BÚSQUEDA Y MINERÍA DE INFORMACIÓN

PRÁCTICA 4

Sistemas de recomendación y análisis de redes sociales

Tomas Higuera Viso
Miguel Antonio Núñez Valle
Pareja 25

Pregunta 1

- Hemos realizado el ejercicio de **estructuras de datos y recomendación simple**.
- Hemos implementado el **filtrado colaborativo kNN**.
- Hemos implementado la **recomendación basada en contenido**.
- En la ampliación de algoritmos hemos implementado tanto **vecinos próximos orientados a ítem** como el algoritmo basado en **contenido por vecinos próximos**.
- Hemos creado las tablas con las métricas y tiempos de ejecución adjuntadas más abajo, aunque no hemos podido ejecutar nuestra implementación de algunos recomendadores con los conjuntos de datos grandes.
- Hemos creado las dos **redes sociales simuladas**, así como la representación y análisis de todas las topologías.
- Hemos implementado las **métricas de análisis de topologías de red**.

Pregunta 2

En este ejercicio analizaremos los resultados obtenidos con los distintos recomendadores. Como hemos explicado no hemos podido ejecutar los recomendadores con todos los conjuntos de datos, ya que nuestro ordenador tardaba mucho tiempo en realizar las pruebas con los mismos. Los datos que no hayamos podido obtener los dejaremos representados en la tabla con NaN.

Tabla conjunto de datos: *Toy dataset*

	Rmse	**Precision**	**Recall**	**Tiempo**
Random	2.486	0.04	0.4	3ms
Majority	2	0.04	0.4	0ms
Average	0.5	0.02	0.2	1ms
User-based kNN	2.226	0.04	0.4	11ms
Normalized user-based kNN	0.508	0.02	0.2	0ms
Item-based NN	1.553	0.04	0.4	2ms
Centroid-based	2.369	0.04	0.4	1ms

Item-based NN (Jaccard)	10.794	0.04	0.4	19ms

Tabla conjunto de datos: *MovieLens latest-small datatest*

	Rmse	**Precision**	**Recall**	**Tiempo**
Random	2.703	0.003	0.001	1.293s
Majority	519.562	0.155	0.780	1.29s
Average	1.092	8.197E-4	3.447E-4	1.394s
User-based kNN	22.237	0.258	0.16	47.14s
Normalized user-based kNN	0.879	0.012	0.008	42.547s
Item-based NN	NaN	NaN	NaN	NaN
Centroid-based	3.844	0.063	0.031	38.639s
Item-based NN (Jaccard)	NaN	NaN	NaN	NaN

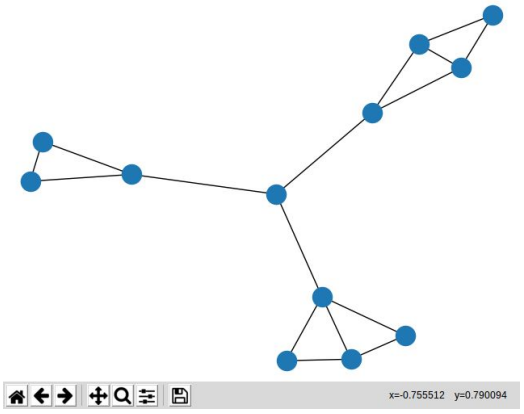
Tras realizar las tablas podemos observar que con conjuntos de datos pequeños (la primera tabla), los resultados que obtenemos con nuestros recomendadores no son muy buenos, sin embargo cuando trabajamos con conjuntos de datos mayores (segunda tabla), los resultados que obtenemos son considerablemente mejores.

Como hemos dicho al principio no hemos podido ejecutar los recomendadores basados en ítem, ya que los tiempos de ejecución eran muy grandes.

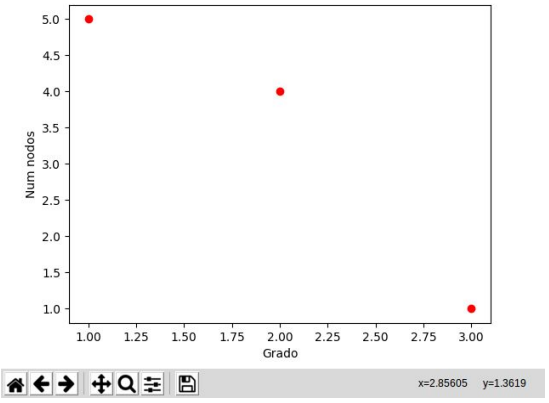
Pregunta 3

A continuación representaremos las diferentes topologías de red que se nos proporcionan, así como las generadas por nosotros.

small1.csv

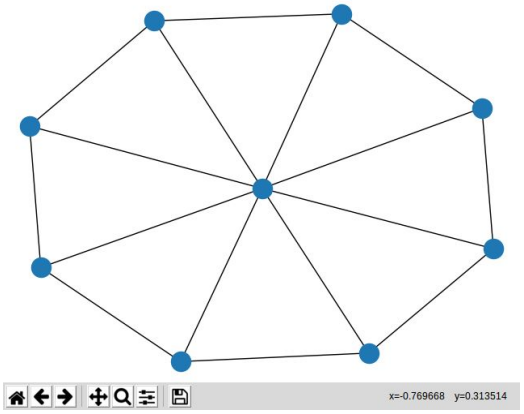


Representación de la topología

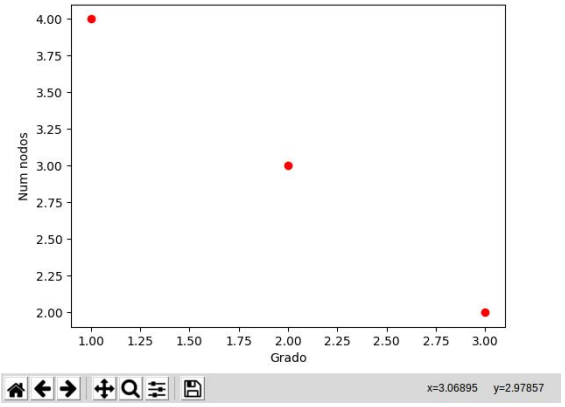


Grafica con x = grado; y = num_nodos

small2.csv

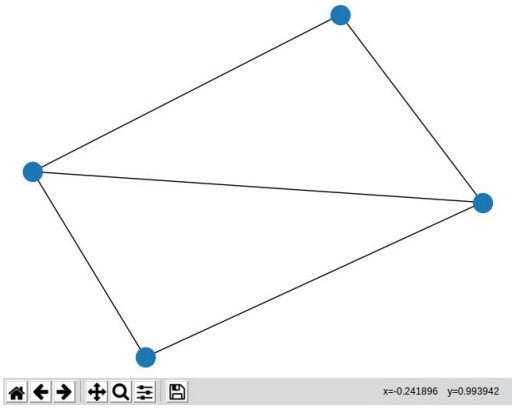


Representación de la topología

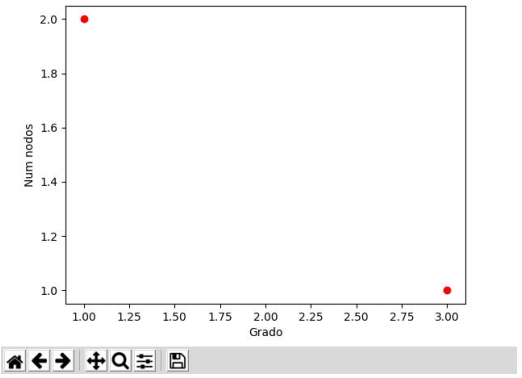


Grafica con x = grado; y = num_nodos

small3.csv

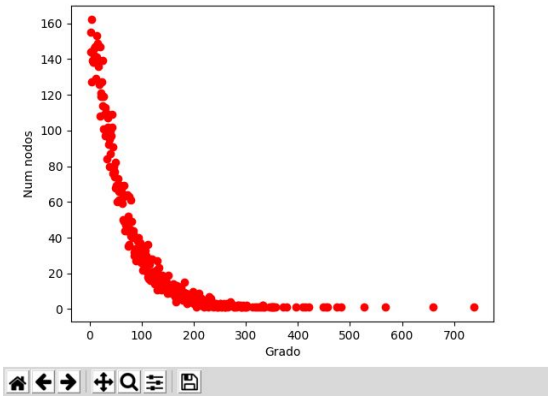


Representación de la topología

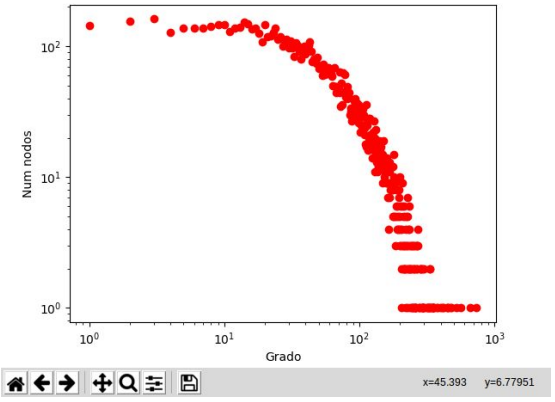


Grafica con x = grado; y = num_nodos

twitter.csv

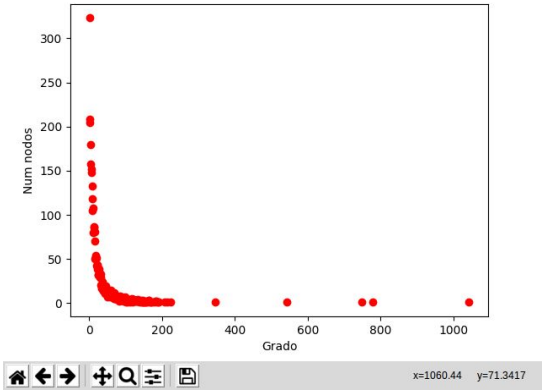


Grafica con x = grado; y = num_nodos

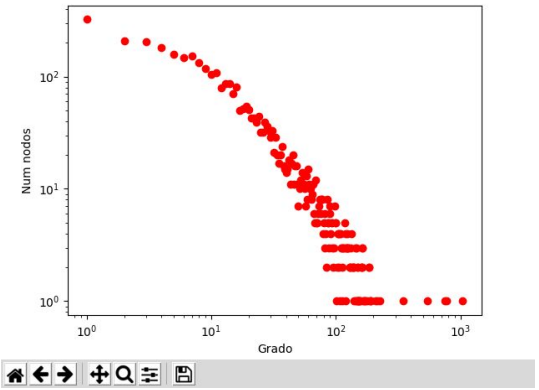


Grafica con x = grado; y = num_nodos con escala log

facebook.csv

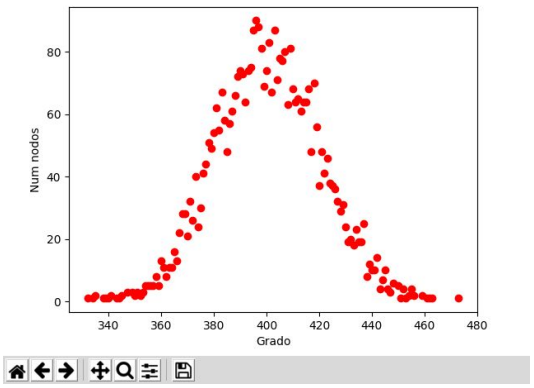


Grafica con x = grado; y = num_nodos

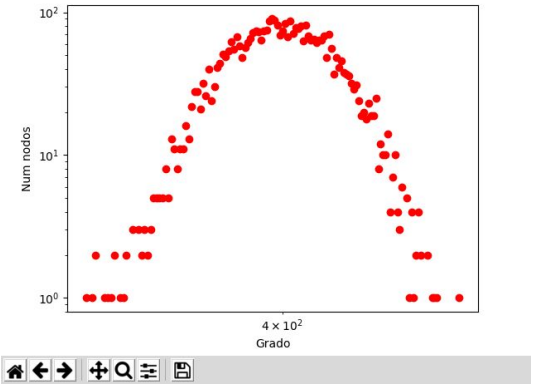


Grafica con x = grado; y = num_nodos con escala log

erdos.csv

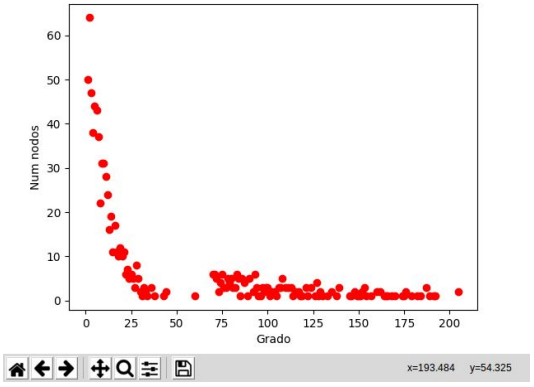


Grafica con x = grado; y = num_nodos

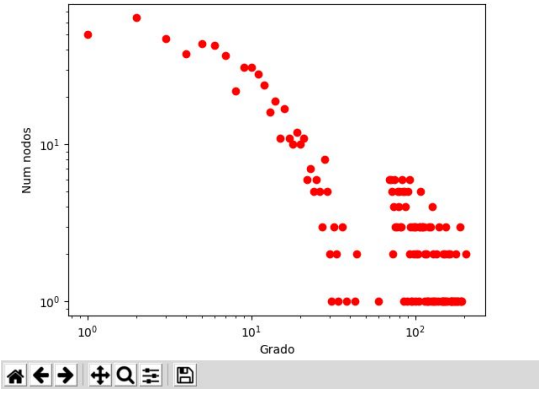


Grafica con x = grado; y = num_nodos con escala log

barabasi.csv



Grafica con x = grado; y = num_nodos



Grafica con x = grado; y = num_nodos con escala log

Conclusiones:

Donde mejor se observa una distribución *power-law* es en las representaciones de twitter, facebook. En la distribución del conjunto de datos de barabasi, también puede apreciarse este tipo de distribución aunque menos clara que en los otros dos conjuntos.

Sobre la paradoja de amistad, la explicaremos en base a los conjuntos de datos de twitter y facebook, ya que son los conjuntos con más datos. Esta paradoja se explica en estos conjuntos diciendo que hay muchas personas con pocos amigos y pocas personas con muchos amigos, lo que al final se resumen con que es fácil ser amigo de personas con más amigos que que con menos, ya que estas topologías representan una distribución *power-law.*