SNA PROJECT UNIT 3

MARIANO MIRANDA SANCHEZ

DATA ENGENIREENG 8°B
UNIVERSIDAD POLITECNICA DE YUCATAN

Introduction

El enfoque principal de este proyecto es el analisis de una red social usando Link Prediction Analysis a los datos obtenidos, los datos recopilados son reunidos utilizando tecnicas de WebScrapping o con la API de la red social utilizada, para este trabajo se usara la API de Reddit la cual tiene la facilidad de ser gratuita y de utilizar.

El conjunto de datos utilizado es una red de usuarios que comentaron en una publicación de Reddit acerca de franquicias de Nintendo, aquí podemos entender que cada nodo es un usuario y que cada enlace es un comentario.

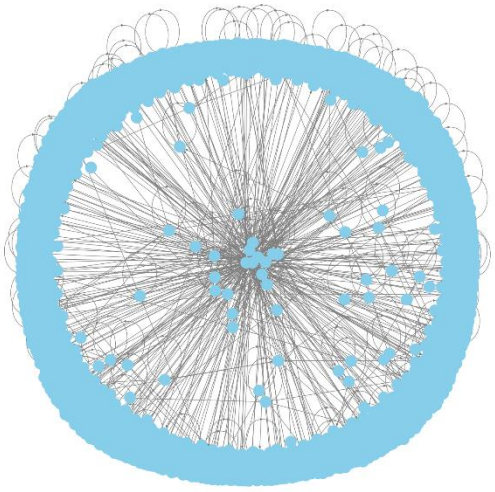
Maping process

Para poder llegar a la recopilación de los datos, se utilizo la biblioteca praw para poder conectar con la pagina y así poder extraerlos, para este caso en especifico, se uso la API de Reddit en la que para poder acceder a la información, se dio la autenticación de las credenciales de la API de desarrollador, que son el ID del usuario, ID secreto y el nombre de usuario de este mismo, en la búsqueda de los comentarios se encontraron algunos los cuales estaban eliminados, estos deben ser eliminados para evitar problemas al momento de extracción y que puedan ser guardados sin problema alguno.

Se almacenaron en un archivo '.json' debido a que es mas sencillo que al momento de leer datos extraídos de paginas web los lea en ese formato, en comparación a un '.csv' el cual da en su mayoría problemas al almacenar los datos recopilados.

Basic characteristics and Visualization

Las conexiones de la red que se muestra a continuación muestras comportamientos extraños a diferencia de otros casos visto en grafos.



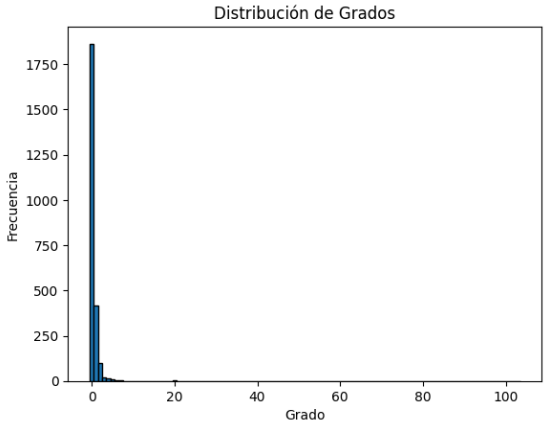
El mapeado de la red tiene ciertas características que se ven a simple vista, estos mayormente no forman grupos como pueden ocurrir en otros grafos, los nodos que se encuentran en los extremos son usuarios los cuales no tienen una gran cantidad de interacciones, a comparación de los nodos que se encuentran mas en el centro del grafo, los cuales tienen una mayor cantidad de interacciones que el resto, esto nos da a entender que la zona central representa una mayor cantidad de interacciones, siendo esta la razón de por que hay una cantidad pequeña de nodos en comparación del resto, la cantidad de comentarios con mayor cantidad de respuestas son muy bajas que los comentarios sin respuestas.

El grafo en si es muy denso al tener muchas conexiones, siendo que la interacción entre usuarios es numerosa, siendo como se mencionaba anteriormente, aquellos nodos que se encuentran en el centro son los mas activos en comparación del resto de nodos.

El numero de nodos de la red es de 2430, con un total de 625 links siendo estas, el total de respuestas que ha habido entre usuarios, muchos de estos regresan a interactuar consigo mismo debido a que son comentarios de respuestas que han recibido en su comentario. El Grado promedio es de 0.511403 siendo bajo para este caso y algo extraño, ya que nos dice que en promedio cada usuario tiene un poco mas de medio enlace, todo esto nos indica que aun con un cierto numero de nodos, las respuestas entre estos no son

grandes. El coeficiente de agrupamiento global es de 0, siendo al tendencia de los nodos inexistentes, lo que significa que no hay grupos o subgrupos conectados entre si ya que en su mayoría todos están en un mismo grupo o no hay suficientes para poder formar otros grupos. La densidad de la red es de 0.0001, siendo este un valor extremadamente bajo, lo que indica que solo una pequeña parte de las posibles conexiones que hay, sean reales, esto es normal en redes sociales donde hay una cantidad enorme de interacciones y las conexiones son esporádicas y no todos los usuarios interactúan entre si.

La grafica de Degree distribution muestra que la mayoría de los usuarios tienen pocas interacciones pero habiendo algunos usuarios los cuales tienen una mayor actividad que otros, estos mientras mas se alejen son mas escasos.



En el análisis de medidas de centralidad buscando los top 5, varia los resultados dependiendo del campo que se hable, en el primero de Degree centrality el nodo principal tiene un valor de 0.042, siendo quien tiene el mayor numero de conexiones directas en la red. En Betweennes centrality el principal es tiene 3.522, lo que significa que tiene una mayor frecuencia de caminos mas cortos entre otros nodos, siendo un intermediario clave en la red. Closeness Centrality su nodo principal tiene un valor de 0.0018, siendo que este usuario esta relativamente cerca de todos los demás nodos en los extremos de la red y es capaz de alcanzar a otro nodos rápidamente.

Link Predictions

Finalmente se da una predicción de los 5 enlaces mas probables, los cuales representan pares de nodos que probablemente formaran conexiones futuras, esto basado en los vecinos compartidos de estos nodos, los resultados obtenidos fueron de cierta forma diferentes a comparación de los restantes, siendo los primeros lugares los nodos cuyo valor fue de 1.5 en el que tiene la mayor probabilidad de conseguir una nueva conexión gracias a la conexión de sus vecinos y a su propia cercanía, así mismo los restantes que consiguieron un valor de 0.5 sin nodos los cuales se mantienen en una distancia relativamente alejada.

Top 5 Enlaces Más Probables:
'Ok-Service-1127', 'AnimusFlux' = 1.5
'merpderpherpburp', 'queasy_logophile' = 0.5
'Brisby820', 'Special-Menu-3231' = 0.5
'queasy_logophile', 'AppleHumper' = 0.5
'justforthis2024', 'm0thership17' = 0.5

Siendo estos los nodos y los usuarios que son los mas probables a formar una conexión.