



PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE
ESCUELA DE INGENIERÍA
DEPARTAMENTO DE CIENCIA DE LA COMPUTACIÓN

Tarea N°1:
IIC3633 Sistemas Recomendadores
(2020-2)

Nombre: Guillermo Echagüe A.

Profesor: Dennis Parra

Fecha: 21/09/20

Contenido

Introducción	1
Objetivo	1
Desarrollo.....	1
1. Exploración de datos.....	1
2. Recomendación basada en interacciones	6
2.1. Actividad 1: ALS con gradiente conjugado	6
2.2. Actividad 2: Bayesian Personalized Ranking (BPR).....	8
2.3. Comparación de resultados	10
3. Recomendación basada en contenido	11
3.1. Actividad 1: Recomendación basada en contenido	11
Conclusiones	18
Anexos	19

Introducción

El desarrollo de esta actividad del curso de Sistemas Recomendadores, consiste que por medio de datos obtenidos de un dataset de Printeres, y el entrenamiento de una red profunda modelo resnet, por medio de sus embeddings, lograr implementar diversos modelos de sistema recomendador, como lo es ASL, BPR y basado en contenido visual.

Este trabajo, intenta poder aplicar y aterrizar diversos modelos, y métricas, algunos vistos en clases y otros leídos desde diversas publicaciones atingentes a los contenidos del curso.

Objetivo

En esta tarea se trata de poner en práctica los conocimientos sobre Sistemas Recomendadores. En particular, experimentaran con recomendación basada en feedback implícito y basada en contenido. Se recomienda utilizar las librerías que serán pyreclab, desarrollada por Gabriel Sepúlveda. Se puede experimentar con otras bibliotecas en algunos casos, por ejemplo, implicit para el algoritmo BPR.

Es obligatorio agregar un archivo README.md al repositorio de la tarea que permita entender la estructura de archivos y detalles necesarios para replicar los experimentos realizados.

Desarrollo

1. Exploración de datos

El siguiente análisis se realizó con el dataset de training.

- Grafique la distribución de usuarios con número de interacciones, identifique los 5 usuarios más activos en el dataset de imágenes. Comente la forma de la distribución y qué porcentaje de las interacciones han sido hechas por estos 5 usuarios.

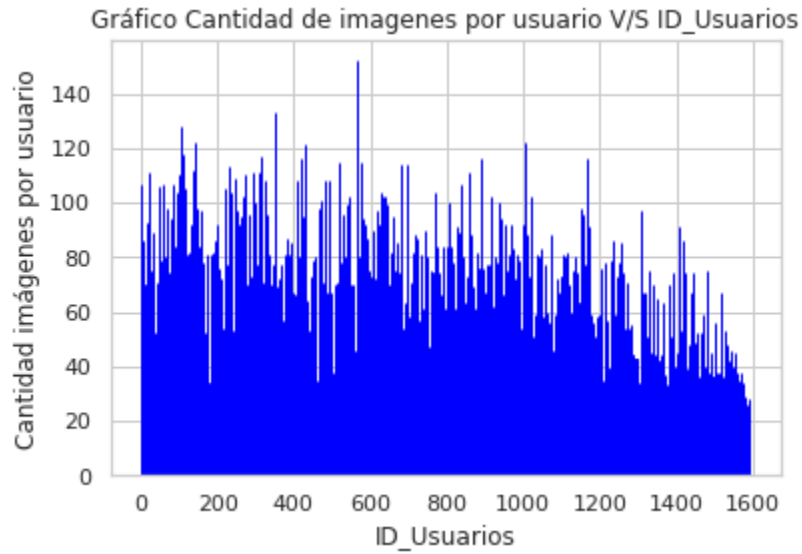


Grafico con la cantidad de interacciones de los usuarios

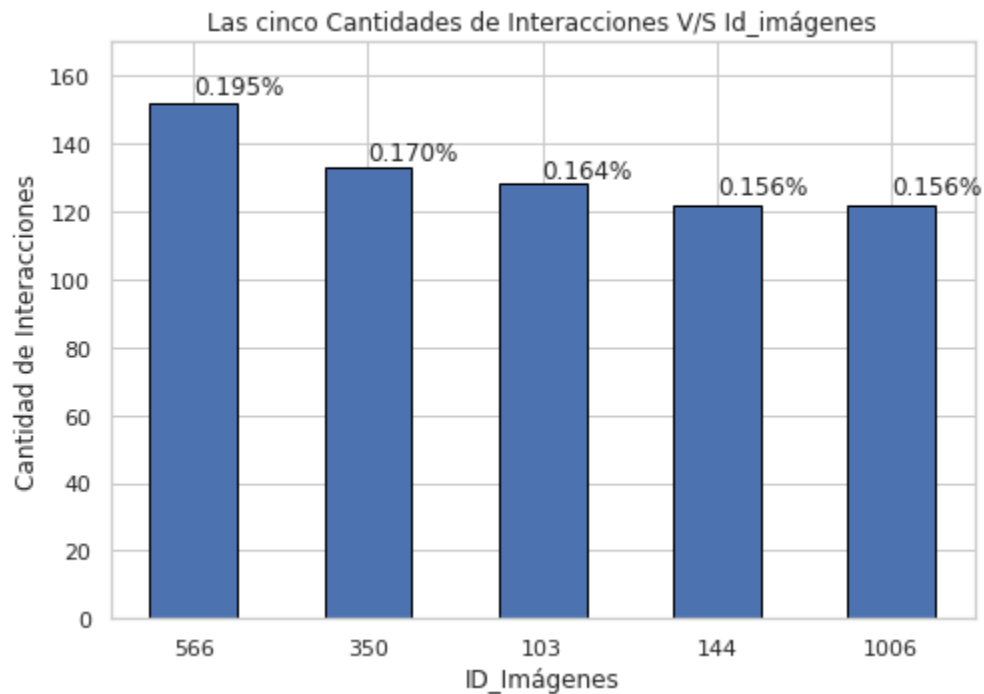
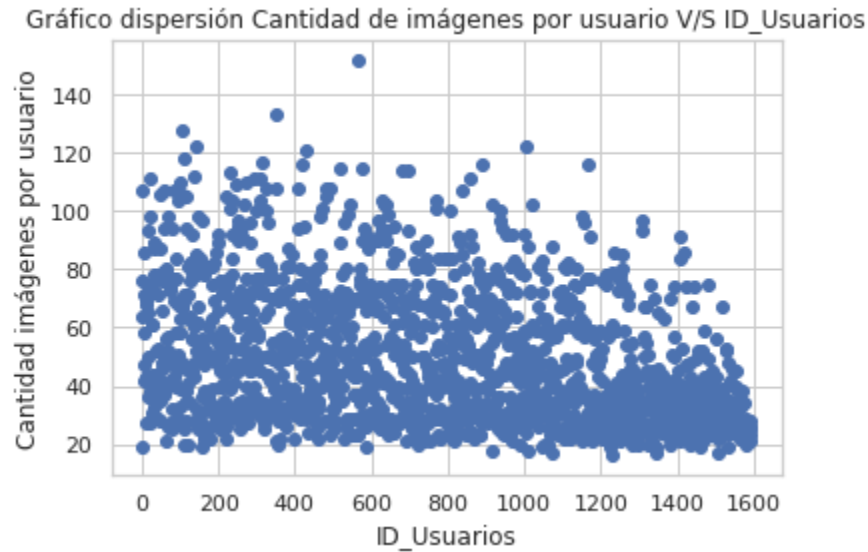


Grafico con los 5 usuarios que más interactuaron

ID_Usuario	566	350	103	144	1006
Cantidad	152	133	128	122	122
Porcentaje	0.195%	0.170%	0.164%	0.156%	0.156%

Tabla con los porcentajes de los 5 usuarios que más interactuaron



Como se observa en el gráfico de dispersión, la mayor cantidad de usuarios está concentrado entre los 20 a 60 imágenes, por lo que no es un comportamiento que permita obtener buenos resultados en las métricas utilizadas.

- Grafique la distribución de imágenes por número de interacciones, Identifique las 5 imágenes que han sido más vistas en el dataset de imágenes. Comente la forma de la distribución y qué porcentaje de las interacciones han sido sobre estas 5 imágenes.

En esta sección, como el tamaño del numero de ID de imágenes es muy grande, se le asignara un valor numérico entre 0 y 71026.

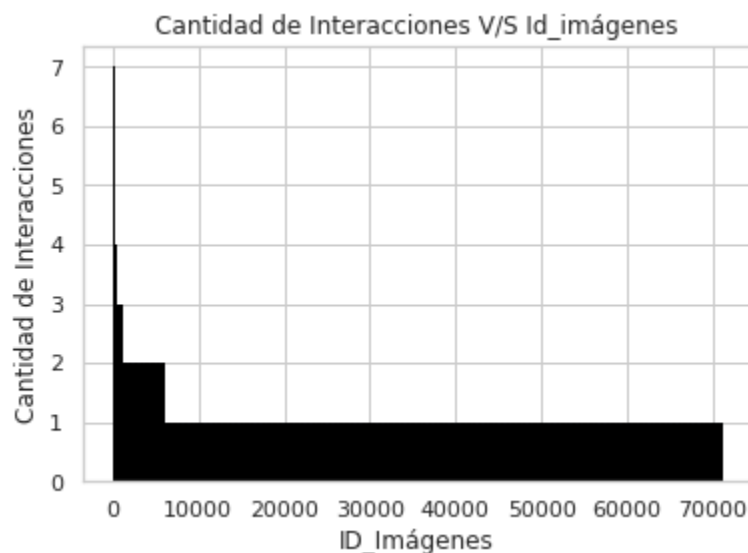


Gráfico con la cantidad de interacciones de las imágenes

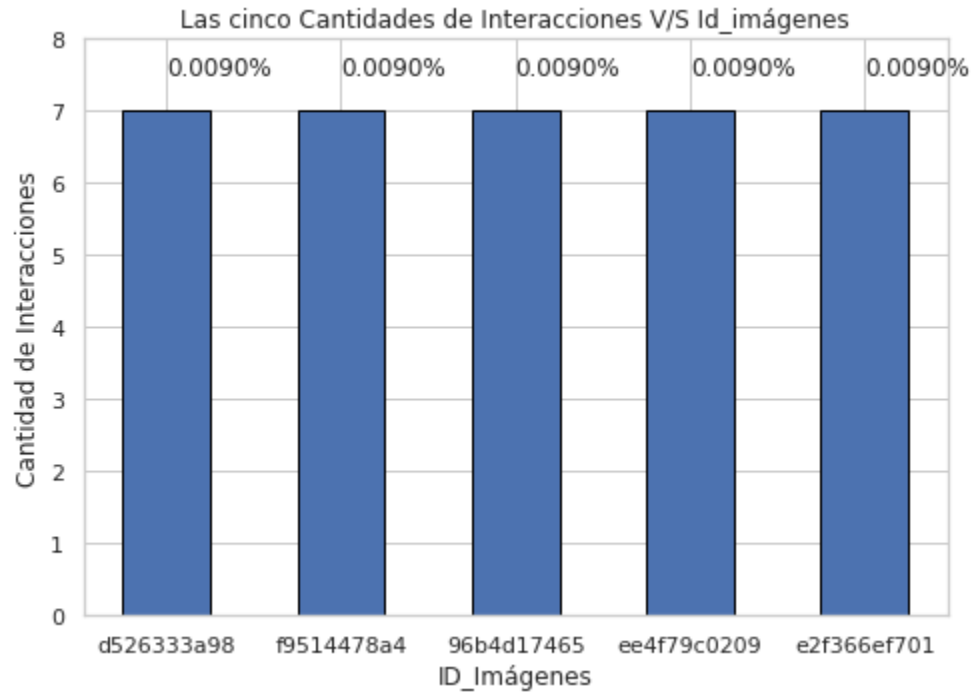
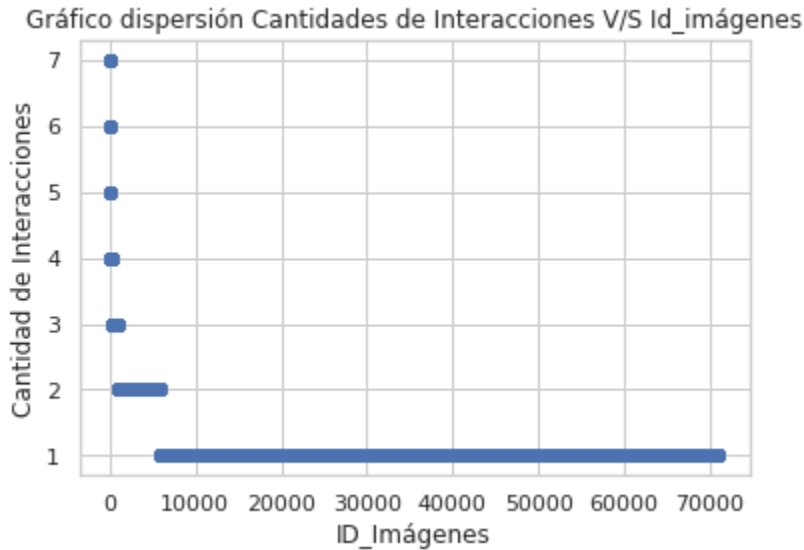


Gráfico con las 5 imágenes que más interactuaron

ID_Imágenes	Cantidad	Porcentaje
d526333a98d58328d47265a4e067323f	7	0.0090%
f9514478a48522018a1fe464c4baf8a7	7	0.0090%
96b4d174650edf3f1bdf0bfa12d75dd3	7	0.0090%
ee4f79c0209898b80839534b97c36eaf	7	0.0090%
e2f366ef70185c50be8df2e1ccd3fc56	7	0.0090%

Tabla con las 5 imágenes más solicitadas



El gráfico de dispersión en este caso muestra que los usuarios solo hicieron interacción con una sola imagen por lo que esto nos indica que en este dataset, al realizar recomendaciones no se obtendrán buenos resultados de métricas.

- Tabla con número de usuarios distintos, número de items distintos, promedio y desviación estándar de imágenes por usuario, promedio y desviación estándar de usuarios por imagen, y densidad del dataset (o sparsity) en cuanto a interacciones.

imágenes por usuario		Interacciones por imagen	
Media	48.9786566	Media	1.09849775
Error típico	0.40522789	Error típico	0.00137407
Mediana	43	Mediana	1
Moda	26	Moda	1
Desviación estándar	22.8729516	Desviación estándar	0.36620288
Varianza de la muestra	523.171915	Varianza de la muestra	0.13410455
Coeficiente de asimetría	0.93953345	Coeficiente de asimetría	4.92254013
Rango	136	Rango	6
Mínimo	16	Mínimo	1
Máximo	152	Máximo	7
Suma	156046	Suma	78023
Sesgo	1.004621399	Sesgo	0.268970459
Cantidad de usuarios	1593	Cantidad de Imágenes	71027

Tabla con los estadísticos del dataset de entrenamiento

Según los valores obtenidos anteriormente, se puede apreciar que los datos de interacción imágenes, estos son muy concentrados en poca cantidad, ya que tienen una moda y media 1.

Esto quiere decir que no se repite mucho la interacción con imágenes distintas.

2. Recomendación basada en interacciones

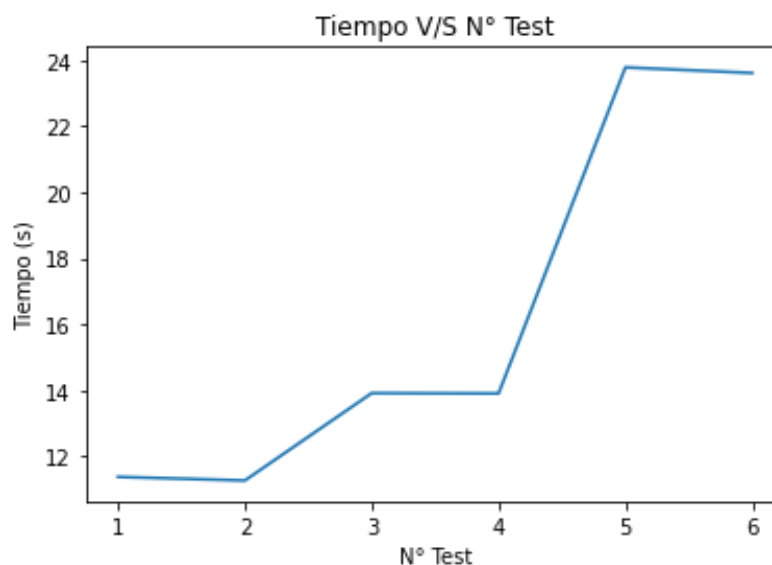
Muestre análisis de sensibilidad de resultados para métricas MAP@20 y nDCG@20 en función de tiempos, hiperparámetros, learning rate (0.001, 0.01), factores latentes (50, 100, 200) y regularización (0.01, 0.1). Grafique cada uno y comente.

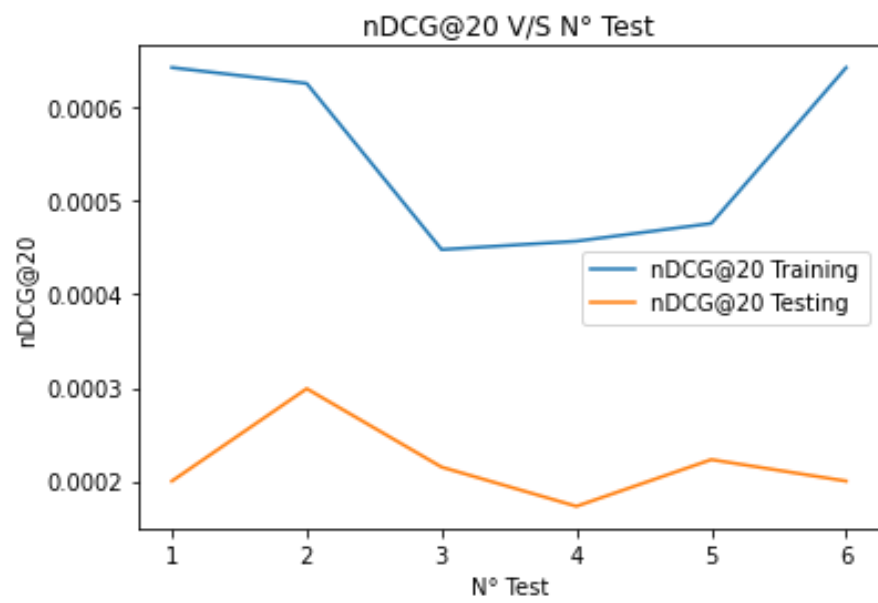
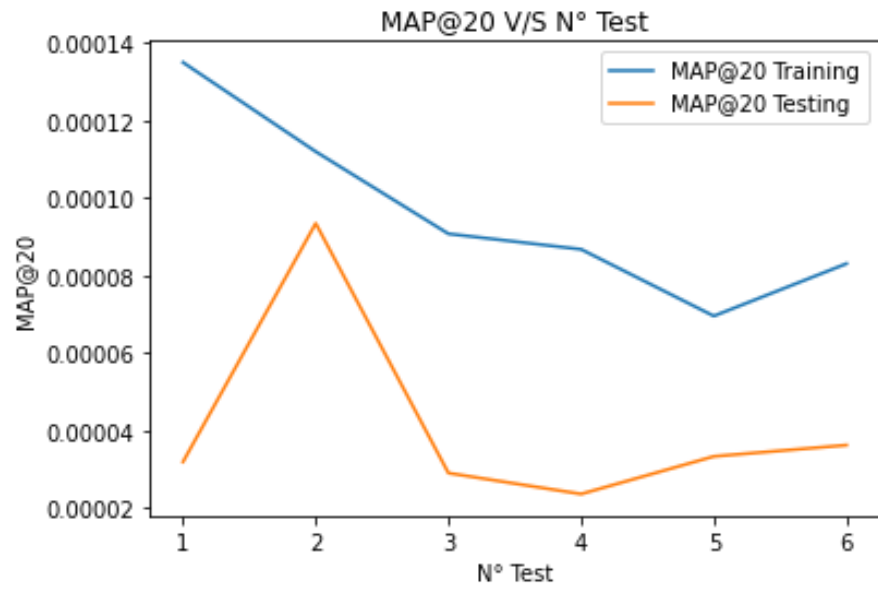
2.1. Actividad 1: ALS con gradiente conjugado

Muestre análisis de sensibilidad de resultados para métricas MAP@20 y nDCG@20 en función de tiempos, hiperparámetros, learning rate (0.001, 0.01), factores latentes (50, 100, 200) y regularización (0.01, 0.1). Grafique cada uno y comente.

Experimentos	ALS1	ALS2	ALS3	ALS4	ALS5	ALS6
Factor Latente	50	50	100	100	200	200
Regularizador	0.1	0.01	0.1	0.01	0.1	0.01
Tiempo	11.37	11.26	13.91	13.90	23.79	23.62
MAP@20 Training	1.35E-04	1.12E-04	9.08E-05	8.68E-05	6.96E-05	8.31E-05
MAP@20 Testing	3.20E-05	9.35E-05	2.91E-05	2.37E-05	3.34E-05	3.63E-05
nDCG@20 Training	6.43E-04	6.26E-04	4.48E-04	4.57E-04	4.76E-04	6.43E-04
nDCG@20 Testing	2.00E-04	2.99E-04	2.15E-04	1.73E-04	2.23E-04	2.00E-04

Los valores de color rojos en la tabla significan que son los máximos valores en cada fila y el color azul son los mínimos obtenidos.



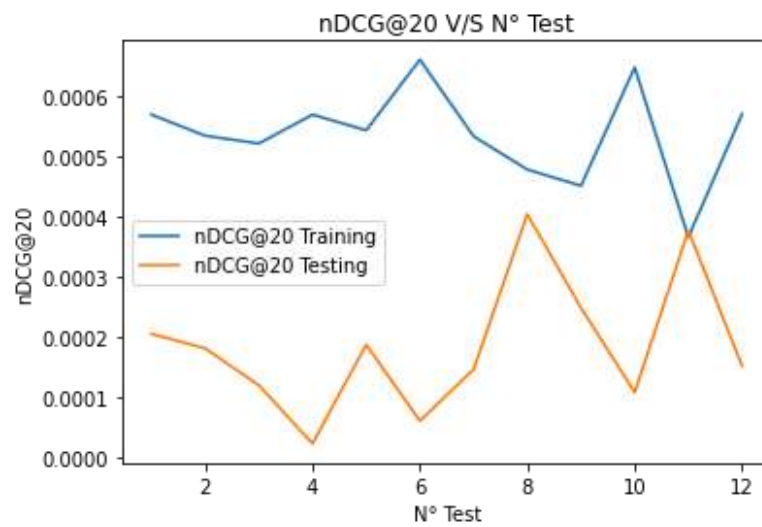
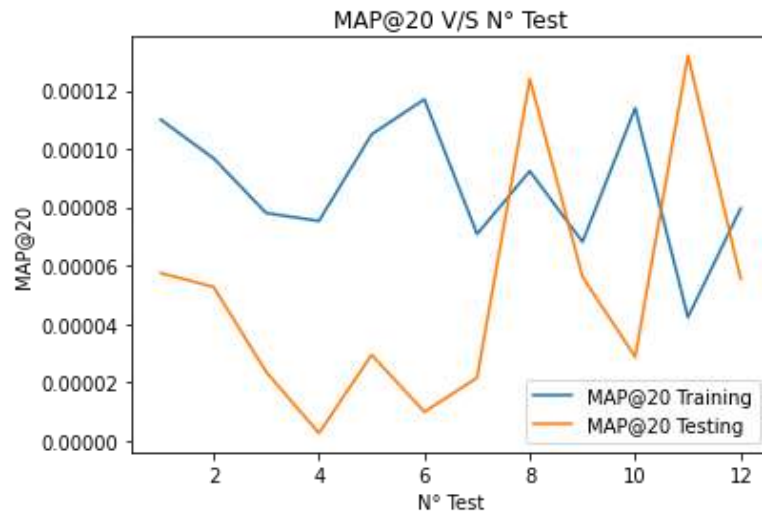
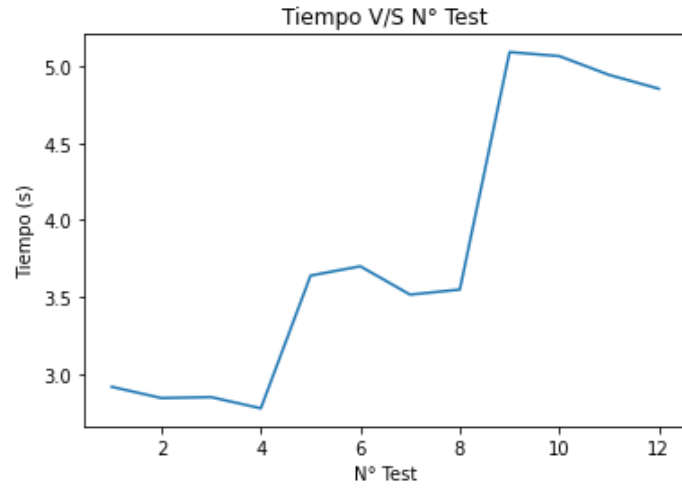


2.2. Actividad 2: Bayesian Personalized Ranking (BPR)

- Haga una tabla comparativa de las métricas para los 2 métodos utilizados en el mejor de los casos luego de modificar hiperparámetros. Identifique claramente los valores de los hiperparámetros para cada método.

Experimentos	BPR1	BPR2	BPR3	BPR4	BPR5	BPR6
Factor Latente	50	50	50	50	100	100
Regularizador	0.01	0.01	0.1	0.1	0.01	0.01
Learning rate	0.001	0.01	0.001	0.01	0.001	0.01
Tiempo	2.916	2.843	2.849	2.776	3.639	3.699
MAP@20 Training	1.10E-04	9.68E-05	7.80E-05	7.53E-05	1.05E-04	1.17E-04
MAP@20 Testing	5.74E-05	5.27E-05	2.35E-05	2.54E-06	2.94E-05	9.83E-06
nDCG@20 Training	5.69E-04	5.34E-04	5.21E-04	5.69E-04	5.43E-04	6.60E-04
nDCG@20 Testing	2.06E-04	1.82E-04	1.21E-04	2.42E-05	1.88E-04	6.18E-05
Experimentos	BPR7	BPR8	BPR9	BPR10	BPR11	BPR12
Factor Latente	100	100	200	200	200	200
Regularizador	0.1	0.1	0.01	0.01	0.1	0.1
Learning rate	0.001	0.01	0.001	0.01	0.001	0.01
Tiempo	3.515	3.548	5.091	5.065	4.943	4.853
MAP@20 Training	7.08E-05	9.24E-05	6.81E-05	1.14E-04	4.23E-05	7.96E-05
MAP@20 Testing	2.15E-05	1.24E-04	5.61E-05	2.87E-05	1.32E-04	5.55E-05
nDCG@20 Training	5.33E-04	4.78E-04	4.51E-04	6.47E-04	3.66E-04	5.70E-04
nDCG@20 Testing	1.47E-04	4.04E-04	2.49E-04	1.09E-04	3.77E-04	1.53E-04

Los valores de color rojos en la tabla significan que son los máximos valores en cada fila y el color azul son los mínimos obtenidos.



2.3. Comparación de resultados

Hiperparámetros	ALS	BPR
Factor Latente	50	200
Learning rate		0.1
Regularizador	0.01	0.001
Resultados	ALS	BPR
Tiempo	23.8158641	4.19167352
MAP@20 Training	1.23E-04	6.61E-05
MAP@20 Testing	2.61E-05	9.18E-05
nDCG@20 Training	6.60E-04	4.79E-04
nDCG@20 Testing	1.60E-04	3.69E-04

Se eligieron los hiperparámetros que entregaban los mejores valores de métricas tanto para MAP@20 y nDCG@20, aunque no son los mejores en cada categoría, se eligieron los que entregaban mejores valores en su conjunto.

3. Recomendación basada en contenido

3.1. Actividad 1: Recomendación basada en contenido

En esta actividad el objetivo es obtener una representación vectorial de imágenes con las que interactuó el usuario y recomendar basándose en métricas de similaridad de éstas con otras imágenes.

Antes de recomendar, se sugiere reducir la dimensionalidad de los embeddings de imágenes. Normalice los vectores (estandarización z-score) y luego aplique PCA, generando embeddings de tamaño 20 y 50. Compare los resultados de MAP@20 y nDCG@20 para ambas dimensiones.

• Z-Score

- Es una fórmula que permite la estandarización de una distribución.
- Esta fórmula se calcula restando al dato, la media de la distribución y dividiendo el resultado por la desviación típica (la distancia que tiene dicho dato respecto de la media).
- El resultado que arroja esta fórmula, permite conocer cuánto de lejos está un dato concreto respecto de la media.
- Es una técnica que sirve para normalizar los datos lo que nos permite comparar conjuntos de datos distintos, bien porque esté en unidades diferentes, bien porque se trate de volúmenes muy dispares.

- $\text{normalización} = (x - \text{media}) / \text{desviación típica}$

```
[57] def normalize_zscore_inplace(feamat):  
    means = feamat.mean(0)  
    stds = feamat.std(0)  
    for i in range(stds.shape[0]):  
        if stds[i] == 0:  
            stds[i] = 1  
        feamat -= means  
        feamat /= stds
```

```
[58] index2id=tabla_id
```

```
[59] n_artworks = len(index2id)  
    #n_artworks=1593  
    resnet50_dim = resnet50_feamat.shape[1]
```

```
    resnet_prueba_feamat = np.empty(shape=(n_artworks, resnet50_dim ))  
    for i in range(n_artworks):  
        resnet_prueba_feamat[i][:resnet50_dim] = resnet50_feamat[i]
```

Aplicación de la normalización Z-score a los embedding

sklearn.decomposition.PCA

- Principal component analysis (PCA)
- **class** sklearn.decomposition.PCA(n_components=None, *, copy=True, whiten=False, svd_solver='auto', tol=0.0, iterated_power='auto', random_state=None)
- Generando embeddings de tamaño 20 y 50.
- Compare los resultados de MAP@20 y nDCG@20 para ambas dimensiones.

<https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html>

- Aplique PCA, generando embeddings de tamaño 20 y 50.

```
[15] # Project into a 20 PCA feature space
pca20_resnet_featmat = PCA(n_components=20).fit_transform(resnet_prueba_featmat)

print("Cantidad de features despues de PCA: ", pca20_resnet_featmat.shape[1])
```

☞ Cantidad de features despues de PCA: 20

```
[16] # Project into a 50 PCA feature space
pca50_resnet_featmat = PCA(n_components=50).fit_transform(resnet_prueba_featmat)

print("Cantidad de features despues de PCA: ", pca50_resnet_featmat.shape[1])
```

☞ Cantidad de features despues de PCA: 50

Aplicación de PCA para generar embeddings de tamaño 20 y 50

Para recomendar utilice el algoritmo de recomendación de imágenes basada en contenido descrito en la ecuación (6) de este paper. Compare las 3 formas de scoring que se indican en el paper.

$$score(u, i)_X = \begin{cases} \max_{j \in P_u} \{sim(V_i^X, V_j^X)\} & (maximum) \\ \frac{\sum_{j \in P_u} sim(V_i^X, V_j^X)}{|P_u|} & (average) \\ \frac{\sum_{r=1}^{\min\{K, |P_u|\}} \max_{j \in P_u}^{(r)} \{sim(V_i^X, V_j^X)\}}{\min\{K, |P_u|\}} & (average\ top\ K) \end{cases}$$

Muestre ejemplos de recomendación con imágenes reales para algún usuario en particular, comparando las imágenes con las que este ha interactuado y las que él se le recomendaron. Comente.


En este caso se realizará la predicción de 10 recomendaciones al usuario_ID = 566, ya que es el usuario que más activo estuvo en el dataset, por lo que es el usuario que más opciones presenta a la hora de comparar.

Lo primero que se realiza es como se diferencian los resultados al cambiar la métrica de trabajo

Comparación con dos métricas diferentes

```
find_similar_images(pca20_resnet_featmat, metrics=['cosine', 'euclidean'], query_id=566, topk=10)
```

query_id = 566



----- retrieved with metric = cosine -----
 unexpected r.status_code = 404 for url = <http://niebla.ing.owc.cl/iic3633-2020/18b07a6816d358033e0f98173be413a2.jpg>
 unexpected r.status_code = 404 for url = <http://niebla.ing.owc.cl/iic3633-2020/18b07a6816d358033e0f98173be413a2.jpg>
 0) id = 566
 1) id = 1368 not found
 2) id = 1359
 3) id = 1053
 4) id = 531



Imágenes con ID=566 y métrica cosine

----- retrieved with metric = euclidean -----
 unexpected r.status_code = 404 for url = <http://niebla.ing.owc.cl/iic3633-2020/18b07a6816d358033e0f98173be413a2.jpg>
 unexpected r.status_code = 404 for url = <http://niebla.ing.owc.cl/iic3633-2020/333066c091165a2088cce550510890aa.jpg>
 unexpected r.status_code = 404 for url = <http://niebla.ing.owc.cl/iic3633-2020/18b07a6816d358033e0f98173be413a2.jpg>
 unexpected r.status_code = 404 for url = <http://niebla.ing.owc.cl/iic3633-2020/333066c091165a2088cce550510890aa.jpg>
 0) id = 566
 1) id = 1368 not found
 2) id = 531
 3) id = 1082
 4) id = 1133
 5) id = 738 not found
 6) id = 1375
 7) id = 1053
 8) id = 1262
 9) id = 1414




Imágenes con ID=566 y métrica euclidean

En ambas métricas no encontró ninguna coincidencia con lo que vio este usuario.

Métrica 1: Cosine	Métrica 2: Euclidean
722c46813f6089b477e5f0e71d53b857	e2940a5b7183ba05d388c08a6014630a
e2940a5b7183ba05d388c08a6014630a	18b07a6816d358033e0f98173be413a2
18b07a6816d358033e0f98173be413a2	b656aca81017c0abe79f3c634ad5b769
436586997626907aa64f4fe0c04542bf	bdfdb7d6667695cae16e753baa031ad8
4d780590fccd5671c3a19ad98364926a	436586997626907aa64f4fe0c04542bf
072ebe6f3035c78b48aaa1baac31e435	f7755f380783e9b25d63b25732e4f1cf
333066c091165a2088cce550510890aa	2828176a88cdad687f5f5c0be45fef80
173a4c19133680e26c09412f0a8d52ac	7355d852c6afae96dde7c4b62b1a8732
bdfdb7d6667695cae16e753baa031ad8	f3d27a6e6a4769d0d7a20365e3b15c13
058255733e4700b4e11326f5b5e2b846	0b3d9edfc400b9a2a361a0d0f1ec024e

Tabla de recomendaciones realizada por el primer recomendador

Recomendaciones realizadas al usuario ID = 566 para el SCORE propuesto.

- SCORE MAX

```
[76] recommend_score_MAX_list(pca50_resnet_featmat, query_id=566, topk=10, metrics='cosine')
```

```
[643, 621, 956, 1242, 66, 929, 233, 63, 47, 1566]
```

```
recommend_score_MAX(pca50_resnet_featmat, query_id=566, topk=10)
```

```
user_id = 566
```

```
query_id = 566
```



***** Recommended items *****

```
unexpected r.status_code = 404 for url = http://nicbia.ing.usc.cl/ii:3633-2820/3ed08b4777f6bf36cc837e812f5a1c5_5og
```

```
unexpected r.status_code = 404 for url = http://nicbia.ing.usc.cl/ii:3633-2820/3ed08b4777f6bf36cc837e812f5a1c5_5og
```

01 id = 643



11 id = 621

21 id = 956



31 id = 1242



41 id = 66



Métrica score MAX
958ed8dd46c9968c4a0220f32dad6ed1 a5ee5f2a0fd4cdb5a4e7aaab07d63ac6 d1f8ced69658b517c2e05395d1b5e0c0 1ceddef812965cead71953e468568e7a 2551ef4a5d8a4efd3c99624e32af8c42 76b4353b159332f0c99383138bd4b1fa cc0ddcc15fe67cc5fb1e6c81dc6efba4 6cd87e92d8619fca44e5fce4e843fc5d 4904b8519c467706e3fdd94f67db1410 3edb9b4777fdbf3dcdc037b912f941c5

Recomendaciones con Métrica score MAX

• SCORE AVERAGE


```
[63] recommend_score_average_List(pca20_resnet_featmat, query_id=560, topk=15, metrics='cosine')
```

```
[1, 13, 4, 2, 18, 7, 12, 14, 16, 1552, 1592, 1599, 1585, 1528, 1587]
```

```
recommend_score_average(pca20_resnet_featmat, query_id=560, topk=15, metrics='cosine')
```


```
user_id = 560
```

```
query_id = 560
```




***** Recommended items *****


```
00 id = 1
```




```
13 id = 13
```




```
21 id = 4
```



```
30 id = 8
```



```
40 id = 10
```



Métrica score AVERAGE
f9514478a48522018a1fe464c4baf8a7 a1602405d009e31497d9e7547ef9b4df e2f366ef70185c50be8df2e1ccd3fc56 455d5f404c882b7d46cf356f404bd795 fcb385c996d6f6eef6fe0d10a992b5b5 1a92746ae6815bbbba2770a8ee29ee69d 9b6e8834a85759970e37b38844647bf5 3f89bd22a2a8e35b6603ff15ac239f34 4f7b2dcbd0f3edb1f04b7bf33eebdd6b 58251ea6d5cce3ebccd2d9c5e4a89270 829c6d361bda483006a83f1f597b2eaf 169c797b6acec6efe72a685cb957ac04 d2dd564d4c89abb818c58969559c5084 1c411378efad77bc07fab0f866e46b9


Recomendaciones con Métrica score AVERAGE

- SCORE AVERAGE TOP K


```


recommend_score_average_topk_list(pca20_resnet_featmat, query_id=566, topk=10, K1=350)
[13, 18, 0, 7, 8, 10, 15, 11, 1592, 1591]
recommend_score_average_topk_list(pca20_resnet_featmat, query_id=566, topk=10, K1=350)
user_id = 566
query_id = 566


```





***** Recommended items *****

0) id = 13


1) id = 18


2) id = 0


3) id = 7


4) id = 8


Métrica score AVERAGE TOP K
a1602405d009e31497d9e7547ef9b4df
4a2d646a5b22e8eae8f57089133b4921
d526333a98d58328d47265a4e067323f
1a92746ae6815bbba2770a8ee29ee69d
455d5f404c882b7d46cf356f404bd795
fc385c996d6f6eef6fe0d10a992b5b5
942cc1dfd343f0959d4e9b8d24aa3524
749c374ee744bdf1c1c48748d14cd8bd
58251ea6d5cce3ebccd2d9c5e4a89270
829c6d361bda483006a83f1f597b2eaf

Recomendaciones con Métrica score AVERAGE TOP K

Con el mejor de los métodos (y su mejor conjunto de hiperparámetros) genere 10 recomendaciones por cada usuario del dataset de testing. La lista de recomendaciones debe ser entregada en este formato:

Conclusiones

Dentro del principal problema presentado en este trabajo, fue con el data set de entrenamiento que se contaba para el análisis, ya que las imágenes que mayor cantidad de interacciones fueron 7 (0,0090%), lo que causo que los valores obtenidos de MAP@20 y nDCG@20 son muy bajos, en el orden de magnitud de 10^{-5} . Aunque se modificaron los hiperparámetros de los modelos analizados, tanto ALS como BPR presentaban valores muy pequeños.

Los modelos implementados, sobre todo los sistemas por contenido, no lograron entregar ninguna coincidencia al seleccionar 10 recomendaciones, aunque se probó con el usuario que más interacciones realizó en el dataset, ID_Usuario 566. Aunque algunas métricas como el de contenido, se logra apreciar en algunos casos que hace una buena predicción, por lo parecido que eran las imágenes, se puede suponer que tal vez, sea otra la razón por la cual el usuario elige ver una imagen y la forma en que el data se comportaba también pudo haber afectado esto.

Esta actividad fue interesante a la hora de poder tener una experiencia de como implementar y probar algoritmos para poder generar una estrategia de como poder enfrentar un desafío, personalmente la metodología de tener el desafío y luego aprender el contenido no es muy atractiva para mí, pero muchas veces más de lo que uno quisiera son la forma más segura de que son los problemas y desarrollos.

Anexos

Se adjuntan las recomendaciones en archivos *.JSON. No me quedo claro si eran 10 o 20 recomendaciones, por esta razón adjunte ambas.

- Recomendador modelo ALS

[Recomendaciones propuestas segun formato\data_model als 50_01_10.json](#)

[Recomendaciones propuestas segun formato\data_model als 50_01_20.json](#)

- Recomendador bpr

[Recomendaciones propuestas segun formato\data_model bpr 200_01_0001_10.json](#)

[Recomendaciones propuestas segun formato\data_model bpr 200_01_0001_20.json](#)

- Recomendador score AVERAGE

[Recomendaciones propuestas segun formato\recommend Score Average 10.json](#)

[Recomendaciones propuestas segun formato\recommend Score Average 20.json](#)