# **SPRINT 3 - MEETING 38** Sistemas de Recomendación (Parte 2)

☐ Grabación **Presentación** 

> Contenido extra Grabación

Sistemas de recomendación 2.0

#### "We're competing with sleep, on the margin. And so, it's a very large pool of time." -Reed Hastings - CEO Netflix.

¡Bienvenido/a a la segunda parte de sistemas de recomendación! En esta

El equipo docente subió contenido extra de la meeting 💠

bitácora profundizaremos en algunos aspectos técnicos de los sistemas colaborativos y haremos un repaso por algunas discusiones y debates acerca de los sistemas de recomendación. ¿Recuerdas lo que aprendiste sobre las empresas pioneras en los 90? Todas

daban sus servicios a través de la web, lo que les permitía -a diferencia de los locales físicos- ofrecer miles de opciones a los usuarios y usuarias. Utilizando sistemas de recomendación, los y las ayudaban en la elección. Podríamos pensar que mientras más opciones le ofreces a la persona consumidora, más probabilidad hay de que concrete una compra, ya que hay más probabilidades de que encuentre el producto que desea. Pero hay una

paradoja: cuantas más opciones de un producto tiene un consumidor, menos probabilidades hay de que lo compre. En el año 2000, los psicólogos Sheena Iyengar y Mark Lepper realizaron un estudio.pdf) sobre el consumo de las personas. Consistía en ofrecer variedades de mermeladas. La mitad de los días ofrecieron un estante con 24 sabores de mermeladas y otros días con 6. Los resultados fueron concluyentes: los días en

los que había 6 variedades, las compras se multiplicaron por 10 en comparación

Los sistemas de recomendación son grandes candidatos a la hora de reducir las opciones. Y aunque el boom y la mayoría de sus implementaciones son en la web: ¿se te ocurren otros usos? Preprocesamiento de datos

Como mencionamos en la bitácora anterior, hay principalmente dos medidas

Coeficiente Jaccard y la Similitud Coseno. Cada uno tiene su particularidad

que podemos utilizar para medir la similitud de ítems y usuarios: el

### y, como de costumbre, la elección va a depender del problema. El tipo de procesamiento de datos va a depender, en parte, de la medida que utilicemos.

a los otros días. Otros estudios confirman la tendencia.

Es importante que prestes atención a lo siguiente. Cuando una persona usuaria no califica un ítem, en realidad en esa columna hay un valor faltante (de cierta manera, los sistemas de recomendación son una forma de resolver un problema de valores faltantes). Si en lugar del valor faltante ponemos un o, esto podría interpretarse como que calificó negativamente ese ítem, cuando en realidad no es así. Por eso, hay que tener cuidado.

Existen dos tipos de procesamiento recurrentes: • Redondear los datos: el Coeficiente Jaccard pierde información importante cuando trabajamos con números continuos. En estos casos, podemos redondear los atributos con números continuos y pasarlos a booleanos. En el ejemplo de las películas, podríamos colocar un 1 en las que fueron rankeadas con 4 y 5, indicando que ese tipo de películas le

• Normalizar ratings: la Similitud Coseno es una mejor medida para

trabajar con atributos continuos. El riesgo es que el sistema interprete que

una película que el usuario o usuaria calificó con 5, está más cerca a una

película calificada con 2 que a una película sin calificar. Si normalizamos

los ratings restándoles el promedio de los ratings dados por el usuario o

usuaria, vamos a tener en negativo las películas que menos le gustaron al

usuario y en positivo las que más le gustaron. Los datos estarán escalados

gustan al/a usuario/a, y un o en las que calificó con 1 ó 2. El 3 suele ser un caso a debatir y algunas personas eligen descartarlo.

y centrados en o.

producto de dos matrices:

m: ítems

comprender el paso siguiente.

4 x 3

4: usuarios

3: ítems

optimización.

**Descomposición UV** Existe una técnica con muy buenos resultados que se parece a las herramientas de **reducción de dimensionalidad**, y en particular, a **SVD**. Si bien los detalles matemáticos pueden ser un poco difíciles, te comentaremos las ideas principales.

La idea fundamental es que podemos aproximar la matriz de utilidad como el

 $u_{n1} u_{n2}$ n x m

n x 2

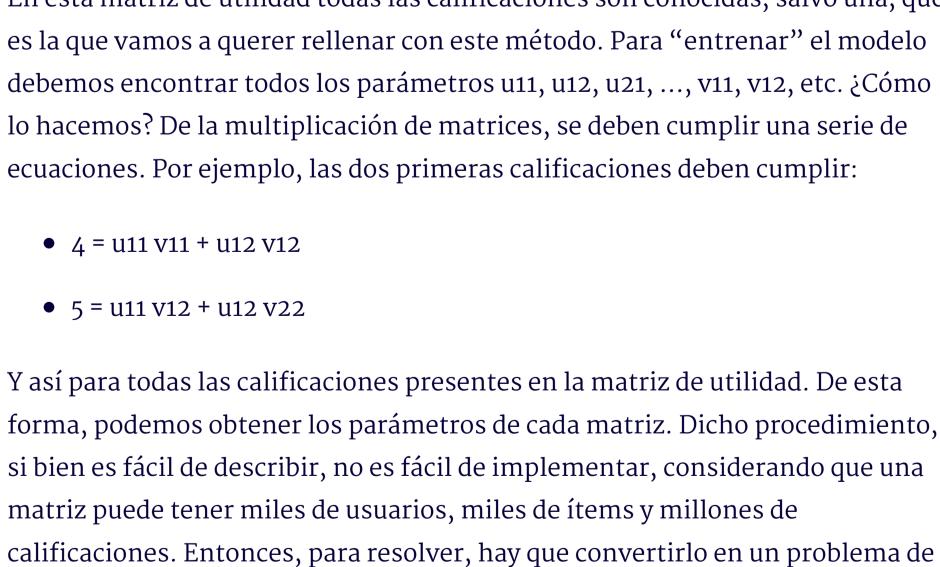
**Para investigar:** si no estás familiarizado/a con el producto (multiplicación)

de matrices, googlea cómo se hace. Es muy sencillo, y lo necesitarás para

Supongamos que originalmente nuestra sistema era el siguiente:

4 x 2 2 x 3

En esta matriz de utilidad todas las calificaciones son conocidas, salvo una, que



de optimización? 1. En primer lugar, necesitamos una función de costo. En este caso, la métrica será el error RMSE entre la verdadera calificación (los elementos de la matriz de utilidad - 4 y 5 en los dos ejemplos de arriba) y las calificaciones predichas (el producto de U y V • u11 v11 + u12 v12 y u11 v12 + u12 v22 en los ejemplos de arriba).

2. Debemos encontrar los parámetros asiados al mínimo de ese costo. Una

gradiente! Arrancamos en un lugar al azar y vamos mejorando

Este proceso es complejo y contiene muchos detalles que no mencionamos.

iterativamente cada parámetro.

búsqueda por grilla no funcionará en este caso, ya que tenemos miles de

parámetros (cada uij y vij). Entonces, ¿qué podemos hacer? ¡Descenso por

¿Recuerdas del Sprint 2 qué cosas necesitamos para convertirlo en un problema

¡También involucra regularización! Presta atención al amplio rango de aplicaciones que tienen las herramientas que te contamos en el segundo Sprint. Esta técnica no está incorporada en Scikit-Learn, sino que debemos usar una nueva librería. Nosotros elegimos Surprise, pero no es la única opción.

¿Y cómo predecimos? Una vez encontrados los parámetros, los podemos usar

para predecir. La idea es aproximar los valores faltantes de la matriz de utilidad

por el valor obtenido en producto de U y V. En el ejemplo que dejamos más arriba, el valor faltante lo podemos aproximar por u31 v12 + u32 v22. Evaluación de Sistemas de Recomendación Aún no hemos dicho nada sobre evaluación de sistemas de recomendación, un aspecto fundamental de cualquier desarrollo. Algunas métricas que podemos

utilizar si estamos prediciendo calificaciones son RMSE, MSE, MAE y cualquiera

calificaciones altas que calificaciones bajas. Pero, en realidad, nos interesa

predecir calificaciones altas, ¡las bajas no nos importan! Es decir, preferimos un

usuario/a en particular) que aquel que predice 2.5 estrellas para una película de

atento/a a esto, ya que según nuestro problema podemos definir una función de

1 estrella, si bien el error es más grande en ese caso. Entonces, conviene estar

evaluación que tenga en cuenta estos detalles.

ítem sin calificaciones.

recomendación

Sesgo de confirmación

propias creencias.

de contenido.

que sirva para problemas de regresión. Sin embargo, hay que tener cuidado:

métricas como RMSE pueden penalizar modelos que predicen mejor

modelo que predice 4.5 estrellas para una película de 5 estrellas (según un/a

## Otras métricas muy utilizadas son aciertos en top-5, top-10, etc. Es decir, cuántos ítems calificados positivamente por una persona usuaria aparecen en ese ranking. Esta es una métrica cómoda, interpretable y útil.

¿Y el sobreajuste y subajuste no son más un problema? Lamentablemente,

cruzada y train-test split. En este caso, una de las formas más comunes de

Límites éticos en los sistemas de

siempre corremos ese riesgo, por lo que deberemos seguir haciendo validación

hacer una separación en entrenamiento y evaluación es sacando calificaciones

al azar de la matriz de utilidad, teniendo cuidado de no dejar a ningún usuario ni

juntos el impacto que tienen en lo que pensamos y en la sociedad. Como sabés, hoy en día podemos realizamos prácticamente todos nuestros consumos a través de la web: • Informarnos: portales de noticias, podcast, twitter. • Sociales: redes sociales, apps para conseguir pareja. • Entretenimiento: videojuegos, deporte, pornografía. • Compras: artículos de limpieza, comida, electrónica, muebles.

En fin, casi todo lo que te puedas imaginar. Todas estas plataformas cuentan

En la bitácora 30, sobre métricas en clustering, incluimos la siguiente frase:

similar pasa con la información que consumimos para reafirmar nuestras

"Cuando una métrica se convierte en objetivo, deja de ser una métrica". Algo

consumos. Algunas personas están alertando sobre los efectos en los individuos

con sistemas de recomendación que, en definitiva, determinan muchos

y en la sociedad. ¡Repasemos algunas de las principales discusiones!

• **Religión**: ya son tendencia los influencers espirituales.

Ahora que aprendiste cómo funciona un sistema de recomendación, pensemos

"El sesgo de confirmación es la tendencia a favorecer, buscar, interpretar, y recordar, la información que confirma las propias creencias o hipótesis, dando desproporcionadamente menos consideración a posibles alternativas. (...) Las personas muestran esta tendencia cuando reúnen o recuerdan información de

manera selectiva, o cuando la interpretan sesgadamente."

están constantemente seleccionando la información por ti. Y como

Sesgo de confirmación según Wikipedia:

que mencionaremos a continuación.

linkeado con este tema.

Imagina dos noticias:

radicalización ideológica.

**Adicción** 

Por las dudas, acá te dejamos un artículo (en inglés) que plantea por qué no podemos culpar a las redes sociales por el aumento de la polarización. Fake news y radicalización Probablemente hayas escuchado sobre las fake news; un fenómeno creciente en el último tiempo. Este trata sobre la masividad que tomaron las noticias falsas

Pero antes de continuar. ¿Es justo atribuirle estos fenómenos a los sistemas de

en los últimos años. Principalmente a través de las redes sociales.

recomendación? ¿Juegan un rol tan importante en la sociedad?. No,

definitivamente no es esa nuestra intención. No estamos juzgando ni

atribuyendo responsabilidades. Pero es importante construir una base de

conocimiento sobre la principales discusiones alrededor de estas tecnologías,

por más que probablemente ya las hayas escuchado. Tal vez, nunca lo habías

"Información de manera selectiva": ¿te suena? Los sistemas de recomendación

generalmente leemos o le damos like al contenido que afirma lo que pensamos,

esto resulta en que algunas plataformas tiendan a mostrarte solamente ese tipo

Algunas personas y organizaciones argumentan que este fenómeno está

causando mayor polarización y conflicto en las sociedades. Desde ya, estas

discusiones están abiertas, son multicausales y se entrelazan con los puntos

• "Un equipo de investigación comienza la fase dos en una vacuna para tratar el HIV." • "Se encuentra la cura para el HIV" ¿Dónde harías click?

Los sistemas de recomendación no evalúan, en principio, la veracidad del

contenido, sino la cantidad de interacciones u otras variables. Las noticias

sensacionalistas suelen tener más de esto. De la misma forma, las noticias

radicalizadas, ya sea por gente que está a favor, como por gente que está en

Acá te dejamos otro artículo que analiza los sistemas de recomendación y la

Como aprendiste, los sistemas de recomendación pueden estar diseñados para

utilizar los intereses de los/as usuarios/as en maximizar su tiempo en pantalla.

El tweet de Netflix es una discusión que aplica a toda la población, pero causa

especial preocupación con los niños y jóvenes que pasan cada vez más tiempo

Finalmente, las técnicas que vimos de sistemas de recomendación pueden ser

aplicadas en otros contextos, a primera vista diferentes. Por ejemplo, pueden

ser utilizadas para encontrar asociaciones entre genes y enfermedades, como

(i)

Cuanto mayor tiempo pasas en la web, más ganancia para la empresa.

Sleep is my greatest enemy. 7:32 PM · Apr 17, 2017

○ 6.1K 
○ 3K people are Tweeting about this

contra. Cuanto más extremo, más suele llamar la atención.

#### hacen en este trabajo y muchos otros. Por eso, es una herramienta muy útil para tener en cuenta, ya que tal vez nuestro objetivo no sea crear un sistema de recomendación, pero lo podemos utilizar para responder una pregunta que nos interese.

Para terminar

en internet.

**Otros usos** 

Como sabes, nuestra intención es siempre abrir la discusión y dar a conocer estos debates. Queda de tu lado profundizar en los temas que elijas, tanto en aspectos técnicos como éticos y sociales. El próximo encuentro empezarás a aprender Procesamiento de Lenguaje Natural, una técnica que te permitirá analizar y trabajar con el lenguaje humano.

Para el encuentro, trae ya hecha tal parte, y completa durante la clase esto...

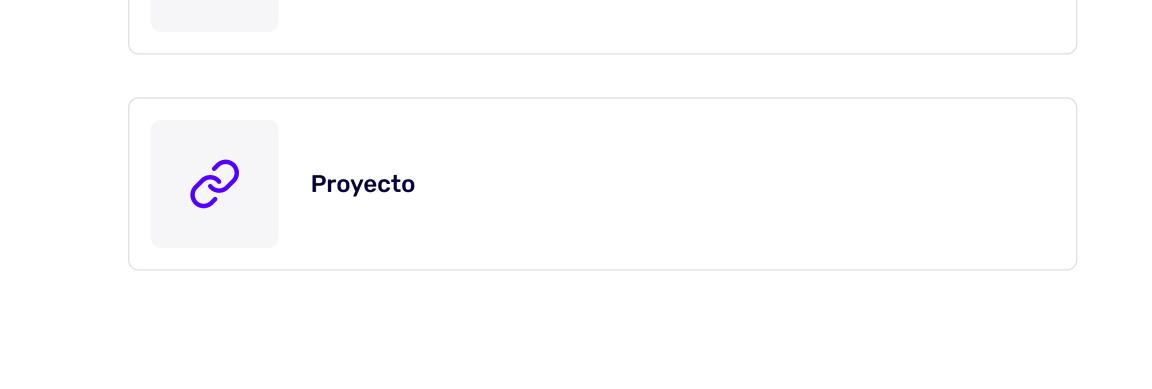
# ¡Prepárate para el próximo encuentro!

**Profundiza** 

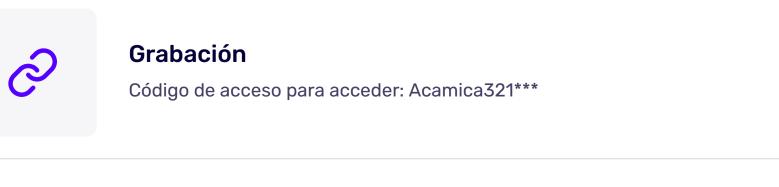
Contenido extra de la meeting

DS\_Bitácora\_38\_SR\_SVD

Empieza a trabajar en el notebook



Te invitamos a conocer más sobre el tema de esta bitácora.



Sistemas de Recomendación (Parte 1)

MEETING 39

Procesamiento Lenguaje Natural (Par...