Comparación de redes de palabras mediante grafos

CARRASCO, Lisandro T.*, DAZA CARO, Yudy Carolina*, GALÍNDEZ MARTÍNEZ, Raúl*, OPAZO, F. Ayelén*
* MAESTRÍA EN EXPLOTACIÓN DE DATOS Y DESCUBRIMIENTO DEL CONOCIMIENTO
FACULTAD DE CIENCIAS EXACTAS Y NATURALES UBA
Buenos Aires, Argentina

***Abstract*—Para el desarrollo de este estudio se utilizó la base de datos SWOW-EN2018: Preprocessed, que contiene palabras en inglés y fue preprocesada para trabajar con una submuestra óptima. El objetivo consistió en comparar el grafo que surge a partir de los datos recolectados durante el experimento *Small World of Words* con el que resulta de la distancia semántica entre las palabras medida con *word2vec*. Entre los resultados más interesantes se encontraron importantes diferencias entre las estructuras y las características que surgen de grafos construidos de manera diferente aún cuando parten de los mismos datos.**

***Index Terms*—words, networks, graphs**

I. INTRODUCCIÓN

Para el estudio de un sistema complejo, como es la lingüística, pueden aplicarse herramientas de análisis de redes y el análisis de grafos. Es posible realizar inferencias acerca de las representaciones mentales de nuestro diccionario de palabras a partir de tareas muy sencillas de asociación semántica y de memoria. Los modelos más sencillos, asumen que las palabras están representadas por nodos y que son conceptos interconectados. Por ejemplo, las palabras que están más relacionadas, estarán a una menor distancia en este diccionario mental, o las palabras polisémicas resultan ser *hubs*, o atajos, que conectan conceptos alejados dando una estructura de *small-world* a la red de palabras.

En este trabajo, se utilizan los datos recopilados por el proyecto *Small World of Words*. Este proyecto experimentó con la asociación libre de palabras: casi 85 mil participantes tuvieron que asociar libremente una palabra clave a las primeras tres palabras que se le ocurrieran en respuesta a esta clave. En total, se manejaron más de 12 mil palabras claves y se obtuvieron más de 3.8 millones de palabras respuestas.

A partir de esta base, se analizarán distintas formas de construir redes de palabras, tanto a partir de las conexiones realizadas por los participantes en el proyecto *Small World of Words* como a partir de las distancias semánticas que pueden generarse a partir de embeddings del tipo *word2vec*, que permiten representar a cada palabra en un espacio vectorial, dentro del cual se puede definir la distancia coseno entre palabras.

La metodología utilizada en este trabajo consiste en la comparación de la red derivada del proyecto presentado y la red semántica de *word2vec*, mediante la caracterización de cada uno, de acuerdo al número de nodos y aristas, peso, conexión, diámetro, densidad, etc. Además, se realizan comparaciones en términos de coeficientes de clustering, distribución de grado, y centralidad, asortividad, entre otras. Se buscan también comunidades con el algoritmo Louvain, que permite obtener un rendimiento rápido y preciso. Finalmente, se comparan las dos redes en cuestión con otras redes prototípicas de diferentes tipos, evaluando las similitudes y diferencias de las redes de estudio con las redes simuladas para considerar si pueden ser categorizadas entre uno de los prototipos.

II. DATOS Y PREPROCESAMIENTO

En este estudio, se utilizó la base de datos preprocesada 'SWOW-EN2018: Preprocessed', que contiene información sobre los participantes del proyecto *Small World of Words* y las tres palabras respuesta que proporcionaron a cada palabra clave que les tocó. Esta base ya se encuentra previamente normalizada, revisada ortográficamente, con correcciones de mayúsculas y acorde al dialecto americano.

En total, se trata de un dataframe de más de 3.8 millones de filas con 13 columnas, aunque solo se seleccionan dos variables y se realizan una serie de pasos para limpiar las observaciones y reducirlas a una escala más manejable e interpretable:

- Se seleccionan solamente las variables 'CUE' y 'R1', correspondientes a la palabra clave que recibió el participante en cuestión y su primera respuesta libre.
- Se eliminan todas las observaciones con valores faltantes.
- Se eliminan todas las palabras con una longitud menor a 2 caracteres.
- Se eliminan las palabras que se encuentren en R1 pero no en CUE, así como también se eliminan las que estén en CUE pero no hayan sido mencionadas en R1.
- Se remueven los pares que tienen la misma palabra por CUE y R1 para evitar *loops*.
- Se filtran las palabras de CUE y de R1 que no estén en el *corpus* del archivo preentrenado de *Word2Vec*.
- Se calcula la frecuencia de aparición de cada par CUE-R1: el 95% de los pares tiene una frecuencia nominal de menos de 10 apariciones. Se excluyen todos estos pares para trabajar únicamente con el 5% de pares que muestra una asociación más firme, a partir de las 10 menciones.
- Se calcula la frecuencia del par relativo a la frecuencia de aparición de R1: un 25% de los pares tiene una frecuencia relativa menor a 0.02, también se eliminan estos casos para continuar con una base más manejable.

978-1-7281-4387-3/19/$31.00 © 2019 IEEE

El dataframe con el que se realizarán los grafos cuenta finalmente con 4.400 observaciones y 3 variables, que son los pares CUE-R1 únicos y sus frecuencias relativas.

III. Grafo SWW

El primer grafo se construye a partir de las palabras únicas obtenidas de los 4.400 pares mencionados previamente, cada palabra se toma independientemente de si sea CUE o R1 y se la considera un nodo de la red. Para definir las aristas, se tiene en cuenta la direccionalidad que surge de los pares, por lo tanto, es un grafo **dirigido** y **pesado**, dado que también tiene un peso basado en la frecuencia relativa. La red no presenta *loops*.

El grafo resultante contiene **3.335 nodos** y **4.400 aristas**. Para realizar un análisis de la red, se considera solamente la componente conexa débil más grande, que presenta 1.932 nodos y 2.789 aristas.

Esta primera red es graficada con los layouts ***Spring***, ***Fruchterman Reingold*** y ***circular***. En ninguno de los formatos puede apreciarse información detallada, dada la importante cantidad de nodos y aristas de la red.

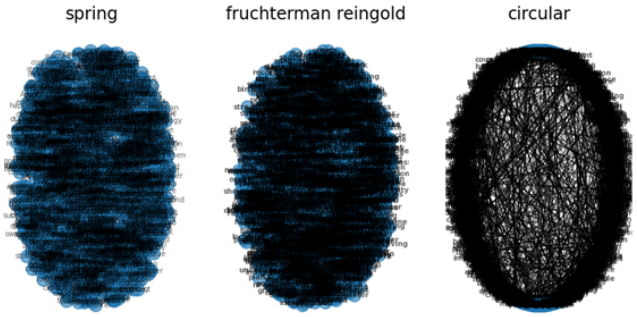

Fig. 1. Visualizaciones de *Small World of Words*.

Para mejorar la visualización del grafo y su información, se toma un subconjunto del grafo, es decir, un subgrafo donde todos los nodos tengan 5 vecinos.

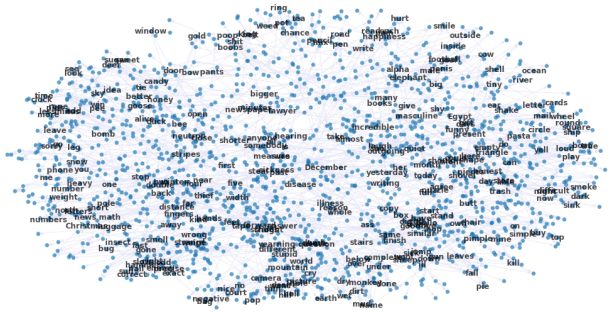

Fig. 2. Subgrafo de *Small World of Words*.

En este subgrafo pueden apreciarse mejor algunos otros detalles, por ejemplo, la cercanía entre palabras que suelen usarse conjuntamente en el lenguaje cotidiano (como *phone*, *me*, *you* y *number*) o que tienen funciones o significados similares (como *pencil* y *pen*).

También se realizó una reducción de la dimensionalidad para visualizar la red con otra perspectiva. Utilizando **TSNE**, puede apreciarse más claramente el agrupamiento de algunas palabras similares (como *freezing* y *refrigerator*) o algunas otras asociadas por la cultura popular (como *Rocky* y *climb*, indisociable de las escalinatas de la película de Stallone). Una observación curiosa que arroja esta reducción de la dimensionalidad es que *climb* se encuentra más cerca de Rocky que de otra palabra gramaticalmente mucho más similar, como *climber*; siendo esta una señal importante de la representación más vinculada al pensamiento libre de los participantes, fuertemente influenciado por la cultura estadounidense, que a la base semántica de las palabras.

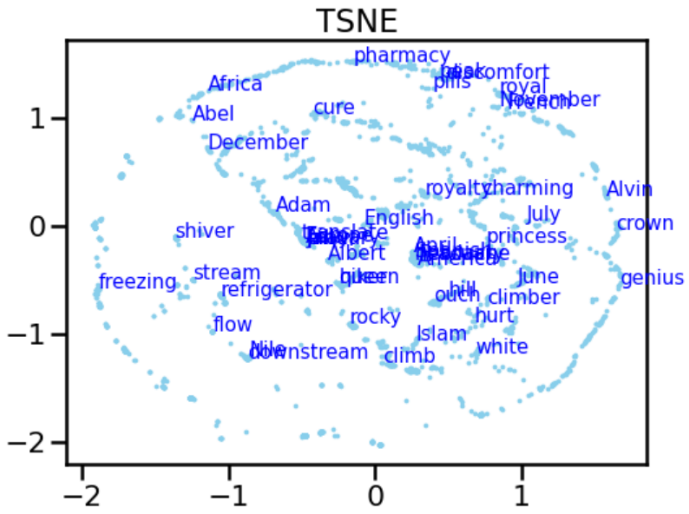

Fig. 3. Reducción de la dimensionalidad con TSNE.

IV. GRAFO W2V

La segunda red fue construida bajo otro criterio. Si bien se utilizó el mismo dataframe final de 4.400 pares con el que se construyó el Gsww, en esta ocasión las relaciones representadas se basaron en la proyección generada en base al embedding del tipo *Word2Vec* preentrenado a partir de un conjunto de GoogleNews. Las palabras del dataframe se transformaron en vectores y la relación entre nodos se dio por la distancia coseno calculada gracias a esta transformación.

Entonces, esta segunda red constituye un grafo **pesado**, cuyo peso se define por la distancia coseno, sin *loops* y completamente **conectado**, pero en esta ocasión es **no dirigido**, dado que no se mantiene el sentido de relación entre CUE y R1.

Si bien este grafo también cuenta con **3.335 nodos**, las **aristas son 555.944**, un aumento muy importante, dado el cambio en la forma en la que se define la relación entre los nodos.

A partir de una visualización del grafo con el formato *Fruchterman Reingold* , puede verse que también hay una alta

densidad y complejidad de la red que no permite apreciar su información.

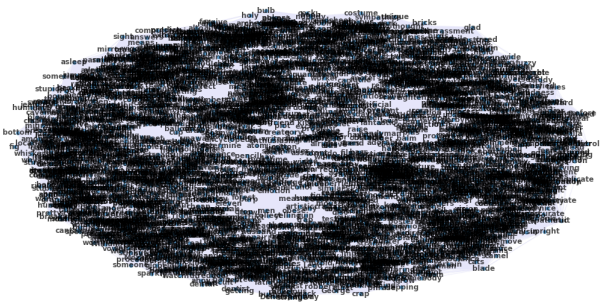

Fig. 4. Grafo de la red W2V con el formato Fruchterman Reingold

Aplicando la misma reducción de dimensionalidad que en el caso anterior, con la técnica de **TSNE**, puede apreciarse como algunas palabras similares se encuentran más cercanas que en la visualización de la red anterior. Por ejemplo, en el caso de *downstream* y *stream*.

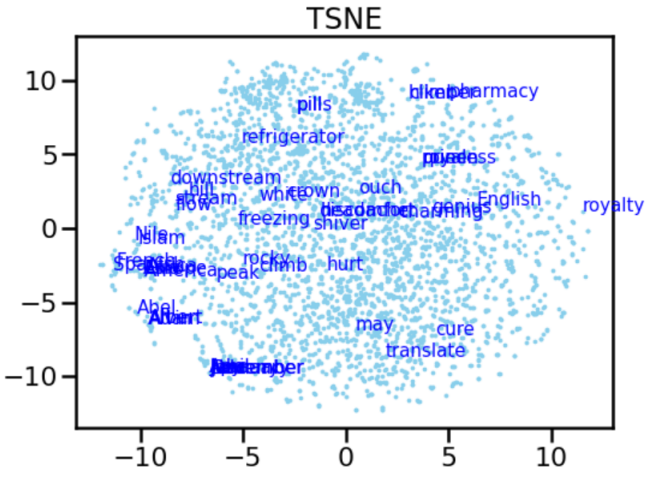

Fig. 5. Reducción de la dimensionalidad con TSNE.

V. CARACTERIZACIÓN DE LOS GRAFOS

Con el objetivo de comparar ambas redes, se calculan algunas métricas que las representan a cada una.

La **densidad** de las redes muestra una primera diferenciación importante: mientras que el primer grafo (de ahora en más, 'grafo SWW') presenta una densidad de 0.0007, el segundo grafo (de ahora en más, 'grafo W2V') tiene una densidad bastante mayor, de 0.0999. Esta métrica expresa la diferencia mencionada anteriormente respecto a la cantidad de aristas que tiene cada red, el enorme salto en la cantidad de aristas que se vio en el grafo W2V generó una red mucho más conectada, donde cada nodo tiene más conexiones con otros, mientras que los nodos en el grafo SWW tiene muchas menos conexiones efectivas de las que potencialmente podría tener (dada la direccionalidad particular de la interacción que surge de la forma de vinculación CUE-R1).

El promedio de los coeficientes de *clustering* de los grados, **el coeficiente de *clustering* global**, también es sensiblemente menor en la primera red, casi un 50% menor. Mientras que el grafo hecho en base a W2V presenta un coeficiente 0.100, en el caso de la red de SWW arroja un resultado de 0.057.

Esto puede interpretarse por la forma de interacción y de construcción que se da en cada red: mientras que en una red social (por ejemplo, en el experimento estudiado de la red de windsurfistas) los nodos -las personas- pueden asociarse e interactuar libremente entre todos los nodos, en este experimento los vínculos se forman de uno a uno, de una palabra clave con otra palabra respuesta. Estas "reglas de interacción" que propone el experimento hace que la interacción entre nodos sea muy baja: las palabras clave del proyecto *Small World of Words* se vincularán entre sí solo cuando los humanos del experimento apelen a una de ellas como respuesta de la otra; mientras que en el caso de *Word2vec* la similitud semántica entre nodos puede dar más chances de interacción entre palabras clave.

La **asortividad** también puede explicar una faceta del tipo de interacción que presentan los nodos en estas redes. En ambos casos, hay nulos niveles de asortividad: para el grafo SWW es de -0.084 y para el grafo W2V, de -0.002. Es decir, no puede pensarse en que haya un criterio de "selectividad" de asociación entre las palabras, dado que no responden a un comportamiento de relacionamiento humano, como en las redes sociales, o incluso natural. En ese sentido, los coeficientes de asortividad confirman que no se puede hablar de un criterio de vinculación entre nodos de mayor grado con otros similares, ni de menor grado en el mismo sentido.

En cuanto a los **grados** de los nodos de ambas redes, pueden verse grandes diferencias y de manera muy clara en sus distribuciones de grado. El rango de grados en una y en otra varía enormemente producto de la forma de confeccionar los grafos, de la forma de relacionarse entre cada nodo que surge de estas decisiones. La mayor densidad y la mayor cantidad de aristas que veíamos en la red W2V se expresa en los niveles de grados más altos que tiene cada nodo, es decir, al haber muchas más conexiones entre nodo y nodo en la segunda red, cada nodo presenta un grado más alto. Esta es una observación importante para el análisis del experimento y la comprensión de la libre vinculación de palabras: mientras que la relación entre una palabra y otra cuando es definida por un humano tiene un criterio más "selectivo" (hay menos conexiones entre palabras); cuando el criterio pasa a depender únicamente de la distancia semántica con W2V, esta selectividad se pierde, trastocando el sentido más asociativo y significativo del lenguaje, generando muchas más conexiones de las que la lógica lingüística de las personas generaría. Parte de esta diferencia en la intensidad de las relaciones, se aprecia en la diferencia que presentan las distribuciones de grado y de **pesos** de las redes.

Por otro lado, hay una diferencia importante en la distribución y no solo en la magnitud. Mientras que en la red SWW se puede ver un histograma asimétrico hacia la derecha que puede coincidir con la distribución de grado de una red

de mundo pequeño; en la red W2V el histograma muestra una distribución de grado muy similar a la normal, asociándose más a una red prototípica *random*.

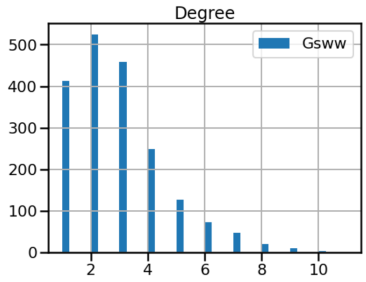

Fig. 6. Distribución de grados de la red *Small World of Words*.

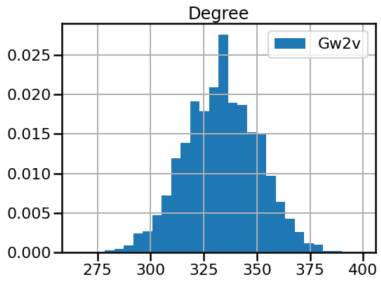

Fig. 7. Distribución de grados de la red *Word 2 Vec*.

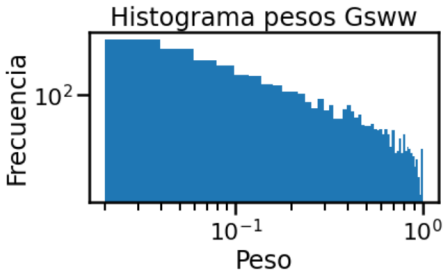

Fig. 8. Distribución de pesos de la red *Small World of Words*.

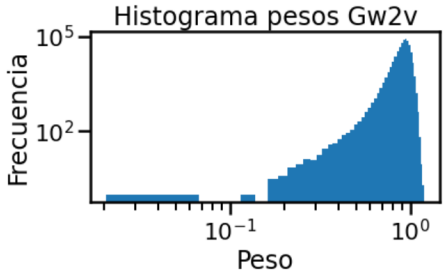

Fig. 9. Distribución de pesos de la red *Word 2 Vec*.

Para profundizar en la influencia o la importancia que tiene cada nodo dentro de la red, exploramos la **centralidad de grado** en las redes.

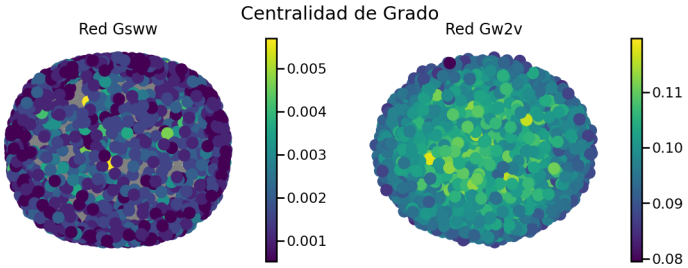

Fig. 10. Centralidad de grado comparada

Los diferentes niveles de centralidad de grado que presenta cada red expresan su diferente comportamiento interno, como cambia el rol local de los nodos en cada una. Mientras que las centralidades de los nodos de la red SWW alcanzan niveles generales más bajos con muy pocos nodos "centrales" o de mucha influencia (algo lógico, pensando en la menor cantidad de enlaces que tiene esta red producto de su interacción y confección), puede verse que los nodos en la red W2V presentan otra escala de valores, mucho más elevados en general, pero con menor variabilidad entre cada nodo; es decir, los nodos son más influyentes en esta red que en la otra, pero su valor de centralidad también es más homogéneo, desprendiendo que varios caminos pasan a través de varios nodos.

VI. COMUNIDADES

Para la detección de comunidades se aplica el **algoritmo de Louvain**. Dado que este algoritmo puede aplicarse únicamente en redes no dirigidas, se transformó la red SWW, que originalmente era dirigida, para poder continuar con el análisis comparado.

Mientras que en la red W2V se detectaron 8 comunidades, en la red SWW se detectaron 88. Una diferencia significativa que tiene que ver con el tipo de interacción y la cantidad de conexiones que tienen los nodos en cada red. Mientras que la mayor cantidad de enlaces entre nodos en la red W2V permite construir comunidades débiles más grandes donde hay una mayor cantidad de conexiones dentro de cada comunidad, en el caso de la red SWW la poca cantidad de enlaces entre nodos hace que la cantidad de comunidades sea mayor y que el tamaño de cada una sea considerablemente menor, dado la cantidad de vecinos con las que se van relacionando.

Comunidades de W2V

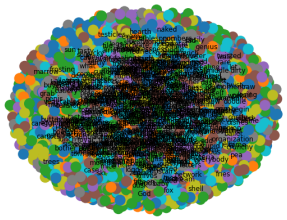

Fig. 11. Comunidades de la red *Word 2 Vec*.

Comunidades de SWW no dirigida

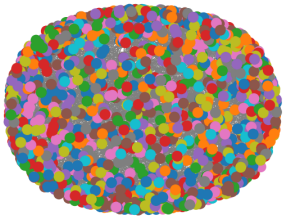

Fig. 12. Comunidades de la red *Small World of Words*.

Los grandes tamaños de la red dificultan tener una visualización clara de las comunidades detectadas, pero también es importante destacar que ante grafos de esta escala, el algoritmo encuentra importantes problemas de resolución, dado que se distorsionan las medidas y la modularidad siempre resultará chica.

Para matizar las dificultades presentadas ante esto, se observa la composición de la comunidad más chica detectada para cada red.

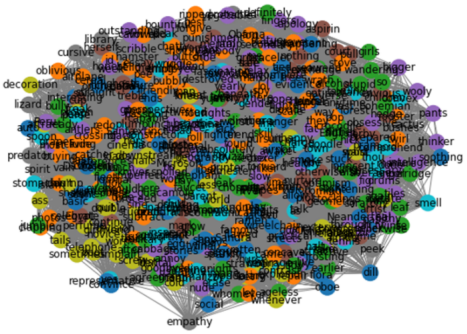

Fig. 13. Comunidad más pequeña del *Word 2 Vec*.

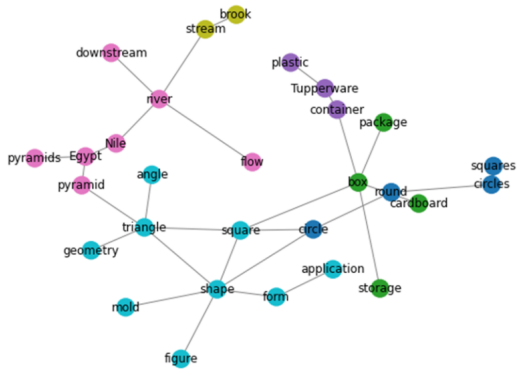

Fig. 14. Comunidad más pequeña del *Small World of Words*.

Si bien las comunidades de W2V siguen siempre de un tamaño elevado para tener una buena interpretación del resultado (la más pequeña contaba con 318 nodos), puede apreciarse levemente una mejor organización de pequeñas comunidades locales en ambas redes. Por su parte, la comunidad de SWW muestra mucha mayor claridad y se puede ver el sentido común que tiene cada agrupamiento realizado por Louvain.

El **índice Rand ajustado** presentó un valor de 0.00013, confirmado la gran disimilaridad que presentan las comunidades de SWW y W2V entre sí.

VII. Redes prototípicas

Con el objeto de comparar las redes obtenidas con las que se obtendrían con nodos de topología similar, pero conectados al azar y construidos con otros modelos, se realizan tres tipos de redes simuladas:

- De mundo pequeño, realizadas con el método de Newman-Watts-Strogatz.
- Aleatoria, construidas con el método de Erdos Renyi.
- Libre de escala, construidas con el método de Barabasi-Albert.

Cada tipo de red fue simulada 100 veces, con la misma cantidad de nodos que los SWW y W2V, así como una cantidad proporcional de nodos y aristas acorde a cada red, para poder comparar con la mayor similaridad posible la red en cuestión con sus prototipos simulados.

Para comparar las redes reales con las redes simuladas, se calcularon los coeficientes de *clustering* para evaluar el histograma de coeficientes de las redes simuladas respecto al *clustering* promedio tanto de SWW como de W2V.

Los resultados para la red SWW muestran una importante distancia entre el promedio de *clustering* de la red y los coeficientes de *clustering* en las redes libres de escala y aleatoria, pero una coincidencia con las redes de mundo pequeño.

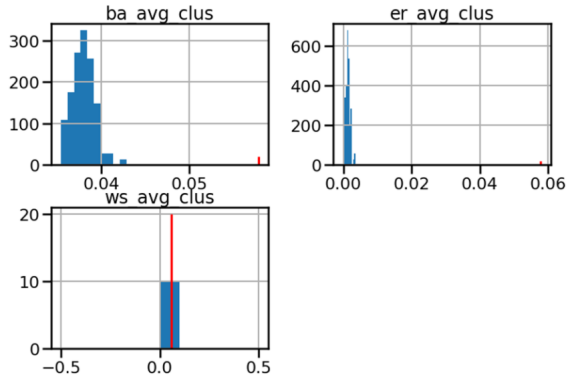

Fig. 15. Coeficientes de *clustering* de las redes

Si bien hay bibliografía previa que afirma que las redes de asociación libre como el grafo SWW analizado hasta ahora tienen una estructura de *small world*, esta comparación entre los coeficientes de *clustering* no resulta suficiente para afirmarlo, pero sí hay otros indicios a lo largo de este trabajo que podrían reforzar esta teoría, por ejemplo, la distribución

de grados mencionada anteriormente, que sí se asemeja a una distribución prototípica de *small world*.

Por su parte, la red W2V ya había presentado características similares a las de una red aleatoria con un histograma de grados muy similar al de una distribución normal.

Ahora, si bien la comparación entre coeficientes de clustering en este caso no permite tener una visión clara, sí es interesante comparar el histograma de caminos mínimos en esta red. En este caso, podemos observar una mayor similaridad de los caminos mínimos de W2V con los simulados en las redes de tipo *random*.

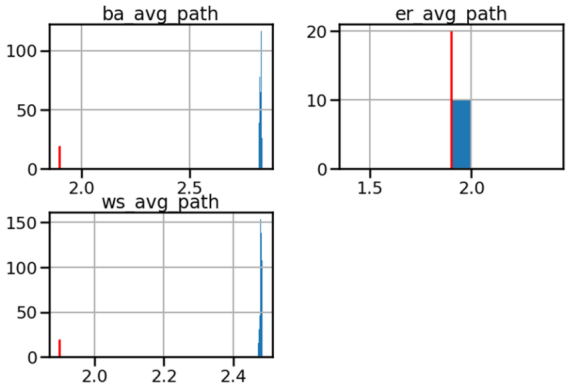

Fig. 16. Caminos mínimos

VIII. Conclusiones

El objetivo de este trabajo consistió en comparar las características de dos redes construidas de diferentes maneras, pero ambas sobre los datos del experimento *Small World of Words*.

Si bien ambas redes compartieron las mismas palabras y, por tanto, los mismos nodos, las distintas formas de confección de cada red generaron diferencias importantes en la información que arrojó cada grafo. En términos generales, puede decirse que son redes disímiles en aspectos estructurales: mientras que la red SWW se asemeja más a un prototipo de red de mundo pequeño, la red W2V es más similar a una red *random*.

Esto implica diferencias que se apreciaron a lo largo de toda la exploración y que se basan en las formas de generar sus relaciones y evaluar su interacción (en la distancia coseno de W2V y las formas de calcular los pesos). Por ejemplo, diferencias en la cantidad de aristas, en los histogramas de grado de cada red, en sus coeficientes de *clustering* y en sus comunidades.

Pero, por otra parte, un punto en común de ambas redes a destacar es la baja asortividad de las dos presentaron. Esto confirma que, en términos de análisis del lenguaje en grandes escalas, ya sea en base a respuestas espontáneas o a análisis semánticos, no se puede hablar de características jerárquicas en el idioma inglés.

En definitiva, estas características diferentes nos permiten entender que la libre asociación de conceptos realizada por humanos tiene una lógica muy diferente a la que puede desprenderse de distancias semánticas a partir de *emdeddings*.

Más allá de las diferencias o los puntos en común de cada red, se aprecia que las técnicas de análisis de grafos permiten obtener información relevante para el estudio del lenguaje y su conceptualización en las personas.

REFERENCES

[1] Barabási, A.-L., et al. Network science. Cambridge university press, 2016.

[2] Elias Costa, M., Bonomo, F., and Sigman, M. Scale-invariant transition probabi- lities in free word association trajectories. Frontiers in integrative neuroscience 3 (2009),19.

[3] Jones, M. N., Kintsch, W., and Mewhort, D. J. High-dimensional semantic space accounts of priming. Journal of memory and language 55, 4 (2006), 534–552.

[4] Mikolov, T., Chen, K., Corrado, G., and Dean, J. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013).

[5] Sigman, M., and Cecchi, G. A. Global organization of the wordnet lexicon. Procee- dings of the National Academy of Sciences 99, 3 (2002), 1742–1747.

[6] Steyvers, M., and Tenenbaum, J. B. The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. Cognitive science 29, 1 (2005), 41– 78.

[7] https://smallworldofwords.org/en/project/home