2727 | 1 | 0.1818 | 0.2727 | 0 | 0.5455 | 0.2727 |



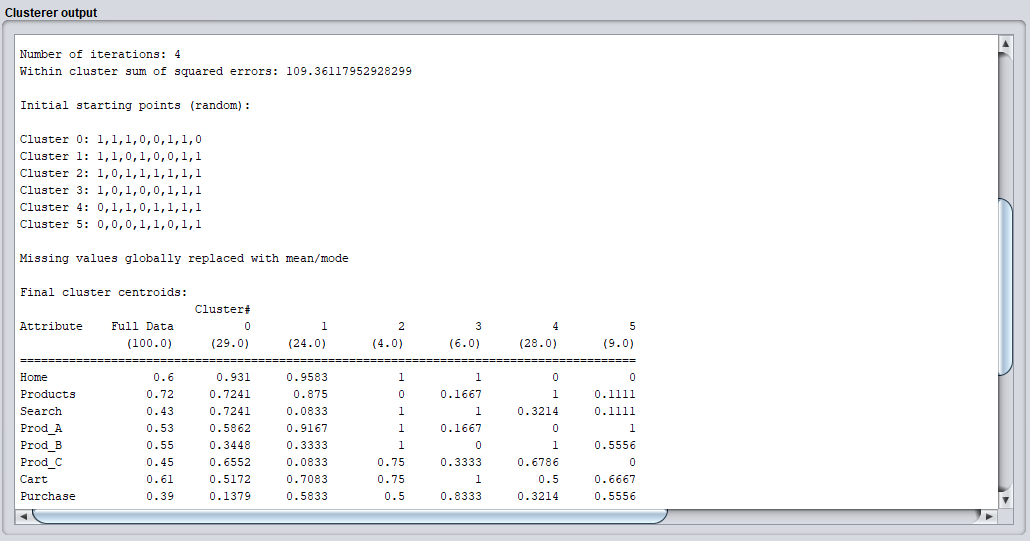
(Hình 1.1. k = 3)



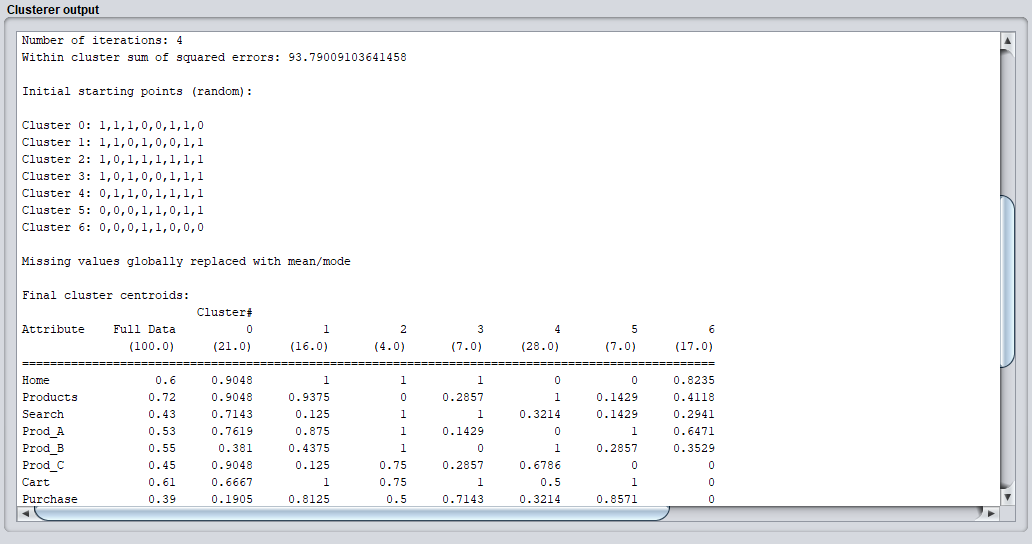
(Hình 1.2. k = 4)



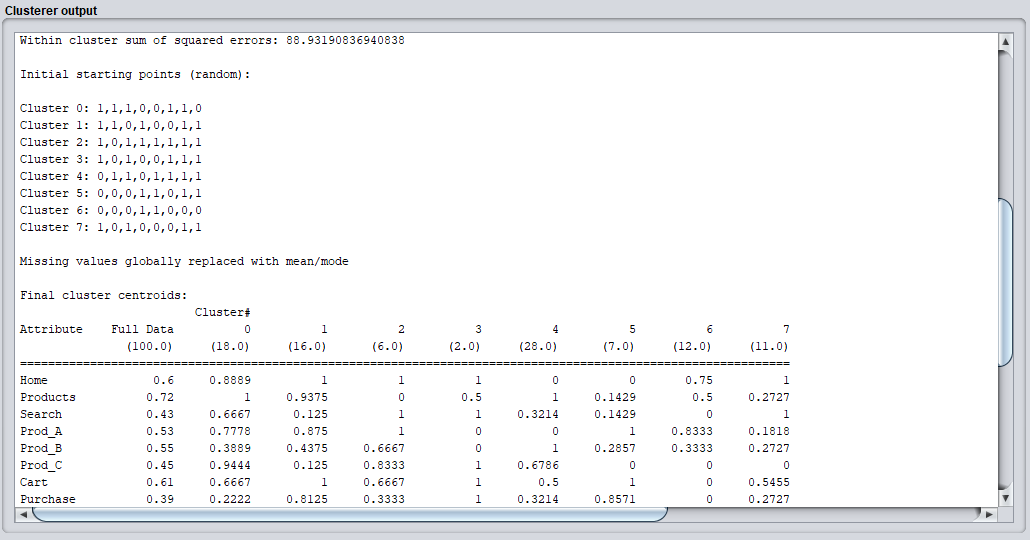
(Hình 1.3. k = 5)



(Hình 1.4. k = 6)



(Hình 1.5. k = 7)



(Hình 1.6. k = 8)

* Từ kết quả bảng trên, chọn 1 kết quả gom k cụm …

1. Giả sử bạn quan sát thấy một người dùng mới đã truy cập các trang là Home => Search => Prod\_B. Bạn sẽ giới thiệu sản phẩm nào đến người này?

* Người dùng này tương ứng với điểm mới (1, 0, 1, 0, 1, 0, 0, 0)
* Danh sách khoảng cách (tính theo norm2 – L2) giữa điểm mới với các cụm

[]

Suy ra: thuộc cụm

* Trong cụm, ta thấy rằng:

Prod\_A = ? Prod\_C = -> chọn Prod\_

**Vậy: Nhóm sẽ giới thiệu sản phẩm …**

1. Tương tự Câu 2., lần này người dùng truy cập các trang Products => Prod\_C

* Người dùng này tương ứng với điểm mới (0, 1, 0, 0, 0, 1, 0, 0)
* Danh sách khoảng cách (tính theo norm2 – L2) giữa điểm mới với các cụm

[]

Suy ra: thuộc cụm

* Trong cụm, ta thấy rằng:

Prod\_A = ? Prod\_B = -> chọn Prod\_

**Vậy: Nhóm sẽ giới thiệu sản phẩm …**

1. Kết quả gom cụm chọn có thể nhận diện được các hình mẫu người dùng dưới đây hay không?

* Người dùng thông thường (window shopper, xem nhiều sản phẩm):
* Người dùng tập trung (biết cần mua rõ sản phẩm gì)
* Người dùng tìm kiếm (sử dụng chức năng search để tìm mua sản phẩm cần mua)

1. Từ kết quả gom cụm mà bạn chọn, có cụm nào thể hiện sở thích mua hàng cụ thể của người dùng đối với sản phẩm đơn lẻ hay nhóm các sản phẩm hay không?
2. Giả sử rằng ABC đã đặt các banner quảng cáo lên một số trang nổi tiếng khác và những banner này trỏ trực tiếp đến trang của các sản phẩm A và B.

* Có, cụm

**Nội dung thực hiện cài đặt**

1. Chạy chương trình với tập dữ liệu session.csv
   1. Kết quả phát sinh được với kết quả của Weka

Ảnh có chứa ảnh chụp màn hình, màn hình, đen

Mô tả được tạo tự động

(K = 3 – Hình 2.model\_3.png)

Ảnh có chứa ảnh chụp màn hình, màn hình, đen, tường

Mô tả được tạo tự động

(K = 4 – Hình 2.model\_4.png)

Ảnh có chứa ảnh chụp màn hình, màn hình, đen, tường

Mô tả được tạo tự động

(K = 5 – Hình 2.model\_5.png)

Ảnh có chứa ảnh chụp màn hình, màn hình, đen, tường

Mô tả được tạo tự động

(K = 6 – Hình 2.model\_6.png)

Ảnh có chứa ảnh chụp màn hình, màn hình, đen, tường

Mô tả được tạo tự động

(K = 7 – Hình 2.model\_7.png)

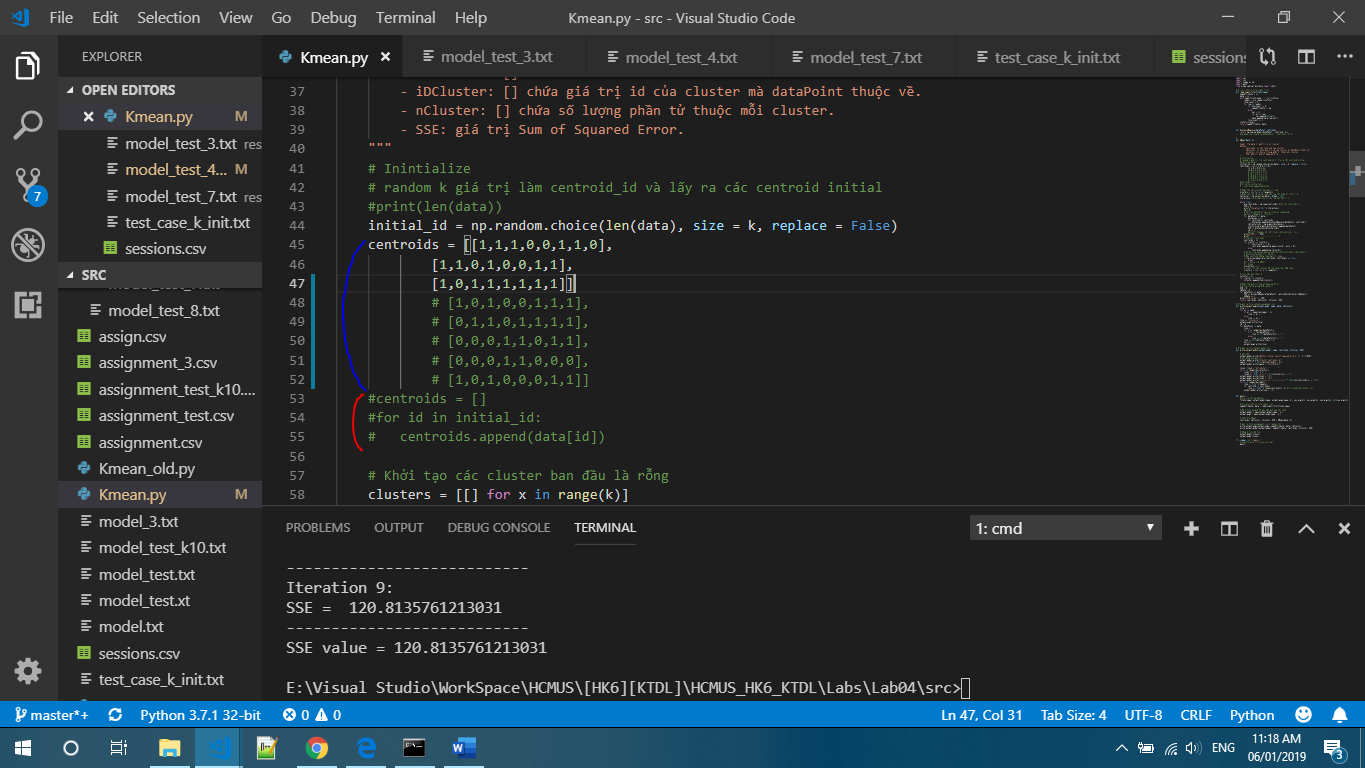
Ảnh có chứa ảnh chụp màn hình, màn hình, đen, tường

Mô tả được tạo tự động

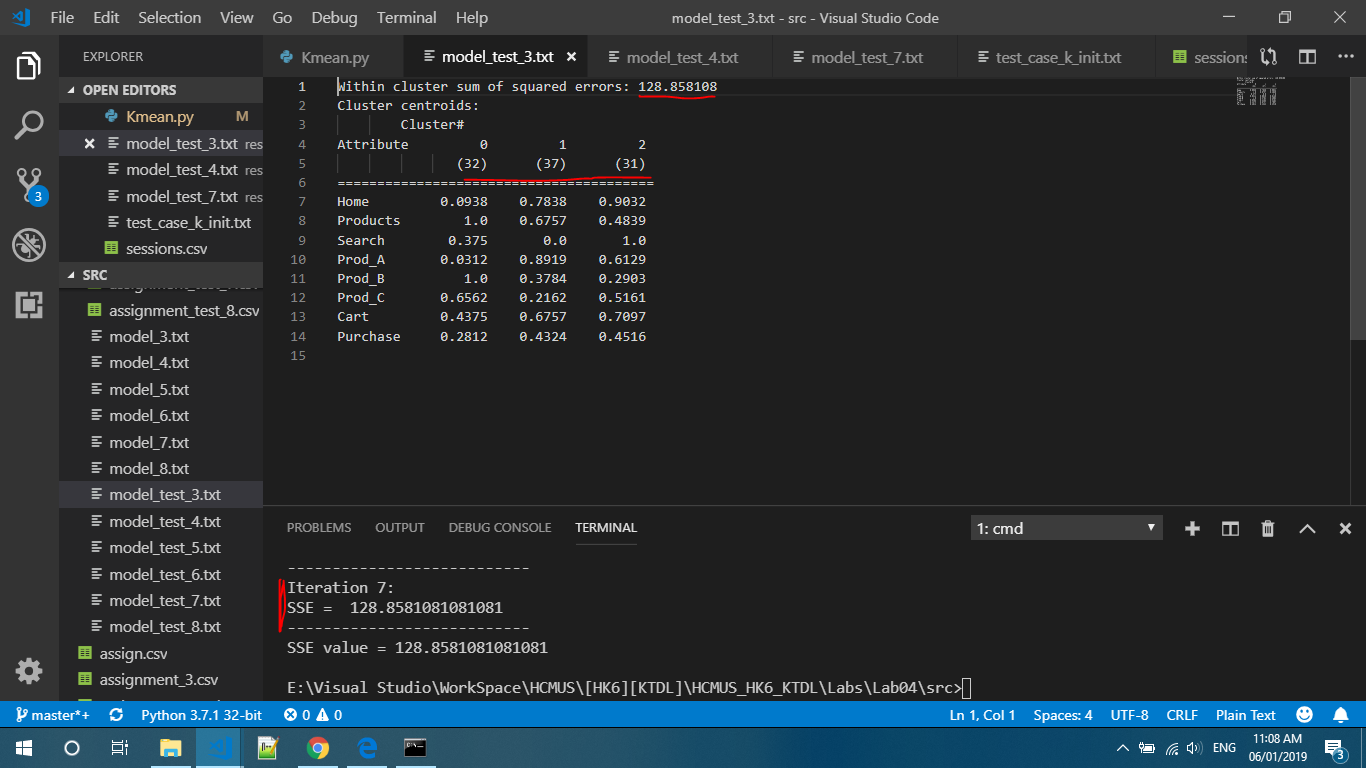
(K = 8 – Hình 2.model\_8.png)

* 1. Kết quả chạy với các **Initial starting points** của Weka

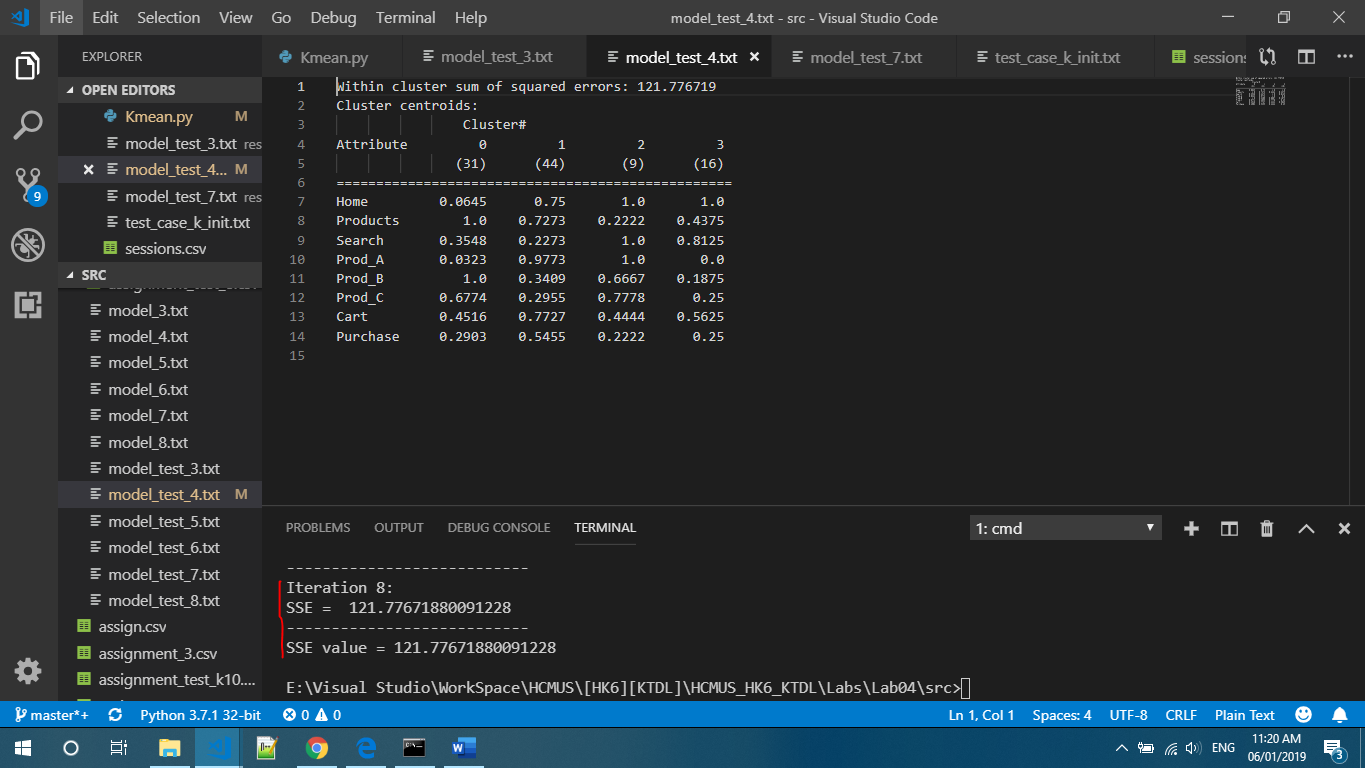
Ở đoạn code này, nhóm thiết lập các initial starting points giống với bên weka để kiểm tra phần cài đặt giải thuật k-mean của nhóm. (*comment phần màu đỏ, copy từ file test\_case\_k\_init.txt ứng với số lượng k cụm muốn phân*).



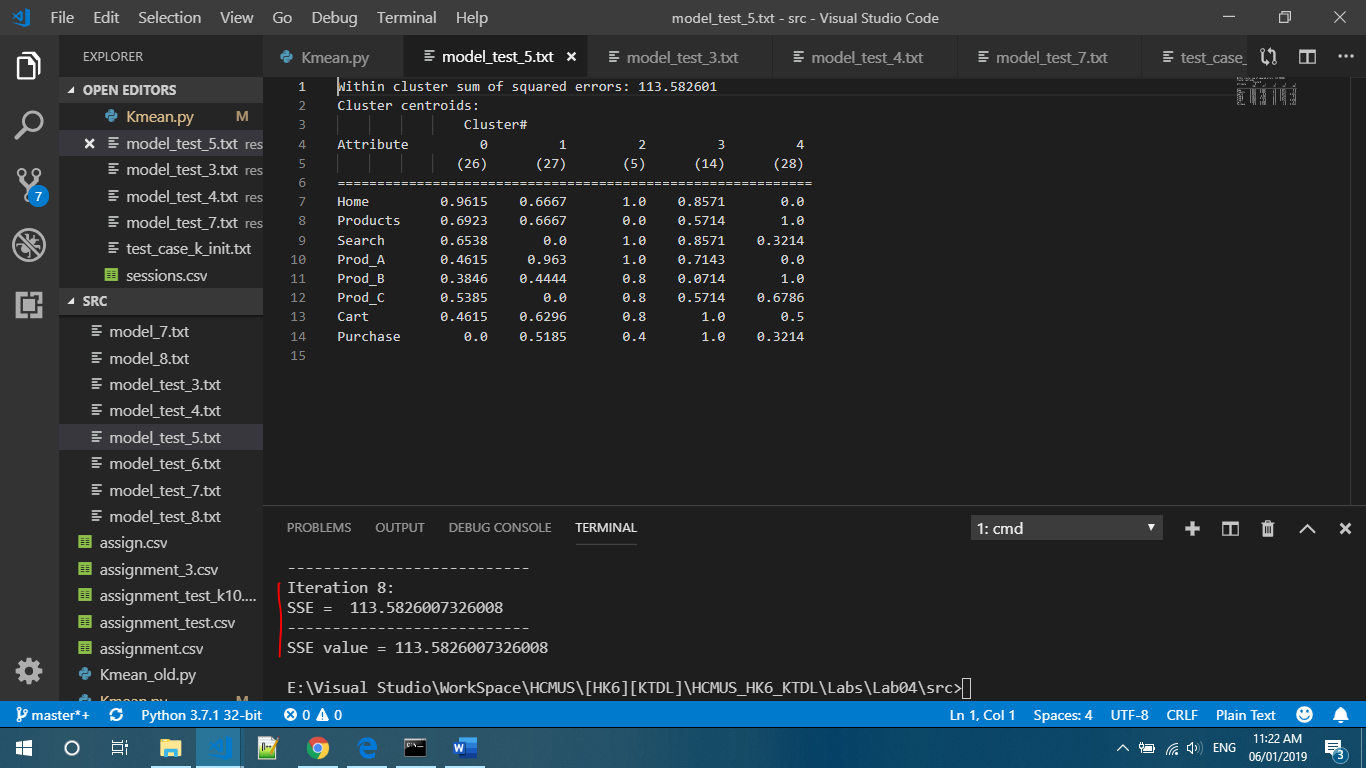
* K = 3 (2.model\_test\_3.png)



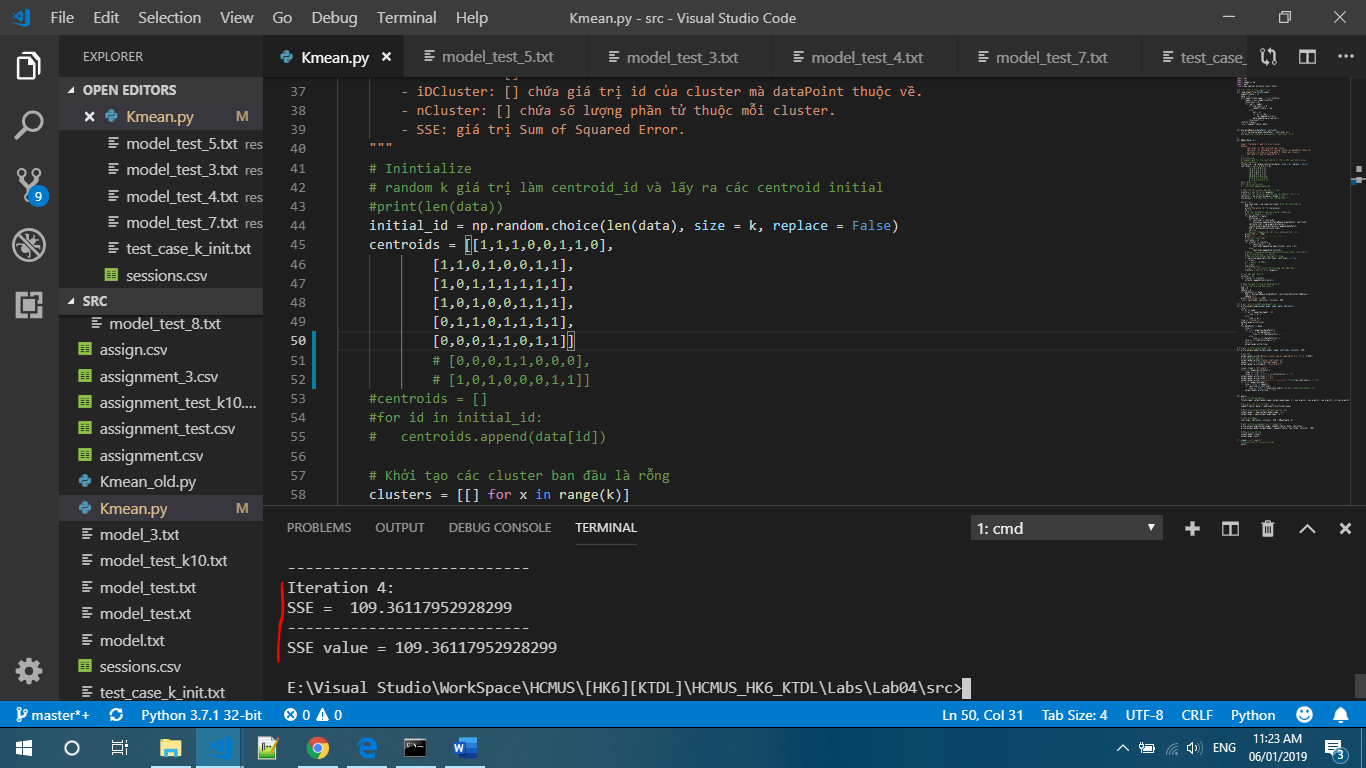
* K = 4 (2\_model\_test\_4.png)



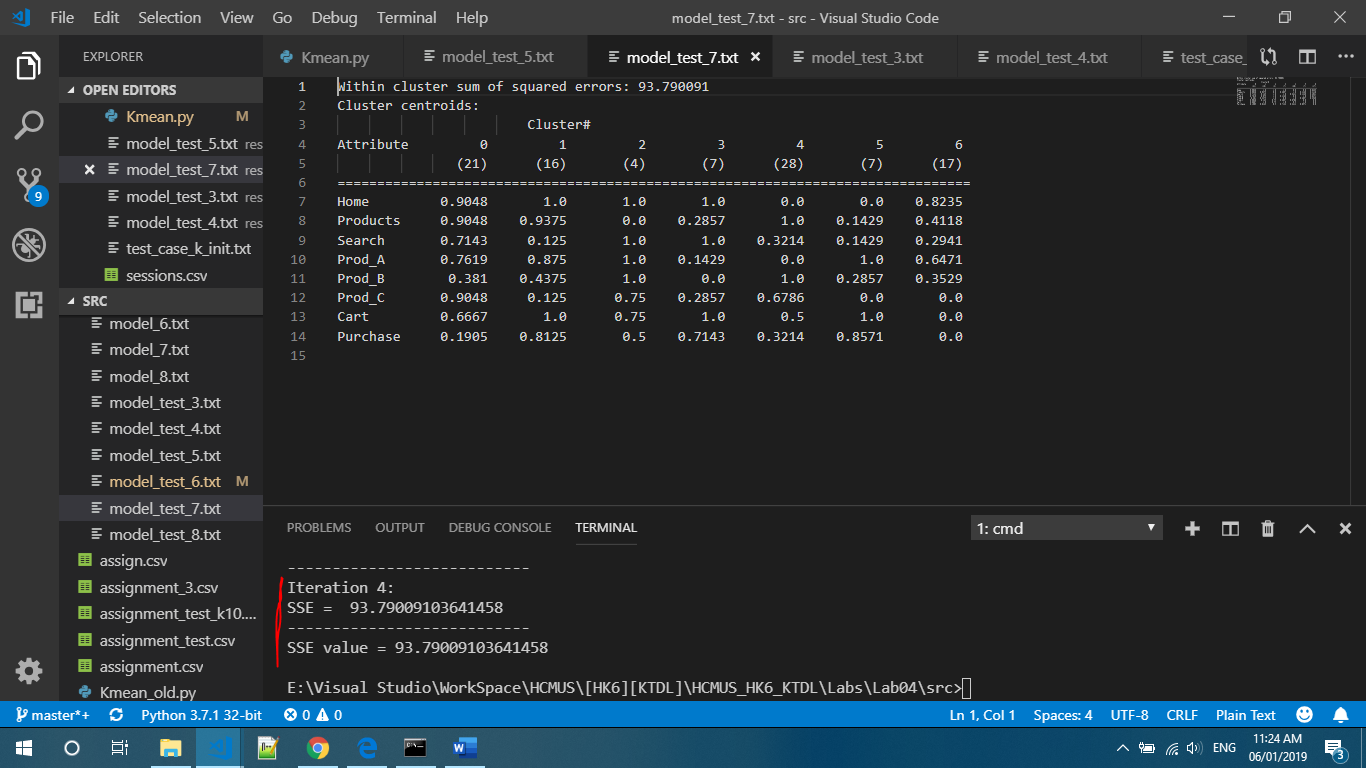
* K = 5



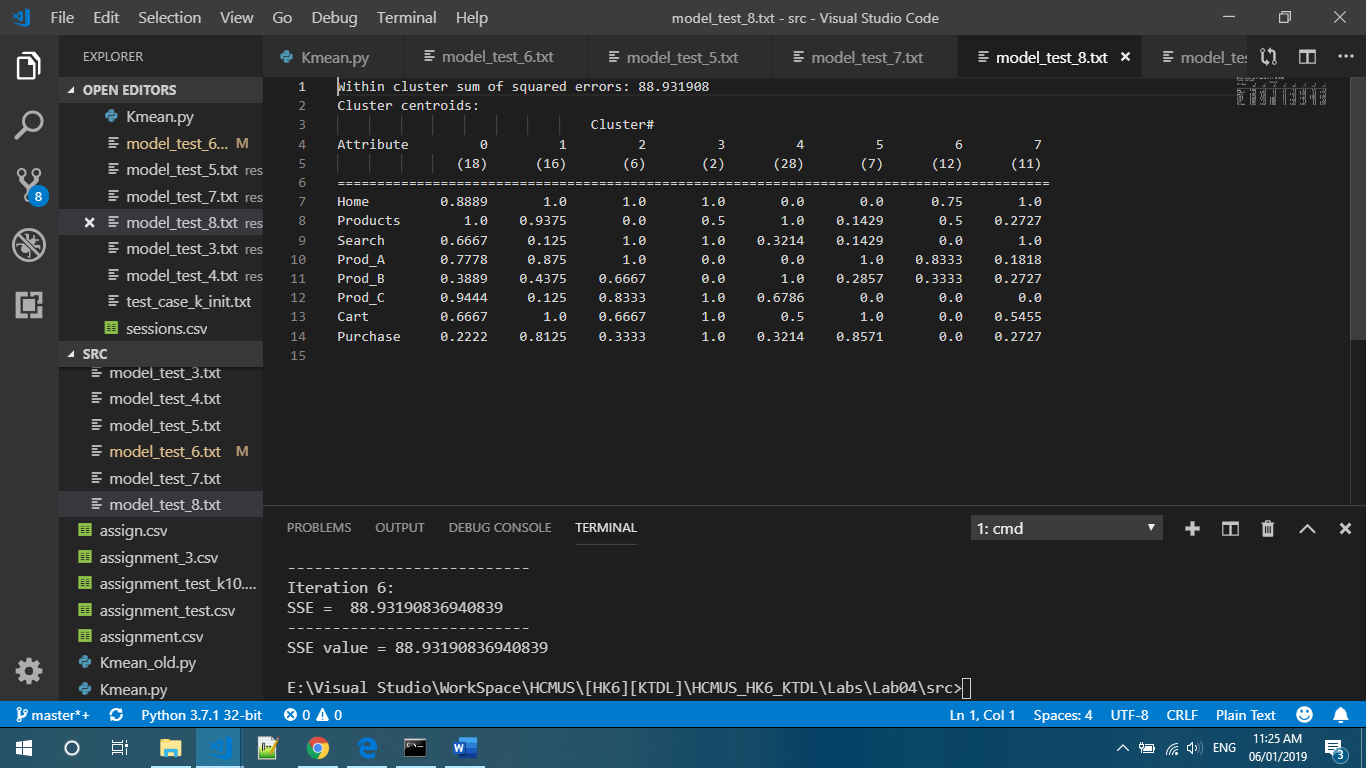
* K = 6



* K = 7



* K = 8



* 1. Nhận xét về cài đặt của nhóm với cài đặt của weka
* Giống:
* Khác: kết quả gom cụm và SSE khác nhau vì các điểm khởi tạo ban đầu (init centroid) của nhóm được lấy một cách ngẫu nhiên từ tập dữ liệu, việc lấy random khác so với weka dẫn đến lấy các điểm khởi tạo ban đầu khác.