

Sistemas de Recomendación:

Reducción de Dimensionalidad
con factorización SVD y
Matriz de Utilidad

Inteligencia Artificial y Aprendizaje Automático

¿Cómo diseñar un algoritmo o sistema de recomendación?



Porque viste "Perros de reserva"



Netflix lanza en octubre del 2006 el reto, con una bolsa de un millón de dólares, de ver quién podría mejorar su sistema de recomendación en al menos un 10%.



Leaderboard

Showing Test Score. [Click here to show quiz score](#)

Display top leaders.

Rank	Team Name	Best Test Score	% Improvement	Best Submit Time
Grand Prize - RMSE = 0.8567 - Winning Team: BellKor's Pragmatic Chaos				
1	BellKor's Pragmatic Chaos	0.8567	10.06	2009-07-26 18:18:28
2	The Ensemble	0.8567	10.06	2009-07-26 18:38:22
3	Grand Prize Team	0.8582	9.90	2009-07-10 21:24:40
4	Opera Solutions and Vandelay United	0.8588	9.84	2009-07-10 01:12:31
5	Vandelay Industries	0.8594	9.84	2009-07-10 00:00:00

<https://web.archive.org/web/20090924184639/http://www.netflixprize.com/community/viewtopic.php?id=1537>

A black background with the Netflix logo in red at the top. Below it, text in white and yellow says "Tenemos 100 millones de miembros y más de 250 millones de experiencias únicas cada día". Then, "¿CÓMO LO HACEMOS?" in large white letters, followed by "Todo depende del perfil con el que se mire" in red. Below that, "Este es tu perfil" with an arrow pointing to a green smiley face icon labeled "Tú". To the left of the icon, text says "Lo que más importa es lo que viste recientemente". To the right, a box says "Nuestros algoritmos vuelven a calcular tus recomendaciones al menos cada 24 horas".

NETFLIX

Tenemos **100 millones** de miembros y más de **250 millones** de experiencias únicas cada día

¿CÓMO LO HACEMOS?

Todo depende del perfil con el que se mire

Este es tu perfil

Lo que más importa es lo que viste recientemente

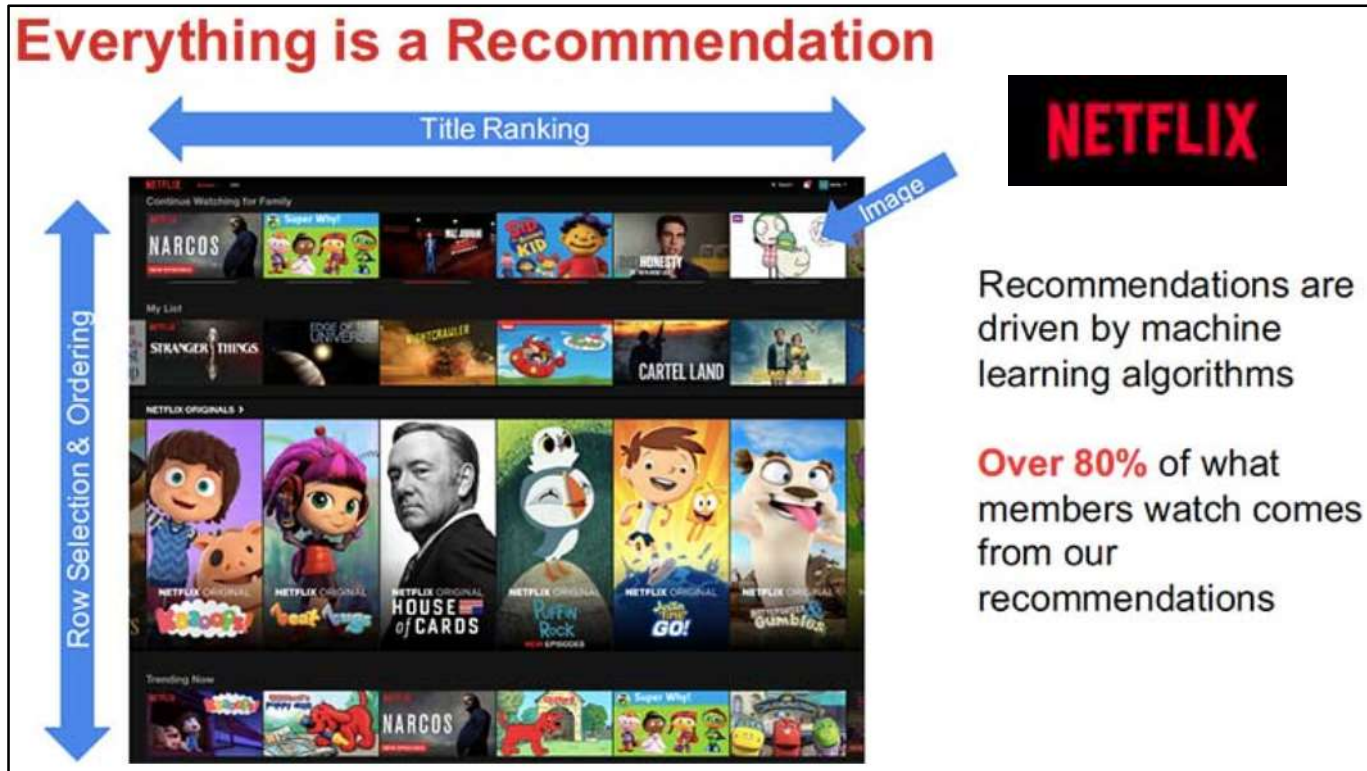
Nuestros algoritmos vuelven a calcular tus recomendaciones al menos cada 24 horas

Tú

Sistemas de Recomendación


























Objetivo: Recomendar al usuario artículos/servicios/lugares de su interés.

Así, se desea personalizar la experiencia del usuario mediante algoritmos de aprendizaje automático (*machine learning*).



Dentro de los Sistemas de Recomendación existen dos problemáticas principales:

- *The Prediction problem*

					
A					
B					
C					
D					
E					

¿Cuál de mis productos que no hay visto un usuario puedo recomendarle?

- *The Ranking problem*



No solamente se desea conocer las preferencias de un usuario en particular, sino que también se desea saber qué k productos recomendarle como primeras opciones en sus diferentes categorías.

Tipos de Sistemas de Recomendación

Existen varios tipos de sistemas de recomendación, en particular los podemos clasificar en los siguientes tres más populares en la actualidad de acuerdo a la cantidad y calidad de los datos que se tienen:

- **Sistemas de Filtrado Colaborativo**
(Collaborative Filtering):

Las recomendaciones se basan a partir de la agrupación de usuarios con gustos similares:

- Sistemas basados en el usuario.
- Sistemas basados en el artículo.
- Basados en factorización matricial.

Para funcionar, estos sistemas requieren tener una gran cantidad de datos de usuarios y de los productos que han comprado en el pasado.

- **Sistemas basados en Popularidad**

- **Sistemas basados en Contenido**
(Content-based systems):

Las recomendaciones no requieren de información relacionada con el pasado, en su lugar, utilizan el perfil del usuario o información proporcionada por él.

- **Recomendadores basados en Popularidad**

Es el método más sencillo de implementar, aunque no dirigido a personalizar la experiencia del usuario.

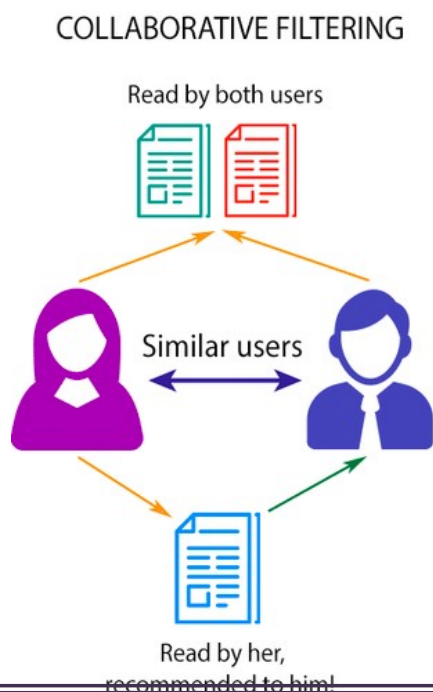
En particular, es una de las opciones que se puede utilizar para enfrentar el *cold start problem*.



- **Sistemas basados en el Filtrado Colaborativo**

Las recomendaciones se basan a partir de la agrupación de usuarios con gustos similares. Es decir, se identifican comunidades de usuarios con preferencias similares, usando inclusive diferentes criterios pueda agruparlo en diferentes comunidades.

Este tipo de recomendadores plantea mayores retos, ya que trata de identificar un tipo de “inteligencia colectiva” entre grupos de usuarios, usando información no solamente de los productos de su preferencia, sino también de lo que escriben en sus redes sociales, preferencias en otras áreas como política, deportes, etc.



- Sistemas basados en el usuario:
(*user-based filtering*)

- Sistemas basados en el producto:
(*item-based filtering*)

Customers who bought this item also bought



Lo que tienen en común estos sistemas es que requieren de un historial de datos relacionados con su actividad en el pasado.

- En general se pueden desarrollar sistemas mixtos, basados tanto en el usuario como en el producto.

Inspired by your browsing history



Related to Items You've Viewed

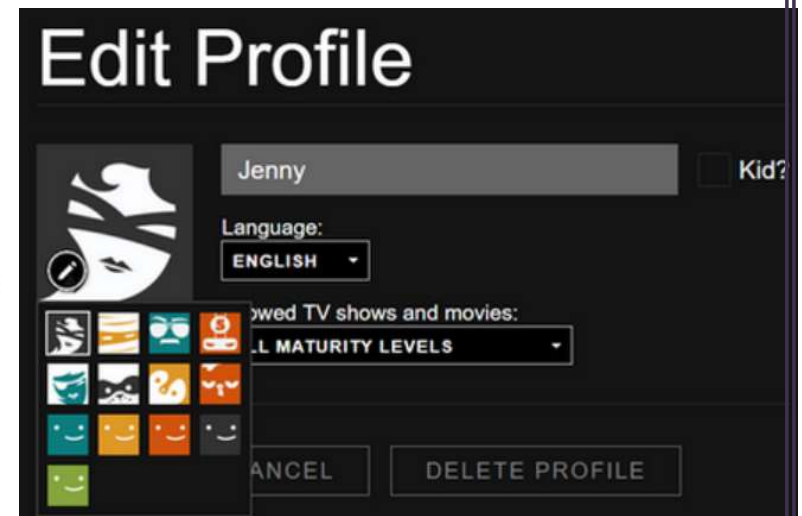
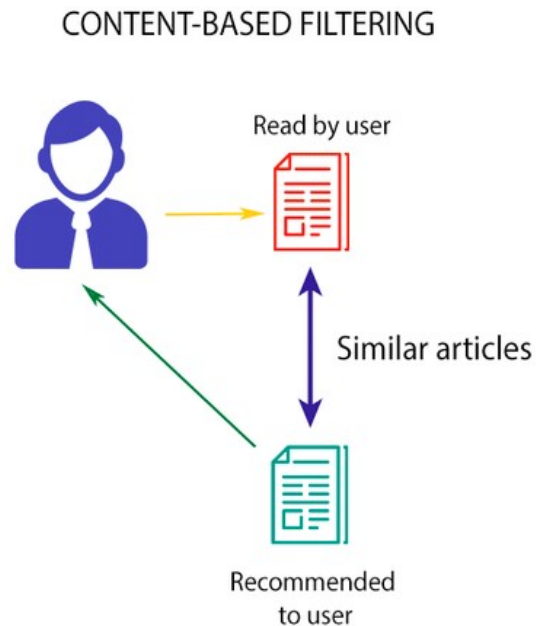


- **Recomendadores basados en Contenido**

Las recomendaciones se basan, no usando el nulo o poco historial que se tenga del usuario, sino que utilizan su perfil y las preferencias que indicó. Se basan principalmente en analizar las características de productos que el usuario ha indicado que le gustan para recomendarle productos similares.

Dentro del área de Machine Learning este problema se puede abordar como un problema de clasificación, donde con base a las características de los productos que el usuario ha mostrado preferencia (*like, dislike*), se trata de inferir cuál otro nuevo producto le agradará.

En general, estos recomendadores estarán sugiriendo productos similares a los que ha mostrado interés el usuario, pero no de otras áreas o categorías.



Cold Start Problem

Los sistemas de recomendación requieren información del usuario o de los productos para hacer una recomendación.

Por ejemplo: los productos que el usuario a comprado o visto, la información personal que pueda compartir; los productos más vendidos de acuerdo a la región o fechas del calendario, o en relación a otros productos comprados en conjunto, etc.

¿Cómo llevar a cabo dichas recomendaciones cuando no se tiene dicha información?

¿Cómo promover que dicha información se genere lo más rápido posible?

- **Sistemas Híbridos**

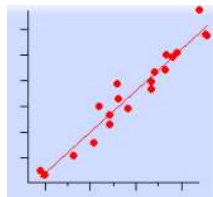
- En general no se utiliza un único sistema de recomendación.
- Por ejemplo con usuarios nuevos se puede iniciar con las recomendaciones por popularidad o el basado en contenido a través de su perfil.
- Posteriormente, cuando se tenga mayor información e historial del usuario, se puede empezar a perfilar al usuario e identificarlo con algunas comunidades para usar filtrado colaborativo.
- Siempre se mantienen vigentes los tres casos, para aprovechar lo mejor de ellos en cada usuario.

Medidas de Similitud entre Usuarios/Artículos

$$\text{sim}(\vec{u}, \vec{w})$$

En particular, podemos mencionar estas dos medidas de similitud de dos vectores en un espacio R^n :

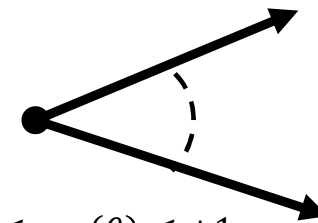
- **Coeficiente de Correlación de Pearson**



$$-1 \leq \rho \leq +1$$

$$\rho = \frac{\text{cov}[\vec{u}, \vec{w}]}{\sqrt{\text{Var}[\vec{u}] \text{Var}[\vec{w}]}}$$

- **Coseno del Ángulo**



$$-1 \leq \cos(\theta) \leq +1$$

$$\cos(\theta) = \frac{\langle \vec{u}, \vec{w} \rangle}{\|\vec{u}\| \|\vec{w}\|}$$



movielens

Non-commercial, personalized movie recommendations.

[sign up now](#)

or [sign in](#)

grouplens

[about](#)

[datasets](#)

[publications](#)

[blog](#)

MovieLens

GroupLens Research has collected and made available rating data sets from the MovieLens web site (<http://movielens.org>). The data sets were collected over various periods of time, depending on the size of the set. Before using these data sets, please review their README files for the usage licenses and other details.

Help our research lab: Please [take a short survey](#) about the MovieLens datasets

Datasets

[MovieLens](#)

[WikiLens](#)

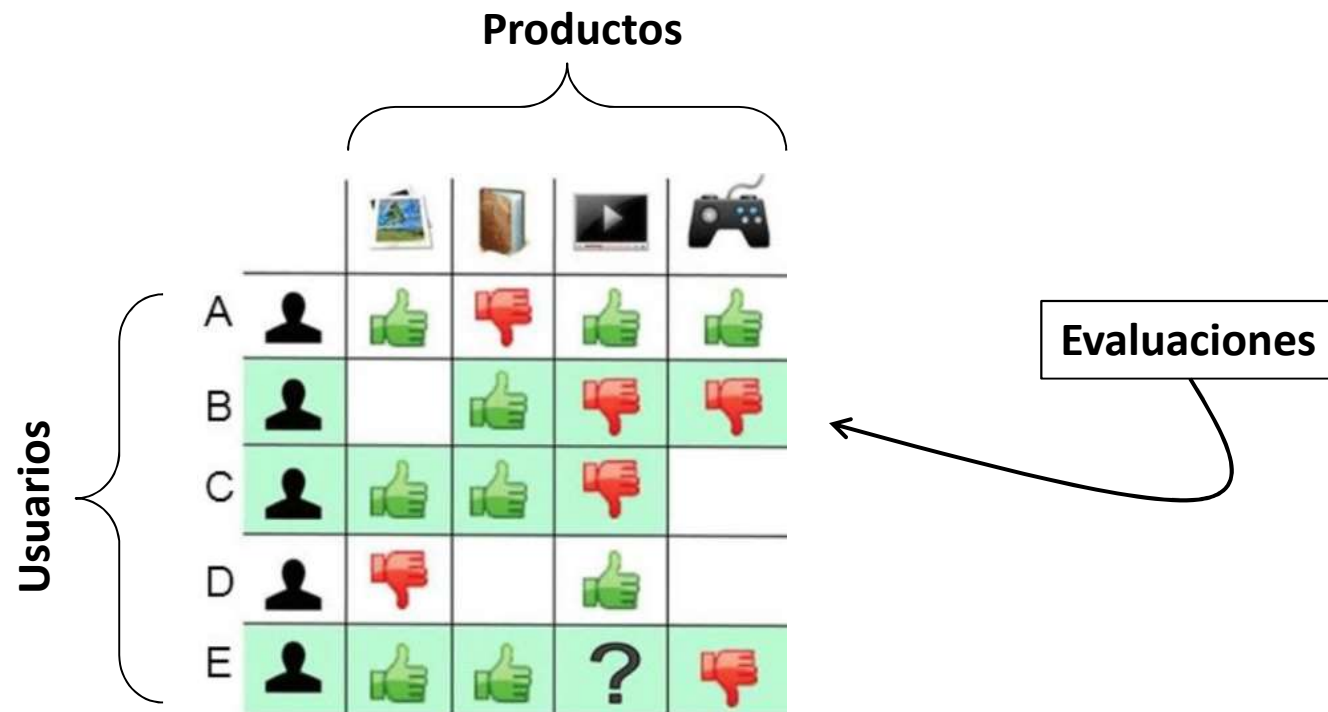
[Book-Crossing](#)

[Jester](#)

<https://grouplens.org/datasets/movielens/>

La información que requiere para empezar a construir sistemas de recomendación se basa en estos tres elementos: el usuario, el producto que adquiere y la calificación que le asigna:

```
> str(data)
'data.frame':  100000 obs. of  3 variables:
 $ userID : int  196 186 22 244 166 298 115 253 305 6 ...
 $ movieID: int  242 302 377 51 346 474 265 465 451 86 ...
 $ rating : int   3 3 1 2 1 4 2 5 3 3 ...
>
```



Matriz de Utilidad

De manera general los sistemas de recomendación se apoyan en conjuntos de datos que se encuentran divididos de acuerdo a las siguientes dos categorías:

- Usuarios
- Artículos

Cada usuario j tiene preferencia por algún artículo k , y lo evalúa de acuerdo un valor r_{jk}

Llamaremos **Matriz de Utilidad** a aquella matriz que tiene la información de usuarios y artículos. Por ejemplo, podríamos tenerlos en el siguiente formato: cada renglón j representa a un usuario, el cual ha evaluado mediante r_{jk} al artículo que se encuentra en la columna k .

	MATRIZ DE UTILIDAD					
	a1	a2	a3		ak	
u1	r11	r12	r13			
u2	r21	r22	r23			
u3	r31	r32	r33			
u4						
uj					rjk	

Usuarios u_j : suscriptores de MovieLens

Artículos a_k : películas de la base de datos de MovieLens 100K

En general esta es una matriz dispersa (*sparse matrix*) y se deben aplicar los algoritmos adecuados para su procesamiento (en general las librerías lo hacen de manera automática).

Teorema de la Descomposición SVD

Sea la matriz $A_{m \times n}$ con valores singulares

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$$

$$\sigma_{r+1} = \sigma_{r+2} = \cdots = \sigma_n = 0$$

entonces $A_{m \times n}$ puede expresarse como:

$$A_{m \times n} = U_{m \times m} \Sigma_{m \times n} V_{n \times n}^T$$

donde $\Sigma_{m \times n} = \begin{bmatrix} D_{r \times r} & 0_{r \times (n-r)} \\ 0_{(m-r) \times r} & 0_{(m-r) \times (n-r)} \end{bmatrix}$

$$D_{r \times r} = \begin{bmatrix} \sigma_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_r \end{bmatrix}, \text{ matriz diagonal de valores singulares no cero.}$$

$$V_{n \times n} = [v_1 \quad \cdots \quad v_n], \text{ formada con los eigenvectores columna de } A^T A.$$

$$U_{m \times m} = [u_1 \quad \cdots \quad u_m], \text{ donde } u_k = \frac{1}{\sigma_k} A v_k$$

Teorema de aproximación SVD / Descomposición SVD Truncada

A partir de la descomposición SVD:

$$A = U\Sigma V^T$$

Se tiene que:

$$A = \sigma_1 u_1 v_1^T + \sigma_2 u_2 v_2^T + \cdots + \sigma_r u_r v_r^T$$

Esta descomposición será importante para la obtención de lo que llamaremos más adelante las variables latentes.

Y podemos entonces aproximar la matriz A con respecto a los $d \leq r$ sumandos más grandes.

$$A \approx \sigma_1 u_1 v_1^T + \sigma_2 u_2 v_2^T + \cdots + \sigma_d u_d v_d^T$$

De donde la cantidad de información contenida en dicha aproximación está dada como:

$$\frac{\sum_{k=1}^d \lambda_k}{\sum_{k=1}^r \lambda_k}$$

Y por lo tanto la reducción de información con dicha aproximación es:

$$1 - \frac{\sum_{k=1}^d \lambda_k}{\sum_{k=1}^r \lambda_k}$$

Los pasos para obtener la descomposición SVD de una matriz $A_{m \times n}$ es como sigue:

1. Obetener $B_{n \times n} = A^T A$
2. Obtener los n eigenvalores λ_k y sus n eigenvectores e_k de B .
3. Los m valores singulares no cero de A son $\sigma_k = \sqrt{\lambda_k}$. Formamos la matriz diagonal $\Sigma_{m \times n}$.
4. La matriz con los n vectores normalizados $v_k = \frac{e_k}{\|e_k\|}$ forman la matriz $V_{n \times n}$.
5. La matriz con los m primeros vectores $\frac{1}{\sigma_k} A v_k$ forman la matriz $U_{m \times m}$.

Finalmente se tiene:

$$A_{m \times n} = U_{m \times m} \Sigma_{m \times n} V_{n \times n}^T$$

Observa que las tres matrices de la descomposición SVD se obtienen de los eigenvalores y eigenvectores de $A^T A$.

¿Cómo nos puede ayudar la descomposición SVD al aplicarla a la Matriz de Utilidad?

	MATRIZ DE UTILIDAD				
	a1	a2	a3		ak
u1	r11	r12	r13		
u2	r21	r22	r23		
u3	r31	r32	r33		
u4					
uj					rjk

Usuarios u_j : suscriptores de MovieLens

Artículos a_k : películas de la base de datos de MovieLens 100K

Primeramente, usaremos el concepto de Similitud para establecer que usuarios o películas (vistos como vectores renglón o columna) están más relacionadas entre sí.

```
<usuarios: 943>
```

MATRIZ DE UTILIDAD						
	a1	a2	a3		ak	
u1	r11	r12	r13			
u2	r21	r22	r23			
u3	r31	r32	r33			
u4						
uj					rjk	

<943>

[illegible]

¿Cómo podrían utilizarse cada una de estas matriceas de variables latentes?

<películas: 1664>

Descomposición SVD directa:
Traspuesta de UtMX \rightarrow
SVD $\rightarrow u * d * v$

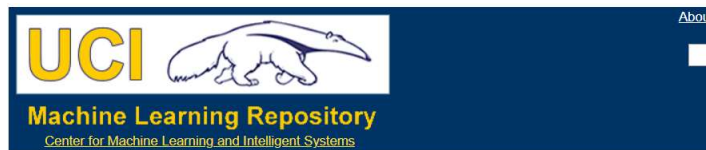
<películas: 1664>

matriz u

Matriz de Variables Latentes de las películas

Actividad de la semana:

Usarás la siguiente base de datos para diseñar un sistema de recomendación con base a la evaluación que usuarios hicieron al servicio de varios restaurantes en México. Te apoyarás en los archivos **rating_final.csv** y **geoplaces2.csv**.



Restaurant & consumer data Data Set

Download: [Data Folder](#), [Data Set Description](#)

Abstract: The dataset was obtained from a recommender system prototype. The task was to generate a top-n list of r

Data Set Characteristics:	Multivariate	Number of Instances:	138	Area:	Computer
Attribute Characteristics:	N/A	Number of Attributes:	47	Date Donated	2012-08-04
Associated Tasks:	N/A	Missing Values?	Yes	Number of Web Hits:	94865

<https://archive.ics.uci.edu/ml/datasets/Restaurant+%26+consumer+data>

	userID	placeID	rating	food_rating	service_rating
0	U1077	135085	2	2	2
1	U1077	135038	2	2	1
2	U1077	132825	2	2	2
3	U1077	135060	1	2	2
4	U1068	135104	1	1	2

rating_final.csv