Clustering Grocery Items

Goal

Online shops often sell tons of different items and this can become very messy very quickly!

Data science can be extremely useful to automatically organize the products in categories so that they can be easily found by the customers.

The goal of this challenge is to look at user purchase history and create categories of items that are likely to be bought together and, therefore, should belong to the same section.

Challenge Description

Company XYZ is an online grocery store. In the current version of the website, they have manually grouped the items into a few categories based on their experience.

However, they now have a lot of data about user purchase history. Therefore, they would like to put the data into use!

This is what they asked you to do:

- The company founder wants to meet with some of the best customers to go through a focus group with them. You are asked to send the ID of the following customers to the founder:
    - the customer who bought the most items overall in her lifetime
    - for each item, the customer who bought that product the most
- Cluster items based on user co-purchase history. That is, create clusters of products that have the highest probability of being bought together. The goal of this is to replace the old/manually created categories with these new ones. Each item can belong to just one cluster.

Data

We have 2 table downloadable by clicking **here**.

The 2 tables are:

```
"item_to_id" - for each item, it gives the corresponding id
```

Columns:

- **Item_name** : the name of the item
- **Item_id** : the id of the item. Can be joined to the id in the other table. It is unique by item

```
"purchase_history" - for each user purchase, the items bought
```

Columns:

- **user_id** : the id of the user.
- **id** : comma-separated list of items bought together in that transaction.

Example

```
    Let's check what one user bought:
```

head(purchase_history,1)

Column Name	Value	Description
user_id	222087	this is simply the user id
id	27,26	this means the user bought together item 27 and 26 in that purchase event. All co-purchased items are listed under the same column and separated by a comma .

```
    Let's check what is item 26 for instance:
```

subset(item_to_id,Item_id==26)

Column Name	Value	Description
Item_name	spaghetti sauce	she bought spaghetti sauce
Item_id	26	spaghetti sauce id is 26.