Redes de palabras

Claudio Collado
cjccollado@gmail.com

Flavia Felicioni
flaviafelicioni@gmail.com

Adrián Marino
adrianmarino@gmail.com

***Abstract*—En este trabajo se busca comparar la relación entre las asociaciones libres entre palabras recolectadas en el experimento *Small World of Words* [DDNP+19] y las asociaciones semánticas obtenidas con el *embedding* word2vec. Tras un procesamiento se construyen dos grafos que son utilizados para calcular distintos coeficientes que permite caracterizarlos y compararlos. Además estos grafos permiten la búsqueda de comunidades y de algún tipo de estructura presente tanto en la red de asociación como en la red semántica.**

***Index Terms*—comunidades, embeddings, word2vec, Small world of words, grahps, centrality, modularity, CUE-R1**

I. Introducción

El análisis basado en redes y grafos es un campo en constante evolución para el estudio de sistemas complejos [Bar13]. La lingüística tampoco escapa a esta tendencia. De acuerdo a [ECBS09], a partir de asociaciones semánticas simples es posible inferir la representación mental del glosario de uso cotidiano. Los modelos más sencillos representan palabras como nodos interconectados, donde por ejemplo las más relacionadas estarán a una menor distancia en este diccionario mental [JKM06], o las palabras polisémicas resultan ser hubs que conectan conceptos alejados [SC02] dándole una estructura de *Small-World* a la red de palabras.

En este trabajo se usa un conjunto de datos extraído del proyecto *Small World of Words* [DDNP+19], donde cada participante contestó hasta tres palabras ante una palabra preguntada, construyéndose así una red de asociaciones libres uniendo los nodos de las palabras que los participantes vincularon. Para construir dicha red se realizaran distintas tareas de pre-procesamiento que serán descriptas en la sección II.

Asimismo, se puede definir una distancia semántica entre palabras a partir de lo que se conoce como embeddings, que es una representación de palabras en un espacio vectorial. Estas distancias son típicamente una distancia coseno. En este trabajo se utilizará el embedding word2vec provisto por la cátedra [KB21] tal como se presenta en II.

En el presente informe se siguen los lineamientos presentados en las referencias para el estudio y exploración de datos de asociaciones y se siguen los puntos de la guía [KB21].

La resolución está accesible en [CFM21] y es utilizada como guía para completar el contenido de las diferentes secciones.

II. Métodos

En este trabajo se utilizan dos datasets, el primero de ellos obtenido con el experimento *Small World of Words* se llama SWOW-EN2018 y esta disponible en [1] y el segundo que se utiliza es el word2Vec pre-entrenado de GoogleNews proporcionado por la cátedra [2].

El dataset *SWOW-EN2018: Preprocessed* contiene palabras en inglés que ya fueron limpiadas de carácteres raros. Además, se le extrajeron las columnas de preguntas *CUE* y primeras respuestas *R1* las que fueron renombradas como **source** y **response** respectivamente y a posteriori se le aplicó un filtrado para eliminar las palabras con menos de 2 letras, aquellas filas con nulos y las **stop words** de varios lenguajes.

A continuación y a partir de GoogleNews se generó un diccionario en el cual la clase es una palabra y el valor es un vector de *embeddings*. Este diccionario sólo fue generado para las palabras que se encuentran en el dataset pre-procesado del experimento *Small World of Words*.

Una vez hecho realizado el procesamiento descripto de ambos datasets se obtuvieron los respectivos grafos denominados *Gsww* y *Gw2v*.

En el caso de *Gsww* se utilizó como peso la expresión (1). En las figuras 1 y 2 se puede apreciar el histograma de las preguntas CUE y las respuestas R1, respectivamente, mientras que en la figura 3 se muestra el peso que es una aproximación de la probabilidad condicional obtenida con (1).

$$weight = \frac{frecuencia(CUE - R1)}{frecuencia(CUE)} \tag{1}$$

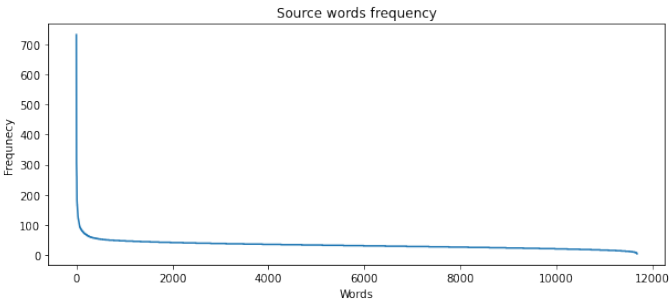

Fig. 1: Frecuencia de palabras CUE (source)

[1] https://smallworldofwords.org/en/project/research
[2] https://s3.amazonaws.com/dl4j-distribution/GoogleNews-vectors-negative300.bin.gz

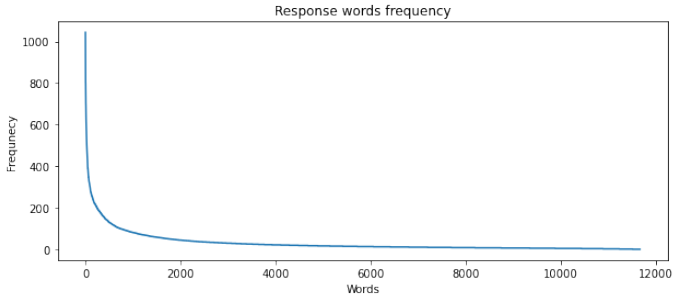

Fig. 2: Frecuencia de la respuesta R1

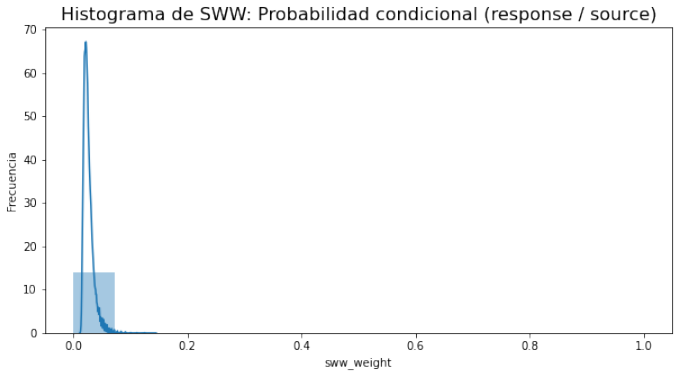

Fig. 3: Probabilidad condicional - Pesos de *Gsww*

Para construir el grafo *Gw2v* se utilizó como medida de similaridad la distancia coseno (calculada a partir de los word embeddings) de las palabras filtradas desde *Small World of Words*.

Dicha distancia se calcula entre dos vectores a y b con la expresión (2). En la figura (4) se puede apreciar el histograma de los pesos calculados para las palabras de *word2vec*.

$$s(x, y) = cos(\phi) = \frac{a.b}{\|a\|\|b\|} \tag{2}$$

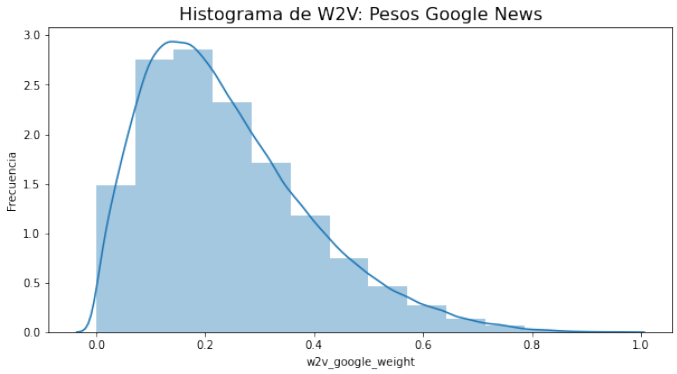

Fig. 4: Histograma de pesos GoogleNews - Pesos de *Gw2v*

Para obtener los resultados de la siguiente sección se utilizaron varias herramientas de la librería NetworkX de Python [AHA+08].

III. Resultados y discusión

A. Tarea 1 - Construcción de los Grafos

El grafo obtenido de *Gw2v* tiene un total de 1118 nodos y 2831 aristas, no es conexo, es pesado, no es dirigido, tiene ciclos y no es un grafo con múltiples aristas.

En la figura (6) se presenta un gráfico con submuestreo de nodos para mejorar la visualización del grafo pesado *Gw2v*. Se puede notar que los nodos en verde más oscuro son aquellos con mayor centralidad de grado (y viceversa). Asimismo en la visualización de la matriz de adyacencia de figura (5) se puede apreciar que la misma es dispersa o rala.

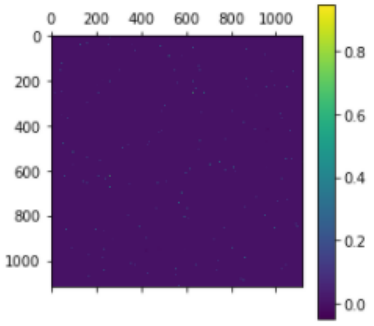

Fig. 5: Matriz de adyacencia

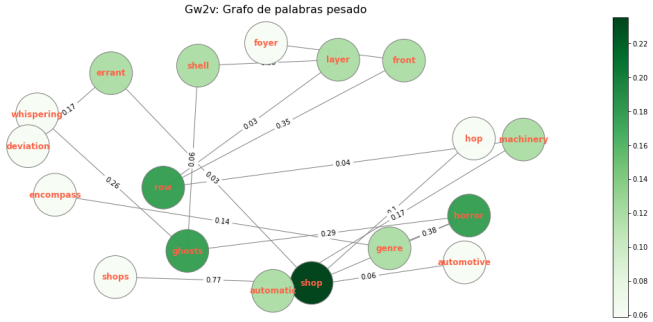

Fig. 6: Grafo pesado de *Gw2v* (submuestreo de nodos)

El grafo obtenido de *Gsww* tiene un total de 1118 nodos y 3122 aristas, es un grafo pesado, dirigido, tiene ciclos y no tiene múltiples aristas.

En la figura (8) se presenta el gráfico con submuestreo de nodos de *Gsww*. Se puede notar que los nodos en verde más oscuro son aquellos con mayor centralidad de grado (y viceversa). Asimismo la matriz de adyacencia de figura (7) también se aprecia dispersa o rala.

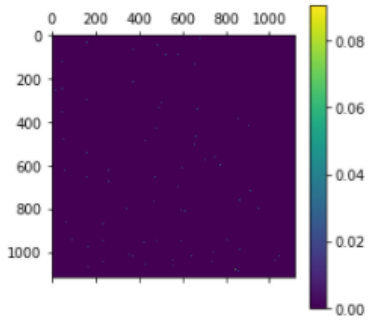

Fig. 7: Matriz de adyacencia

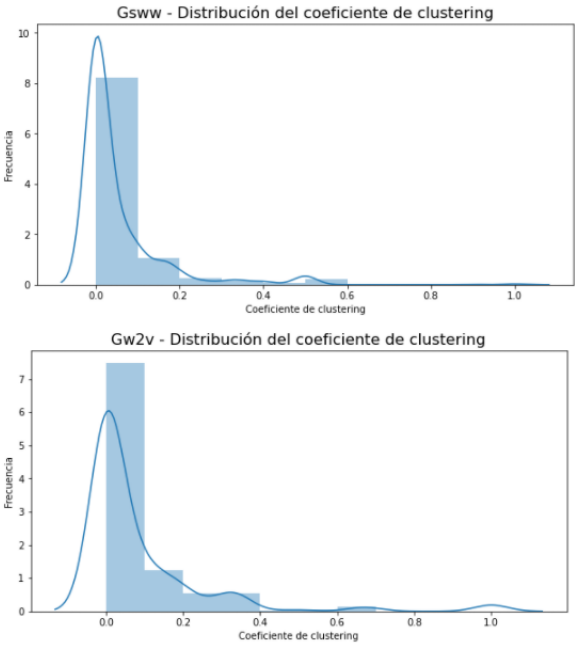

Fig. 9: Coeficiente de Clustering - *Distribución*

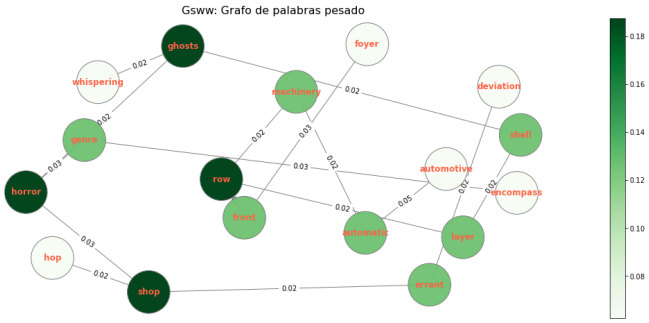

Fig. 8: Grafo pesado de *Gw2v* (submuestreo de nodos)

Al comparar los grafos de (8) y (7) se puede apreciar que varios de los nodos tienen intensidad de color muy similares (ver por ejemplo las palabras deviation, shop, whispering).

A su vez el coeficiente de clustering global para Gsww es de 0.053 y para Gw2v de 0.087, lo cual indica que Gw2v tiene vecinos mas conectados entre si que Gsww.

2) Medidas de Centralidad: En (10) se aprecia la centralidad de grado para ambas redes

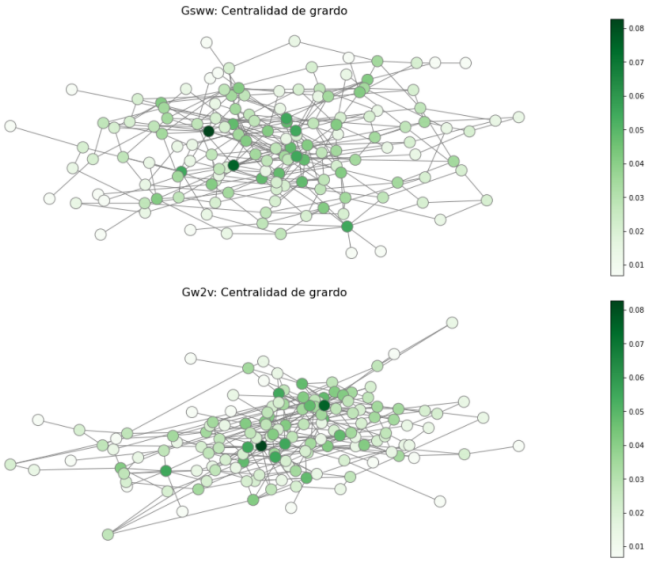

Fig. 10: Centralidad de Grado - *Gsww y Gw2v*

En (11) se aprecia la centralidad de grado para ambas redes

B. Tarea 2 - Caracterización de los Grafos

1) Coeficiente de Clustering: En (9) se aprecia que en ambos grafos existe el mismo patrón. Es decir, se tienen muchos nodos con bajo coeficiente de clustering y pocos con valores a partir de 0.2. Esto nos dice que tenemos muy pocos nodos con vecinos muy conectados entre sí y muchos nodos con vecinos poco conectados.

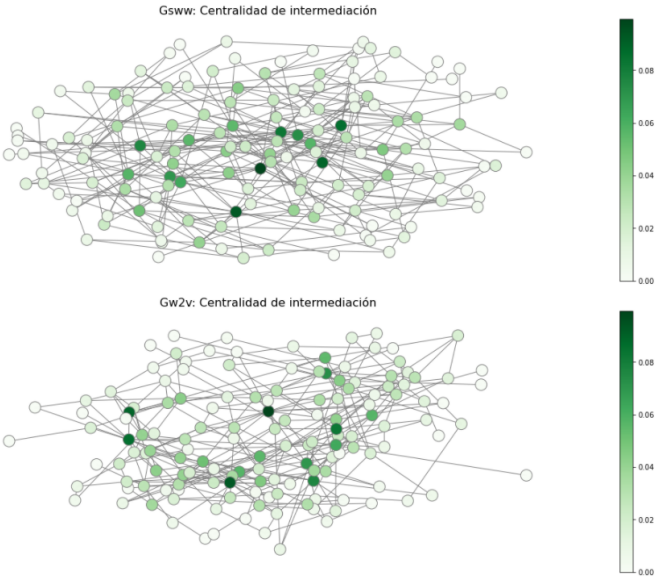

Fig. 11: Centralidad de Intermediación - *Gsww y Gw2v*

En (12) se aprecia la centralidad de cercania para ambas redes

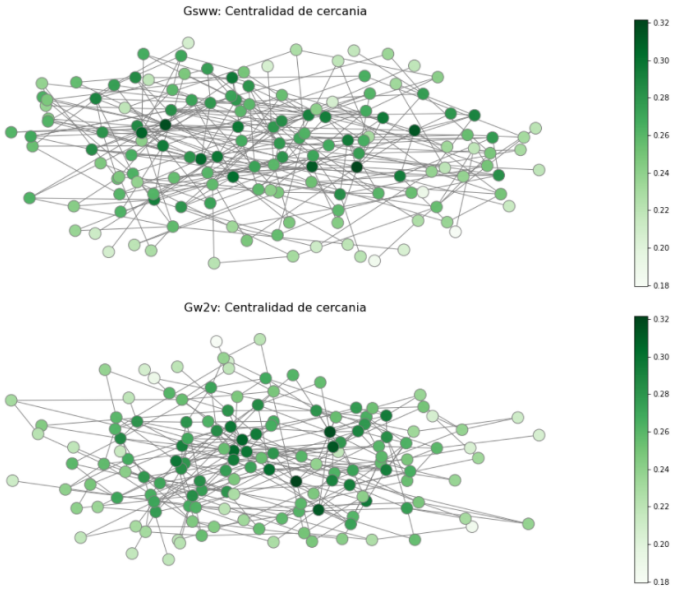

Fig. 12: Centralidad de Cercania - *Gsww y Gw2v*

En (13) se aprecia la centralidad de cercanía para ambas redes

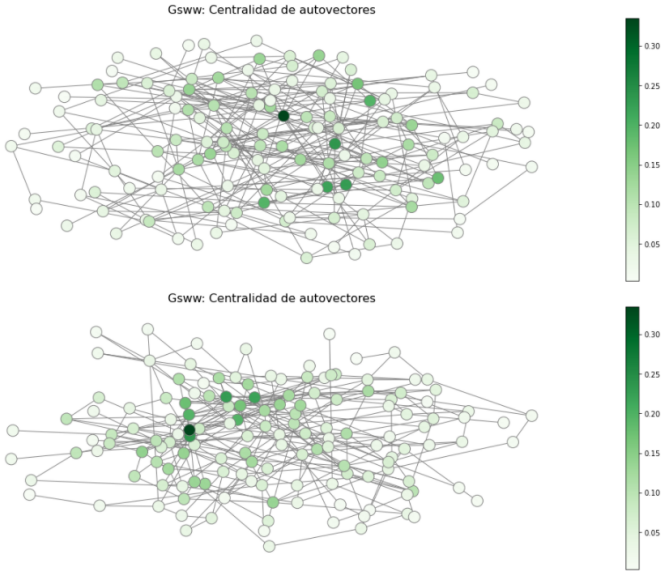

Fig. 13: Centralidad de Autovectores - *Gsww y Gw2v*

En (14) se aprecia la matriz de correlación para las medidas de centralidad

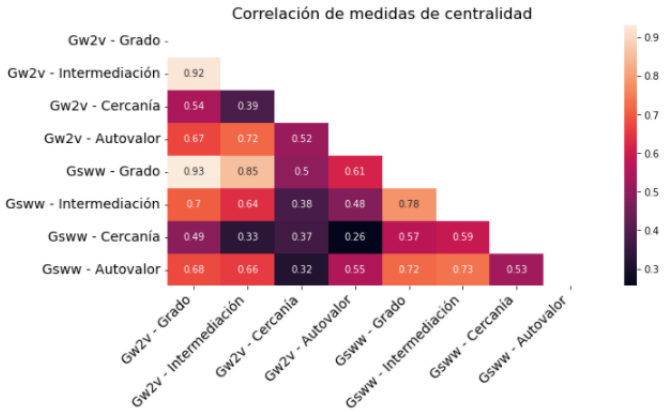

Fig. 14: Matriz de correlaciones - *Centralidades*

3) Distribución de pesos: En (15) se aprecia que las distribuciones de peso para ambos grafos Gsww y Gw2v.

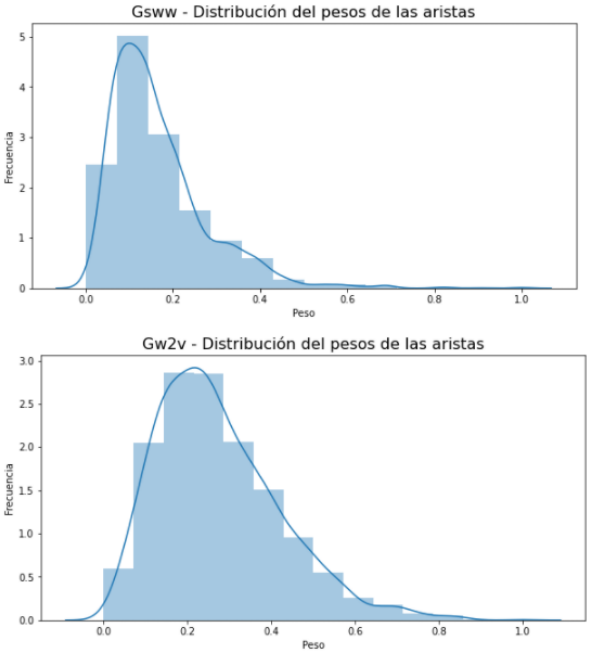

Fig. 15: Distribución de Pesos - *Aristas*

4) Distribución de grado: En (16) se aprecia que para ambos grafos las distribuciones de grado son muy similares pero no iguales, ya que hay palabras que existen en Gsww pero no existen en Gw2v. Por otro lado, se aprecia que tenemos muchos nodos bajo grado (la mayoría de grado 1) y muy pocos nodos de grado mayor a 1. Es decir que tenemos pocos nodos hubs.

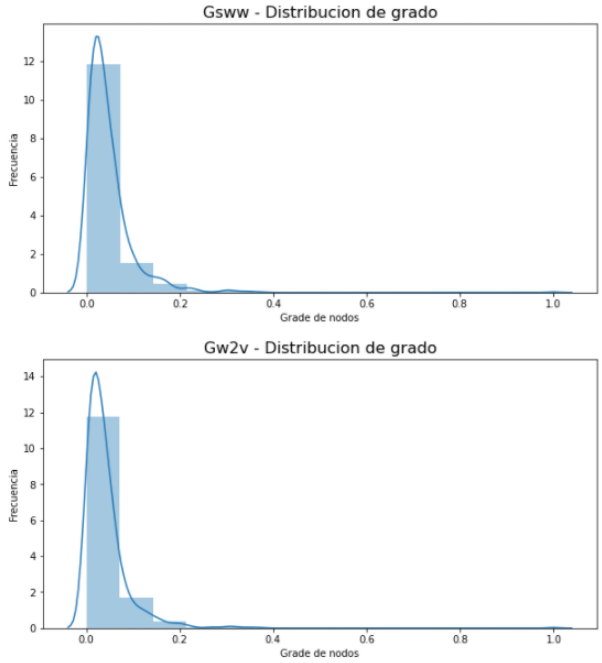

Fig. 16: Distribución de Grado

En (17) se aprecia que si comparamos el grado de ambos grafos vemos que ambas distribuciones son muy similares. La diferencia radica en que la distribución de Gsww disminuye su grado mas lentamente que Gw2v. Esto se debe a que Gw2v tiene menos conexiones ya que no todas la palabra del dataset SWW se encuentran en el embedding.

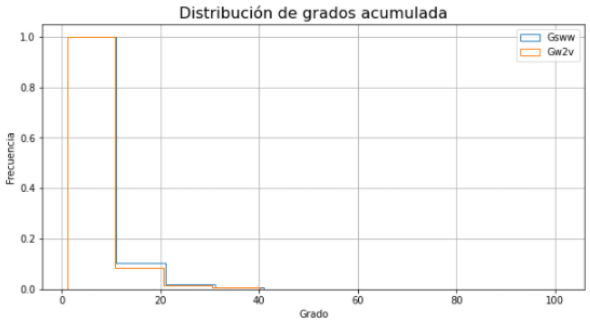

Fig. 17: Distribución de Grado acumulado

5) Coeficiente de Asortividad: El coeficiente de asortividad para Gsww es de 0.018, lo cual indica que es una red asociativa, es decir que, existe un grado de correlacion entre nodos de igual grado. Entonces, los nodos se agrupa por igual grado.

En cambio para Gw2v el coeficiente de asortividad es de -0.063, lo cual indica que es una red disociativa, es decir que, existe un grado de correlación entre nodos de distinto grado. Entonces, los nodos de menor grado se agrupan con los nodos de mayor grado.

6) Camino Mínimo - Diámetro - Densidad: Para ambos grafos se observa que tienen el mismo numero de aristas en su camino mínimo, lo cual es de esperar debido a que tienen la misma topología.

Con respecto al diámetro ambos grafos tienen el mismo valor de 10, lo cual es de esperar debido a que tienen la misma topología.

Finalmente la densidad para el grafo Gsww es de 0.002 mientras que para el grafo Gw2v es de 0.004, lo cual corresponde al doble

C. Tarea 3 - Comunidades

1) Algoritmo Girvan-Newman: Para descubrir comunidades se usa el algoritmo de Girvan-Newman ya que funciona con grafos dirigidos y no dirigidos. Este algoritmo de tipo jerárquico va creando grafos desconectados al remover aristas que se eligen por su alto valor de centralidad de *edge betweenness*.

Para poder aplicarlo se submuestrea la cantidad de nodos a 400 ya que el algoritmo no escala bien (para el tamaño de grafos obtenidos). Los nodos elegidos se seleccionaron de forma que tengan la mayor centralidad de grado.

En la figura 18 se muestra el índice de modularidad obtenido con este algoritmo para cada uno de los grafos. Los máximos valores de modularidad valen 0.496 para *Gsww* y 0.533 para *Gsww* y ocurren para una cantidad de particiones igual a 18 en ambos casos. Ambos valores máximos son muy similares y razonablemente altos.

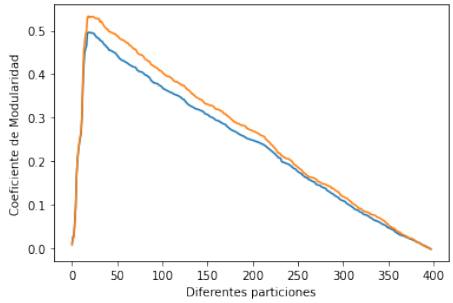

Fig. 18: Coeficiente de modularidad Girvan-Newman *Gsww*: azul, *Gw2v*: naranja

El índice Rand ajustado obtenido es 0.93 que es cercano a la unidad lo que indica que las comunidades de ambos grafos encontradas con el algoritmo de Girvan-Newman son muy similares.

En la figura 19 se muestra el grafo de *Gsww* realizado para un submuestreo de 80 nodos (para simplificar la visualización). Los colores de los grafos capturan la mejor partición predicha por el algoritmo para la comunidad de *Gw2v*. La intensidad de los colores se mantiene y se aprecia una aceptable coincidencia entre ambas comunidades, lo que es compatible con lo observado en el alto valor del índice Rand.

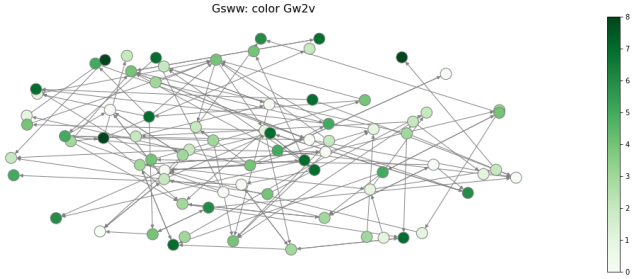

Fig. 19: Grafo Girvan-Newman para *Gsww* con colores en función de *Gw2v*

2) Otros algoritmos y modificaciones de grafos: A los efectos de comparar con otro algoritmo de detección de comunidades se utiliza el algoritmo de Louvain [Bar13] que calcula la partición de los nodos del grafo de forma de maximizar la modularidad.

Como dicho algoritmo sólo puede aplicarse a grafos no dirigidos, lo primero que se hizo es crear una versión del grafo Gsww no-dirigido.

En las figuras 20 y 21 se presenta un esquema simplificado de los resultados obtenidos al aplicar este algoritmo para los grafos Gsww y Gw2v respectivamente. En dichos casos las mejores particiones se obtuvieron para una cantidad de 12 y 14 comunidades, respectivamente.

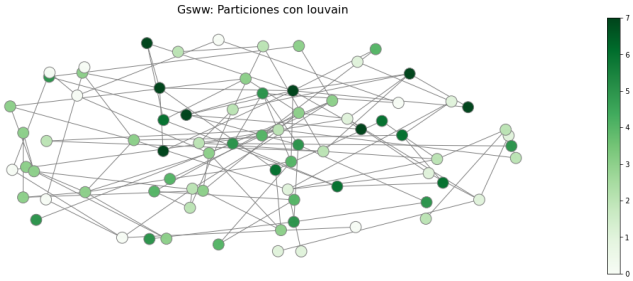

Fig. 20: Grafo algoritmo Louvain para *Gsww*

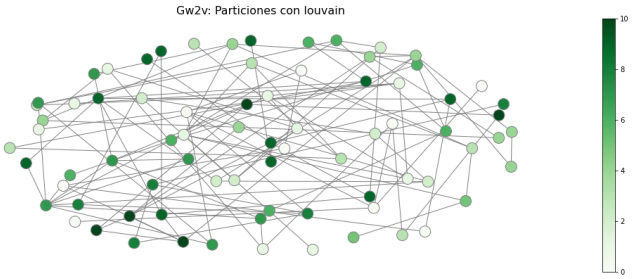

Fig. 21: Grafo algoritmo Louvain para *Gw2v*

Asimismo se aplicó nuevamente el algoritmo de Girvan-Newman pero al grafo no dirigido de Gsww, en la figura 22 se puede apreciar la modularidad obtenida. Su valor máximo alcanzado es de 0.559, ligeramente superior al obtenido con el grafo original que es dirigido. Resulta interesante destacar que el índice Rand obtenido al comparar con la comunidad encontrada con el mismo algoritmo pero aplicado al grafo dirigido es de 0.89 lo que indica que no hay grandes cambios entre las comunidades identificadas.

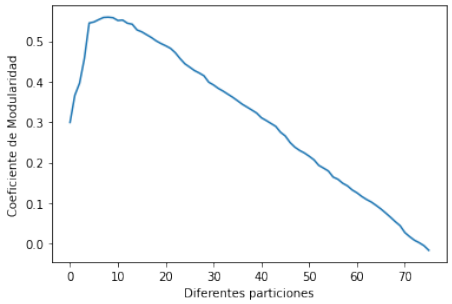

Fig. 22: Modularidad para gráfico no-dirigido del *Gsww* con Girvan-Newman

D. Tarea 4 - Small-world y redes prototipicas

Para esta tarea se consideraron las redes Gsww y Gw2v y se realizaron subgrafos de menor tamaño (100 nodos) para cada tipo, de forma tal de reducir los tiempos de ejecución. A su vez se generaron en forma sintética 500 redes Small-world y Scale-Free con igual cantidad de nodos y aristas, para luego comparar la longitud media de la ruta mas corta y el coeficiente de clustering con respecto a la red de asociaciones (Gsww)

y la red semántica (Gw2v) A continuación se observan los resultados obtenidos para redes de Small-world y Scale-Free

1) Redes Small-world: En la figura (23) se observan los resultados obtenidos para las redes Small-wolrd generadas artificialmente (azul) y los valores correspondientes a la red Gsww y Gw2v (rojo). Además en la figura (24) se observa que solamente la longitud media de Gw2v se encuentra dentro de la distribución obtenida para las redes Small-world, siendo el resto de las características muy diferentes en sus valores numéricos

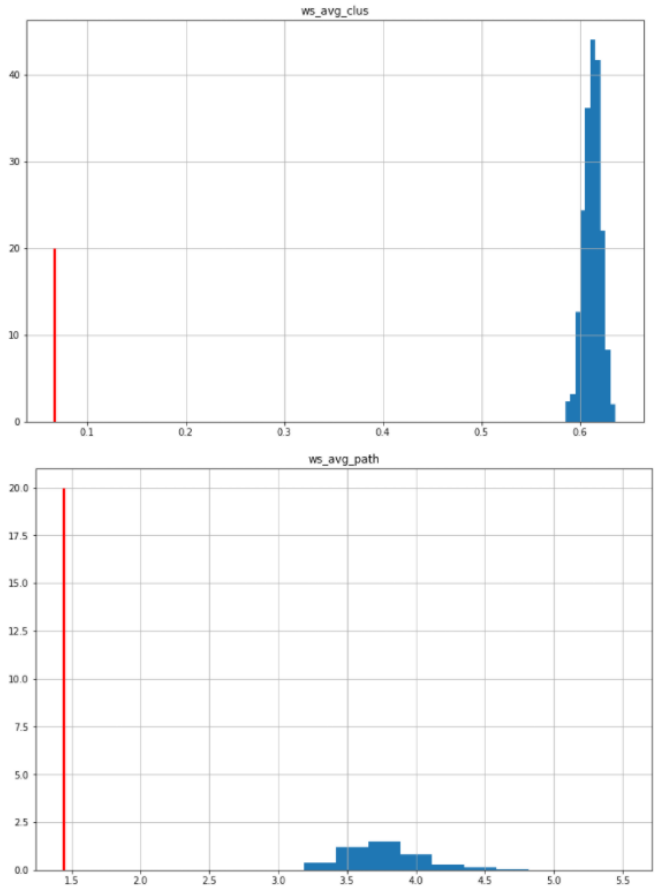

Fig. 23: Coeficiente de Clustering y longitud media - Small world y Gsww

2) Redes Scale-Free: En la figura (25) se observa los resultados obtenidos para las redes Scale-Free generadas artificialmente (azul) y los valores correspondientes a la red Gsww y Gw2v (rojo). Además en la figura (26) se observa que ninguna característica se encuentra dentro de la distribución obtenida para las redes Scale-Free.

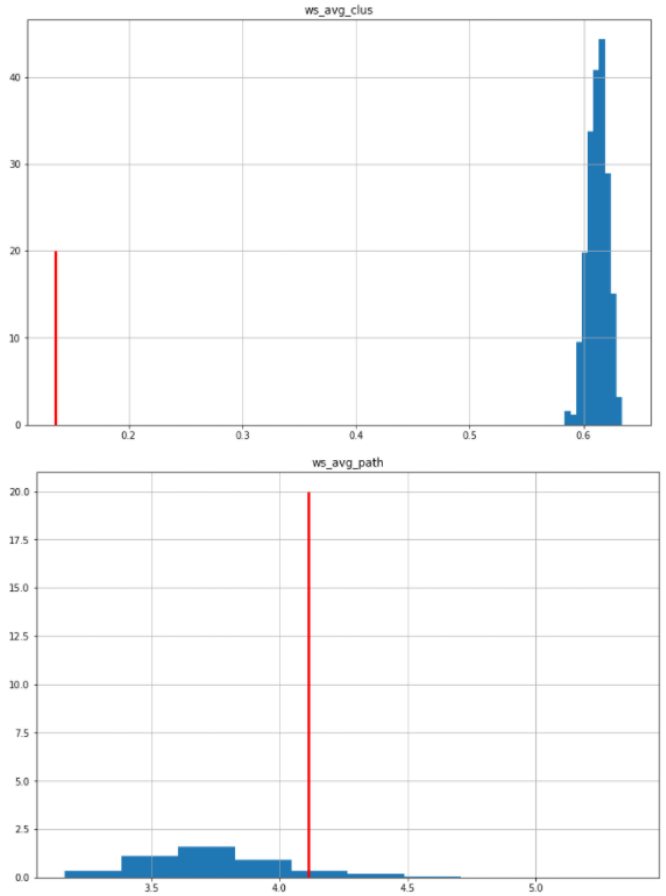

Fig. 24: Coeficiente de Clustering y longitud media - Small world y Gw2v

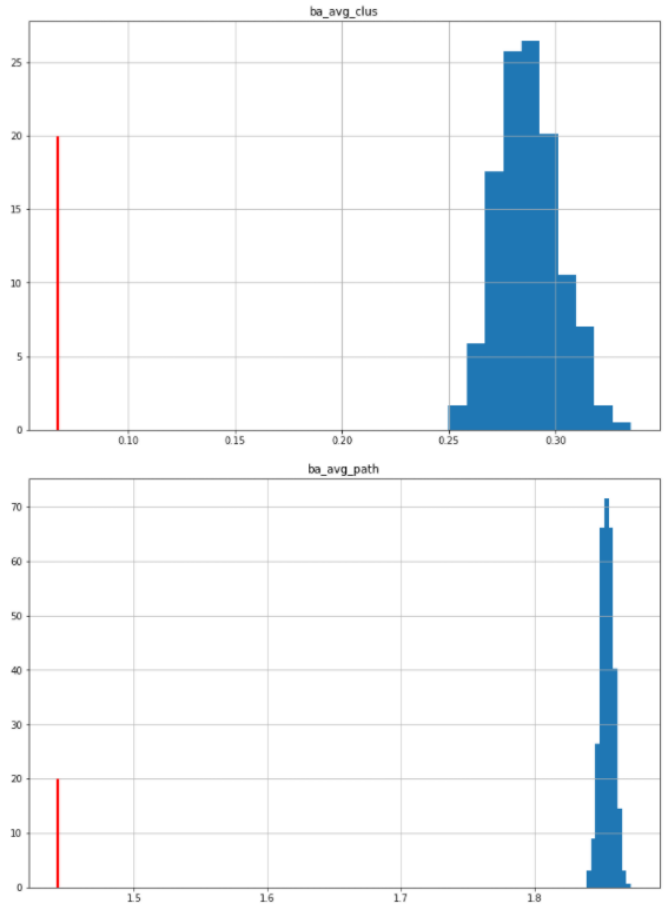

Fig. 25: Coeficiente de Clustering y longitud media - Scale Free y Gsww

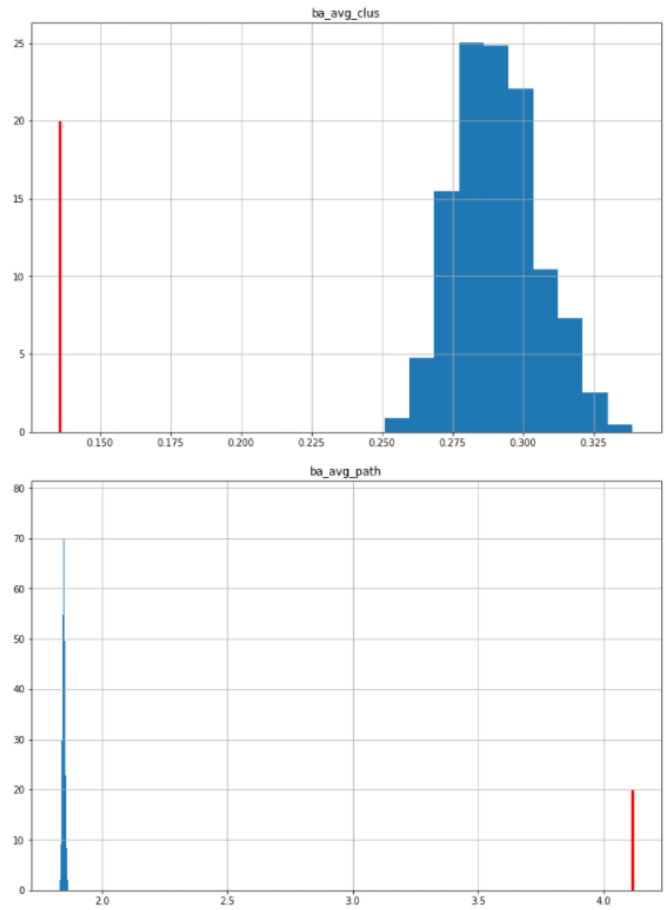

Fig. 26: Coeficiente de Clustering y longitud media - Scale Free y Gw2v

IV. Conclusiones

- La distribución de los coeficientes de clustering resulto ser similar para ambas redes Gsww y Gw2v: Muy pocos nodos con vecinos muy conectados entre si y muchos nodos con vecinos poco conectados.
- El coeficiente de clustering global para Gw2v es el doble que Gsww, lo cual indica que Gw2v tiene vecinos mas conectados entre si que Gsww.
- La distribución de grados indica que hay palabras que existen en Gsww pero no existen en Gw2v. Por otro lado, se aprecia que se tienen muchos nodos de bajo grado. Esto ultimo se da en mayor medida en Gsww.
- El coeficiente de asortividad indica que Gsww es una red asociativa, mientras que Gw2v es una red disociativa.
- Ambas redes tienen el mismo número de aristas de camino mínimo y diámetro. La densidad de Gw2v es el doble que Gsww
- Utilizando los algoritmos de detección de comunidades Girvan-Newman para una versión submuestreada de 400 nodos se pudieron encontrar comunidades muy similares entre sí de acuerdo a lo establecido por el índice Rand ajustado.
- Se observa que ambas redes (Gsww y Gw2v) no presentan características similares a redes del tipo Small-world o Scale-Free

References

[AHA+08] Aric A., Hagberg, Daniel A., Schult, and Pieter J. Swart. Exploring network structure, dynamics, and function using networkx. *in Proceedings of the 7th Python in Science Conference (SciPy2008), Gäel Varoquaux, Travis Vaught, and Jarrod Millman (Eds), (Pasadena, CA USA)*, pages 11–15, 2008.

[Bar13] Albert-László Barabási. Network science. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1987):20120375, 2013.

[CFM21] C. Collado, F. Felicioni, and A. Marino. Link a TP de Redes de palabras. https://github.com/magistery-tps/dm-cyt-tp2/blob/main/notebooks/analysis.ipynb, 2021. Grupo 3 - DMCT.

[DDNP+19] S. De Deyne, D. Navarro, A. Perfors, M. Brysbaert, and G. Storms. The "small world of words" english word association norms for over 12,000 cue words. *Behavior Research Methods*, 51, 2019.

[ECBS09] Martin Elias Costa, Flavia Bonomo, and Mariano Sigman. Scale-invariant transition probabilities in free word association trajectories. *Frontiers in integrative neuroscience*, 3:19, 2009.

[JKM06] Michael N Jones, Walter Kintsch, and Douglas JK Mewhort. High-dimensional semantic space accounts of priming. *Journal of memory and language*, 55(4):534–552, 2006.

[KB21] J. Kamienkowski and L. Bavassi. Tp2: Redes de palabras. *Data Mining en Ciencia y Tecnología*, November 2021.

[SC02] Mariano Sigman and Guillermo A Cecchi. Global organization of the wordnet lexicon. *Proceedings of the National Academy of Sciences*, 99(3):1742–1747, 2002.