# RECOMENDADOR DE CERVEZAS

Trabajo Fin de Máster

Antonio José Arcones Olmo

## INDICE

## 

Introducción 3

Datos de Entrada y Datos Calculados 6

Pipeline 9

Lista de Librerias 10

Resumen de resultados 11

## Introducción

Un sistema de recomendación se compone de un uno o más algoritmos ”inteligentes” que combinados son capaces de ofrecer a sus usuarios una o varias listas de ítems o productos recomendados de forma personalizada.

Los sistemas de recomendación estudian las características de cada usuario y de cada item y como salida ofrecen un subconjunto de items que estima o predice que podrán ser de interesantes para el usuario para cumplir un objetivo de negocio que, normalmente, es la venta o alguna forma de consumo del item.

Existen varios tipos de sistemas de recomendación pero sólo comentaré brevemente los que interesan desde el punto de vista del presente TFM:

1. Sistemas de recomendación basados en la popularidad de los ítems que tienen valoraciones o ratings por parte de sus usuarios.
2. Sistema de recomendación de filtrado colaborativo: el ﬁltrado colaborativo consiste en ver qué usuarios son similares al usuario activo y, a continuación, recomendar aquellos items que no han sido votados por el usuario activo y que han resultado bien valorados por los usuarios similares.
3. Sistemas de recomendación híbridos realizan una “mezcla” con alguna metodología de las recomendaciones de salida de recomendadotes de distintos tipos intentando aprovechar los “puntos fuertes” de cada uno de ellos.

En este TFM, inicialmente se han realizado tareas de análisis y limpieza de datos, extracción de datos de página web (web scrapping) y actualización de datos obsoletos para poder obtener los datasets necesarios para los sistemas de recomendación.

Después, se han implementado tres sistemas de recomendación, dos de filtrado colaborativo y otro basado en popularidad.

Nuestro **sistema de recomendación basado en popularidad** tendrá tres tipos de recomendaciones de salida:

* **Most Rated:** 'Las cervezas que se han valorado mayor numero de veces'.
* **Positive Rated:** 'Las cervezas que tienen valoraciones positivas por encima de un valor' (el valor de valoración de las que se consideran “Buenas” o más).
* **mean\_rate\_beers:** 'La cerveza que tiene la valoracion mas alta de media' (solo se tendran en cuenta cervezas con mayor numero de valoraciones que un umbral fijado).

El primer **sistema de recomendación de filtrado colaborativo de Matriz de Co-ocurrencia**.¶

* Las recomendaciones se calculan teniendo en cuenta el grado de similitud que tienen las cervezas entre si.
* En concreto, la recomendación de una cerveza para un usuario concreto se calculará en dos pasos:
  + Se encontrarán las topN cervezas más similares a una dada como entrada.
  + Después, para un usuario dado y la lista de cervezas que ha valorado, seleccionaremos las topN cervezas de las listas de cervezas mas similares que correspondan a cada cerveza que ha valorado ese usuario.

El segundo **sistema de recomendación de filtrado colaborativo será tipo ALS (Alternating Least Squares) o SVD (Singular Value Decomposition)**:

En los modelos de Filtrado Colaborativo de Factorizacion de Matrices se intentan obtener los factores latentes (vectores) para poder modelizar las interacciones de los usuarios con los items que, a diferencia del la matriz de coocurrencia, los espacios latentes son de menor tamaño (alrededor de 100 dimensiones) y se pueden usar para hacer las recomendaciones.



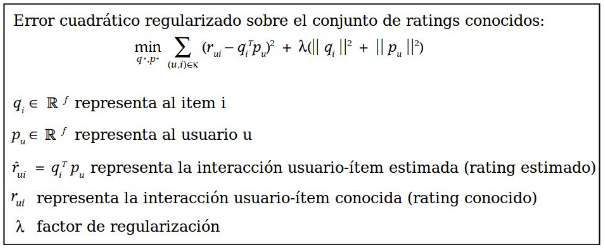
El **sistema de recomendación de filtrado colaborativo SVD (Singular Value Decomposition)**¶será mi opción alternativa si tuviera problemas con ALS.

* El objetivo es reducir la dimensionalidad del espacio de entrada: SVD es casi lo mismo que la Descomposicion Eigen o PCA (Principal Component Analysis).
* Calcularemos la matriz M (r\_uid) de valoraciones (user\_idx x beers\_ids) segÚn esta formula:



El **sistema de recomendación de filtrado colaborativo ALS (Alternating Least Squares)**¶se elige como primera opción porque tiene menos tendencia al sobreajuste (overfitting) de la que tiene SVD debido al problema de la dimensionalidad.

* Es un algoritmo de factorización matricial con regularización que evita el sobreajuste del algoritmo SVD (Singular Value Decomposition) al aumentar su número de dimensiones.
* ALS se usa para calcular las matrices q\_i y p\_u mediante factorización de la matriz de interacciones usuarios-ítems con pesos que si se multiplican da como resultado la matriz de valoraciones estimada r\_ui\_estimada.
* La funcion de perdida (loss) aplicada que se optimizara con ALS es la suma de errores al cuadrado:



* ALS se basa en el hecho de que la optimización se aborda como dos problemas cuadráticos obtenidos fijando, de forma alternativa, los términos p\_u y q\_i, en principio desconocidos. Cuando todos los términos p\_u han sido fijados, el sistema hace iteraciones para optimizar los términos q\_i por mínimos cuadrados y despues realizara iteraciones fijando los términos q\_i para optimizar los terminos p\_u [Koren, Y., Bell, R., Volinsky, C., 2009].

Todos los sistemas de recomendación han sido evaluados con estas métricas:

* **recall@n**.
* **map@n**.

Adicionalmente, se han evaluado las recomendaciones o predicciones de SVD con la medida **RMSE**.

## Datos de Entrada y Datos Calculados

El dataset de entrada inicial **'beer\_reviews.csv'** se ha obtenido en:

<https://query.data.world/s/gib6aa6n3tmtuvrqpbdlhcbgdg7tk4>

El dataset contiene valoraciones de cervezas realizadas hasta el año 2015 por los usuarios de la web **BeerAdvocates** mediante estos parámetros:

* **Appearance (Look) = review\_appearance:** APARIENCIA O IMPRESION VISUAL
* **Smell = review\_aroma:** OLORES
* **Taste = review\_taste:** SABOR
* **Mouthfeel (Feel) = review\_palate:** SENSACIONES EN BOCA Y PALADAR
* **Overall = review\_overall:** VALORACION GLOBAL

A continuación muestro cómo interpretar los rangos de valoración de cada parámetro de “review” basada en la guía de valoraciones de cervezas de BeerAdvocate: <https://www.beeradvocate.com/community/threads/how-to-review-a-beer.241156>

|  |
| --- |
| 4.50 - 5.00 = World-Class (La mejor del mundo)  4.00 - 4.49 = Outstanding (Excepcional)  3.75 - 3.99 = Very Good (Muy Buena)  3.50 - 3.74 = Good (Buena)  3.00 - 3.49 = Okay (Esta bien)  2.00 - 2.99 = Poor (Mala)  1.00 - 1.99 = Awful (Horrible) |

Por otro lado, como se verá después, la mayoría de las cervezas valoradas son cervezas artesanas por lo que el tipo de usuario que realiza valoraciones en BeerAdvocates es mayoritariamente un bebedor americano de cerveza de cervezas artesanas aunque también hayan hecho valoraciones de algunas cervezas de grandes cerveceras.

A partir de **'beer\_reviews.csv'** se han generado otros datasets a lo largo de los distintos procesos ejecutados en el pipeline a través del cual se han hecho ciertos análisis, limpieza de datos, calculo de nuevos campos y creación de los datasets de entrada finales para los recomendadores.

Los datos adicionales calculados que se han ido agregando y forman parte del fichero final de valoraciones que será la entrada de los recomendadores, **'beer\_reviews\_v3.csv'** . Estos campos son:

* **'abv\_strength':** Intensidad del Grado Alcohólico (**‘beer\_abv’** de **'beer\_reviews.csv'**), podría ser un dato para crear el perfil de usuario.

|  |
| --- |
| Nos hemos basado en:  <http://dev.bjcp.org/beer-styles/introduction-to-beer-styles>  Low Strength = intensidad-baja: <4% ABV  Medium Strength = intensidad-estandar: 4-6% ABV  High Strength = intensidad-alta: 6-9% ABV  Strong Strength = intensidad-muy-alta: >9% ABV |

* **'review\_average':** Media de todos los Parámetros de Valoración de una cerveza (nueva medida de valoración).
* **'beer\_advocates\_style':** es el campo que contiene el mapeo de los **'beer\_style'** de **'beer\_reviews.csv'** que NO coinciden con los de 2019 de la web de BeerAdvocates y hay que mapearlos a uno de ellos, que es el campo **'style\_name'** de **'beer\_styles\_v1.csv'**.
* **'beer\_style\_ibu\_avg':** el rango de la medida estandarizada del amargor IBU de cada estilo de cerveza se usa par calcular su media y asignarla a todas las cervezas de ese estilo ya que se carecía de este dato.
* **'beeradvocate\_score':** es la medida de puntuación global para el ranking de cervezas de la web BeerAdvocate basada en dar prioridad o más peso a las puntuaciones de review por este orden: review\_taste, review\_aroma, review\_palate, review\_appearance y review\_overall, para un usuario más generalista, con esta fórmula:

|  |
| --- |
| **('review\_taste' \* 0.40) + ('review\_aroma' \* 0.24) + ('review\_palate' \* 0.10) + ('review\_appearance' \* 0.06) + ('review\_overall' \* 0.20)** |

* **'aroma\_appearance\_score':** es una medida creada para usarla como rating o valoración de entrada para que un recomendador pueda recomendar a los usuarios que den más prioridad o peso a frente a y, así hacer una mejor personalización.

|  |
| --- |
| **('review\_aroma' \* 0.40) + ('review\_appearance'\* 0.24) + ('review\_taste' \* 0.10) + ('review\_palate' \* 0.06) + ('review\_overall' \* 0.20)** |

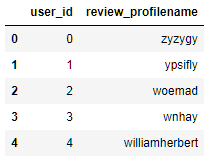
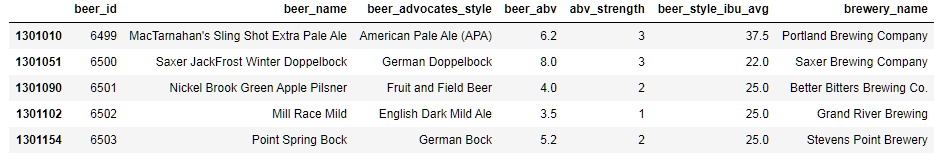
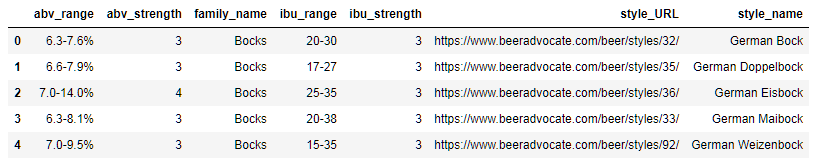
* **‘user\_id’:** identificador de usuario regenerado desde cero.
* **'beer\_id'**: identificador de cerveza regenerado desde cero.

Tanto **'beeradvocate\_score'** como **'aroma\_appearance\_score'**, se eligieron como resultado del análisis hecho en el notebook **“03-Beer\_Parameters\_Analysis.ipynb”**.

El dataset final se llama **“beer\_reviews\_v3.csv”** y esta es una muestra:



Por último listamos una muestra de los datasets adicionales, todos relacionados con , que se han creado con datos adicionales sobre estilos de cervezas **“beer\_styles\_v1.csv”**, usuarios **“users\_v1.csv”** y cervezas **“beers\_v1.csv”**.



Todos estos datasets se leerán y escribirán en el directorio **“./Data”** respecto a aquel en el cual se sitúen los notebooks que forman parte del Pipeline para su ejecución.

## Pipeline

A continuación muestro una tabla donde se ve el **Pipeline** con el que se han llegado a crear y modificar estos datasets junto a los creados en los pasos intermedios:

|  |  |  |
| --- | --- | --- |
| **ENTRADA** | **Notebook (Proceso Implementado “\*.ipynb”)** | **SALIDA** |
| **“beer\_reviews.csv”** | “01.01-Data\_Analysis” | “beer\_reviews\_v1.csv” |
| (1) URLs de web BeerAdvocate | “01.02.-Scraping-Additional-Beers-Data” | “beer\_styles\_v1.csv” |
| “beer\_reviews\_v1.csv” “beer\_styles\_v1.csv” | “01.03.-Mapping-Beer\_Styles” | “beer\_reviews\_v1\_1.csv” |
| “beer\_reviews\_v1\_1.csv” “beer\_styles\_v1.csv” | “02-Users\_And\_Beer\_And\_Beer\_Styles\_Ratings\_Analysis” | “beer\_styles\_v2.csv” |
| “beer\_reviews\_v2.csv” “beer\_styles\_v1.csv” | “03-Beer\_Parameters\_Analysis” | Formulas de Scoring: estudio, comparativa, .. |
| “beer\_reviews\_v2.csv” “beer\_styles\_v1.csv” | “04-Recommender-Input-Data-Creation” | “beer\_styles\_v3.csv” “users\_v1.csv” “beers\_v1.csv” |
| **“beer\_reviews\_v3.csv” “beer\_styles\_v1.csv”  “users\_v1.csv” “beers\_v1.csv”** | **“05-Recommenders-v1”** | **Evaluación y Recomendaciones de Popularidad,  Co-ocurrencia y SVD** |

(1) '<https://www.beeradvocate.com/beer/styles/'>,'<https://www.beeradvocate.com/beer/styles/32/'>, …

## Librerías

He usado Python para codificar los notebooks del Pipeline y además se han tenido que instalar e importar estas librearías de Python:

|  |
| --- |
| import pandas as pd  import numpy as np  import os  Librería beautifulsoup:  import requests  import bs4  import re  import matplotlib.pyplot as plt  import seaborn as sns  %matplotlib inline  import itertools  from sklearn.metrics import mean\_absolute\_error  from sklearn.model\_selection import cross\_val\_score  from sklearn.model\_selection import GridSearchCV  from sklearn.tree import DecisionTreeRegressor  from tabulate import tabulate  from warnings import catch\_warnings, simplefilter  from sklearn.model\_selection import KFold  from sklearn.model\_selection import ShuffleSplit  from math import sqrt |

## Resumen de Resultados

* **“02-Users\_And\_Beer\_And\_Beer\_Styles\_Ratings\_Analysis.ipynb”**.
  + El threshold de NUMERO DE VALORACIONES x CERVEZA elegido es 30.
  + El threshold de NUMERO DE VALORACIONES x USUARIO elegido es 30.
  + Aunque es una conclusión algo intuitiva, se confirma con este análisis que el estilo de la cerveza también influye en la Valoración o posible compra de una cerveza.
* **“03-Beer\_Parameters\_Analysis.ipynb”**..
  + **ABV Y ABV Strength:** se comprueba que la Intensidad de ABV y el ABV tienen un pequeño grado de correlacion con AROMA (28%), PALATE (24%), TASTE (24%), una correlacion que no es desdeñable y demuestra una influencia del ABV en esos tres parametros de valoración de una cerveza y, de forma indirecta en la valoracion OVERALL.
  + **IBU:** no es una buena medida de amargor pero el IBU Medio de cada estilo de cerveza parece que tiene un pequeño grado de correlación con AROMA (23%), APPEARANCE (21%), PALATE (20%), TASTE (19%), una correlación que no es desdeñable y demuestra cierta influencia en esos tres parámetros de una cerveza y, de forma indirecta en la valoración OVERALL..
  + **Análisis de posibles fórmulas de scoring o rating:** en base a los parámetros de valoración de las cervezas: para crear un ranking de cervezas que se usara en un Recomendador tendremos que usar al menos dos distintas para poder abarcar los dos tipos de bebedores principales.
    - BeerAdvocate Score para el bebedor general y el mas experimentado o cualificado para hacer valoraciones (fundadores de BeerAdvocate después de su estudio de negocio de varios años o un jurado de una competición) dan más importancia o peso al AROMA (Smell) frente al PALATE (Mouthfeel).
    - Aroma & Appearance Score Formula 2 para un bebedor de cerveza que diera mas prioridad a AROMA y APPEARANCE.
    - Opcionalmente, podríamos tener **Importance Score Formula** para el público general.
* **“05-Recommenders-v1.ipynb”.**
  + Los **recomendadores de filtrado colaborativo; Matriz de Co-ocurrencia y SVD** para calcular Matriz de Ratings, la lista de 10 primeras cervezas recomendadas estaba formada por aproximadamente el 80% de las 10 cervezas recomendadas por el recomendador de popularidad MostRated (además eran las mejor valoradas) y, además, se incluían 1 o 2 que no formaban parte de ese grupo.
  + Las medidas **recall@n y map@n** creo que no era mala pero eran mejores para el **recomendador de popularidad MostRated** y no había tanta diferencia de valor respecto a los otros dos (creo que el efecto de las cervezas más y mejor valoradas se ha notado aquí).
  + **ALS y SVD en su versión “para modelizar la matriz de preferencias, o sea, si una cerveza le gusta o no a un usuario”**, no se pudieron terminar de implementar porque durante su entrenamiento el kernel de Jupyter-Python lanzaba un **error “Memory Error”** que lo impedía: creo que es un problema de dimensionalidad que impedía la creación de las matrices de trabajo ya que tengo 30.810 user\_ids y 6.504 beer\_ids y, creo que la cuestión es que se crea una **matriz de ratings de user\_ids x beer\_ids = 200.388.240**.