Análisis de grafos aplicados a redes de palabras

Fabiana Alejandra Rossi Magali Rodrigues Pires Eliana Inés Padula Álvaro López Malizia

RESUMEN

El siguiente trabajo práctico resulta de la comparación entre grafos construidos a partir de interacciones entre palabras. Por un lado, se encuentran las relaciones entre palabras que surgen a partir de un experimento de asociaciones libres y por el otro, las distancias semánticas calculadas entre las mismas palabras (el embedding que se utilizó es word2vec).

Para ambos conjuntos de interacciones, se graficaron los grafos correspondientes y se realizó su caracterización. Se efectuaron comparaciones a partir del número de aristas, las medidas de centralidad, el diámetro de la red, las conexiones entre sus elementos y detección de comunidades.

Finalmente, se realizó la construcción de redes prototípicas para el conjunto de palabras devenidas de la experimentación.

El procesamiento de los datos y confección de figuras se realizó utilizando el siguiente código: colab.

I. Introducción

Las redes se caracterizan por ser sistemas complejos en donde diversos elementos se encuentran unidos o interaccionan entre sí. Para cada sistema, existe una representación en forma de red que define las relaciones entre sus componentes.

Existen múltiples ejemplos de sistemas intrincados que abarcan desde la biología,la lingüística, la sociología, la academia y hasta los sistemas de computación y de comunicación contemporáneos (Igual, L. Segui, S. (2017)). A modo de ejemplo, los sistemas biológicos son por excelencia sistemas de complejidad, en donde existen interacciones con el medio externo, rutas metabólicas, procesamiento de señales químicas y eléctricas. Además, en estos procesos se manifiestan propiedades emergentes como el resultado de estas interacciones.

En la actualidad, la sociedad humana se sostiene funcionalmente sobre redes crecientes en complejidad, que cooperan coherentemente entre sí. Desde sistemas eléctricos, señalización de ciudades, posicionamiento global o gestión de residuos. Cada una de estas redes es susceptible de fallar, dando como resultado fallas en cascada en donde, uno o más sistemas aparentemente independientes entre sí son afectados de forma catastrófica (Albert-Laszlo Barabasí 2006).

El objetivo del estudio de las redes, puede ser tan diverso como su composición. Se pueden estudiar para encontrar relaciones entre sus elementos, determinar formas eficientes de recorrerlas o buscar sus puntos débiles para evaluar su robustez frente a ataques.

Con el propósito de abordar el análisis de las redes, su procesamiento y generación de información, se realizó el siguiente trabajo práctico en donde se aborda la lingüística como campo de aplicación para este tipo de herramientas matemáticas y computacionales.

A través de reglas de asociación, pueden inferirse experimentalmente las conexiones entre las distintas palabras que conocemos y usamos de forma rutinaria (Elias Costa, M.). Un modelo sencillo, asume que las palabras pueden representarse como nodos en una red. Los vértices y la distancia entre cada uno de los nodos de esta red mental de palabras, se calculan mediante el proceso de evocación de un ”palabra respuesta” o ”R1”, a partir de una ”palabra entregada” o ”cue”. Aquellos elementos de la red mejor interconectados mediante su relación, surgirán naturalmente con mayor frecuencia y representarán una menor medida de distancia entre sí (Jones, M. N).

Este trabajo utiliza uno de los conjuntos de datos recopilados por el proyecto Small World of Words 1. En dicho experimento, a cada participante le aparece una palabra, y debe completar hasta tres palabras relacionadas. De esta manera se construye una red de asociaciones libres, donde cada palabra es un nodo, y las aristas surgen de las conexiones realizadas por los participantes. Hay distintas formas de construir la red que se analizarán a lo largo del trabajo. A su vez, es posible definir una distancia semántica entre palabras a partir de lo que se conoce como embeddings. Los embeddings permiten representar a cada palabra en un espacio vectorial, dentro del cual se puede definir una distancia entre palabras. Esta distancia es típicamente la distancia coseno, y el embedding que se utilizará en el trabajo es word2vec (Mikolov, T. 2013). Estas distancias se pueden utilizar tanto para interpretar los datos directamente (mirando por ejemplo la cohesión de las comunidades), como para construir una segunda red y compararlas.

II. Datos

Se obtuvieron los datos pertenecientes al proyeto Small Word of Words. Se utilizó el dataset de ”SWOW-EN2018: Preprocessed” que contiene palabras en idioma inglés. Este

dataset fue previamente preprocesado, en donde eliminaron los caracteres raros y palabras sin sentido.

Por otra parte, se utilizó el word2vec, perteneciente a Google. El mismo consiste en una representación vectorial de un modelo de palabras, obtenidas a partir de dataset de noticias de Google (con una cantidad de 100 billones de palabras). El modelo resultante resulta de un vector con 300 dimensiones y 3 millones de palabras y frases (https://code.google.com/archive/p/word2vec/).

III. Metodología

A. Preprocesamiento

Los datos fueron preprocesados. Inicialmente, se decidió trabajar solo con R1, es decir, la primera palabra que la persona respondió ante el estímulo (cue) durante el experimento. Luego, se eliminaron las palabras cuya longitud fuese menor a dos letras, los ”stopwords” y aquellos términos que no aparecieran de forma simultánea en cue y R1. Con el objetivo de conservar sólo las respuestas más frecuentes, se eliminaron primero todos las palabras utilizadas como estímulo (cue) cuya frecuencia se encontrara por debajo del percentil 98.5. Dicha operación se repitió para los pares de palabras normalizados por la frecuencia de cue (cue-R1/cue) que se encontraran por debajo del mismo percentil. Finalmente, se constató que aquellas palabras se encontraran dentro del cuerpo de word2vec.

Para la construcción del grafo de distancias semánticas (word2vec) se conservaron solo las palabras que resultaron del preprocesamiento anterior.

B. Elementos conectados

De manera complementaria, se realizó un filtrado mediante el cual se generó un subgrafo dónde sólo los elementos conectados fueron considerados.

C. Visualizacion

Los grafos fueron representados mediante la configuración ”Spring”.

IV. Tarea 1: Construcción de los grafos

Para el conjunto de palabras seleccionadas del experimento Small World of Words, se construyó un grafo pesado y no dirigido, donde los nodos son las palabras de cue y R1; las aristas, la presencia de los pares cue-R1; y los pesos, la frecuencia de cue-R1/cue. El grafo (Gsww) resultante tiene 219 nodos y 221 aristas (Figura 1).

Para el segundo conjunto de palabras, se calculó la matriz de similaridad y luego la matríz de adyacencia a partir de la cual se construyó un grafo pesado y no dirigido (Gw2v). La primera versión del mismo tenía 219 nodos y 23.871 aristas. Se decidió aplicar un filtro para conservar sólo los enlaces que refieran a distancias entre nodos mayores a 0.92. El número de aristas se redujo a 1918 (Figura 2).

A continuación, se describen brevemente ambos grafos (Tabla 1).

Grafos	Gsww	Gw2v
Número de nodos	219	129
Número de aristas	221	1918
Dirigido	No	No
Pesado	Si	Si
Loops	No	No

Tabla 1: *Características de los grafos*

A. Gsww: Small World of Words

Figura 1: *Representación de grafo de tipo Spring para el dataset obtenido a partir de Small World of Words*

Se observa de manera sencilla la presencia de subcomunidades dentro de la red.

B. Gw2v: word2vec

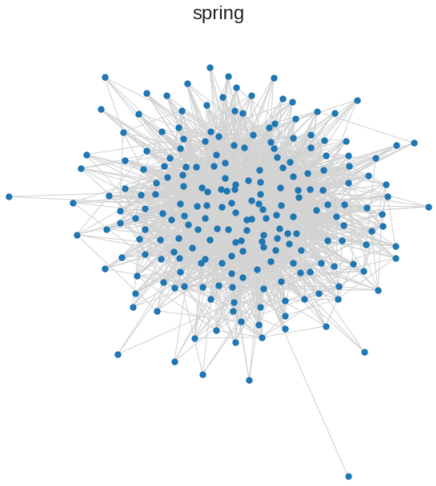

Figura 2: *Representación de grafo de tipo Spring para el dataset obtenido a partir de word2vec.*

La red no presenta una estructura clara, y tampoco se desprenden comunidades fácilmente a partir de la visualización de ésta.

TAREA 2: CARACTERIZACIÓN DE LOS GRAFOS

En este apartado se realizó una caracterización de los grafos construidos en función de su distribución de grado, diámetro, densidad, camino mínimo promedio, medidas de clustering, asortatividad y medidas de centralidad.

En primer lugar, se realizó un análisis de la distribución de grados de cada uno de los grafos estudiados (Figura 3). Si bien ambos grafos son pesados y no dirigidos, el resultado de la Figura 3 muestra una clara diferencia en sus estructuras.

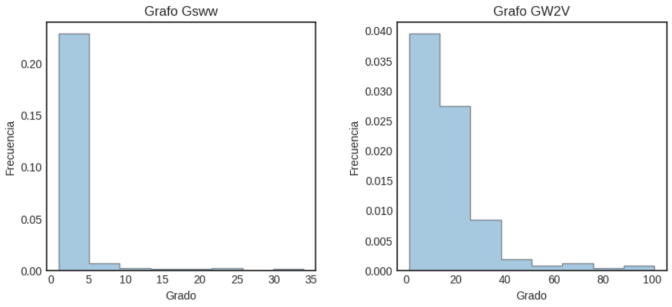

Figura 3: *Distribución de los grados pertenecientes a los nodos de los grafos Gsww (izquierda) y GW2S (derecha).*

En el caso de Gsww, la mayoría de los nodos tienen menos de 10 enlaces cada uno y se observan algunos pocos bastante más conectados (hubs) que tienen entre 20 y 25 o entre 30 y 35 enlaces. Este tipo de red, que se caracteriza por tener pocos nodos muy conectados y el resto de los nodos con grados bajos o medios, se denomina "libre de escala". En comparación, el grafo Gw2v está bastante más conectado: la mayoría de los nodos tienen hasta 40 enlaces cada uno y se observa una mayor cantidad de hubs que en el grafo anterior. En este caso, existen nodos fuertemente conectados que tienen hasta 100 enlaces respectivamente.

A continuación, se calcularon distintos parámetros para caracterizar las redes encontradas. Los resultados se muestran en la Tabla 2.

Medidas	Gsww	Gw2v
Diámetro	19	4
Densidad	0.0093	0.0803
Clustering promedio	0.0051	0.0564
Asortatividad	-0.4769	-0.3679
Camino mínimo promedio	8.26	2.093

Tabla 2: *Medidas principales de las redes de Gsww y Gw2v*

Los resultados de la Tabla 2 indican que la red Gw2v presenta un menor diámetro, mayor densidad y clustering promedio que la red Gsww, lo cuál indica que tiene mayor potencialidad de conexión entre nodos. Se trata de un grafo más compacto e interconectado, lo cuál indica que es más probable que dos palabras estén conectadas en la red semántica Gw2v que en la red de asociaciones analizada (Gsww).

La asortatividad es una medida que permite analizar si los nodos de la misma clase se agrupan sólo entre ellos o se agrupan con nodos de distinta clase. Ambos grafos tienen coeficientes negativos (Gsww: -0.47 / Gw2v: -0.36), lo cual indica que se trata de redes no selectivas (disortativas), donde los vértices con alto grado se conectan preferencialmente con nodos de bajo grado.

A continuación, se estudiaron distintas medidas de centralidad de nodo, que permitieron comparar los grafos Gsww y Gw2v. Dichas medidas indican la relevancia de un nodo en una red. La misma suele estar dada por el modo en que transmite la información, es decir, cuánto colabora en el flujo de la red y con su cercanía al resto de los nodos (lo cuál refiere a la noción de cohesividad del grafo). Las medidas de centralidad que se calcularon para ambos grafos son: centralidad de grado, donde el nodo más importante será el de mayor grado (los hubs); la centralidad de cercanía, que es una medida basada en la distancia, donde los nodos que tomen mayores valores son los que logran transmitir información más rápidamente; la centralidad de intermediación (betweenness), donde más importante será el nodo cuantos más caminos cortos pasen por él; y la centralidad de autovalor, donde el nodo más relevante será el que tiene un mayor autovalor, es decir, será el nodo de mayor grado cuyos vecinos también tengan mayor grado. Dichos resultados pueden encontrarse en el url provisto anteriormente, el cuál contiene el link al Google colab utilizado para generar las redes y sus caracterizaciones. En la Figura 4 se muestra el resultado del cálculo de clustering por intermediación.

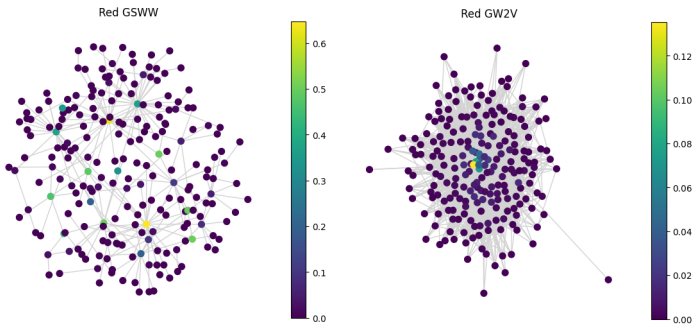

Figura 4: *Gráfico de redes Gsww (izquierda) y Gw2v (derecha), en donde los nodos se colorean de acuerdo a su valor de coeficiente de clustering de intermediación.*

Se puede observar que los nodos más relevantes según la centralidad de grado, son en general los mismos que cobran mayor relevancia según la centralidad de intermediación (betweenness), lo que indica que los nodos por los que pasan mayor cantidad de caminos cortos son al mismo tiempo los más conectados. Es probable que se trate de palabras polisémicas, que permiten conectar conceptos disímiles.

En el caso de la medida de centralidad de cercanía (ver url colab), observamos que en general la mayoría de los nodos de ambas redes tienen valores distintos de cero. Lo contrario ocurrió en el caso de la centralidad de autovalor: para el grafo Gsww sólo un nodo se destacó por su relevancia, mientras que en el caso del grafo Gw2v, se destacaron los nodos que obtuvieron valores más altos al calcular las medidas de

centralidad de grado e intermediación.

TAREA 3: COMUNIDADES

En un paso siguiente, nos propusimos evaluar la presencia de comunidades en las redes generadas. Para ello, se utilizó el algoritmo de Louvain. Este algoritmo utiliza la modularidad como estrategia para detectar comunidades conectadas y densas.

El resultado de la aplicación de dicho algoritmo puede observarse en las Figuras 5 y 6.

Las comunidades encontradas en la red asociativa (Gsww) fueron diez y están bien definidas (Figura 5).

Figura 5: *Gráfico de la red Gsww. Los nodos están coloreados de acuerdo a la identidad de las distintas comunidades encontradas por el método de Louvain.*

Por otro lado, se representó a la red semántica (Gw2v) y se la coloreó de acuerdo a las comunidades obtenidas con el método de Louvain. Asimismo, se comparó la distribución de colores en la misma red, pero considerando las comunidades encontradas para el grafo Gsww (Figura 6). Se puede observar que la distribución de las comunidades en la red Gw2v no es coincidente con las comunidades encontradas en la red Gsww. Asimismo, el método de Louvain encuentra menor número de comunidades en la red de Gw2v respecto la red Gsww.

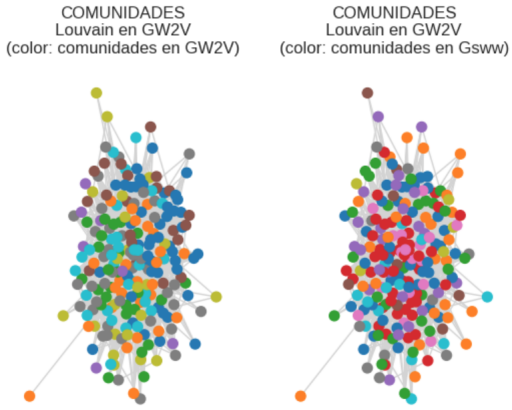

Figura 6: *Gráfico de red Gw2v. Los colores representan las comunidades encontradas mediante el método de Louvain en la red Gw2v (izquierda) o Gsww (derecha).*

C. Opcional 4: Agrupamiento por MDS TSNE

A continuación generamos una visualización de las palabras de la red Gw2v mediante las técnicas de MSD (Multidimensional scaling) y TSNE (t-Distributed Stochastic Neighbor Embedding) (Figura 7).

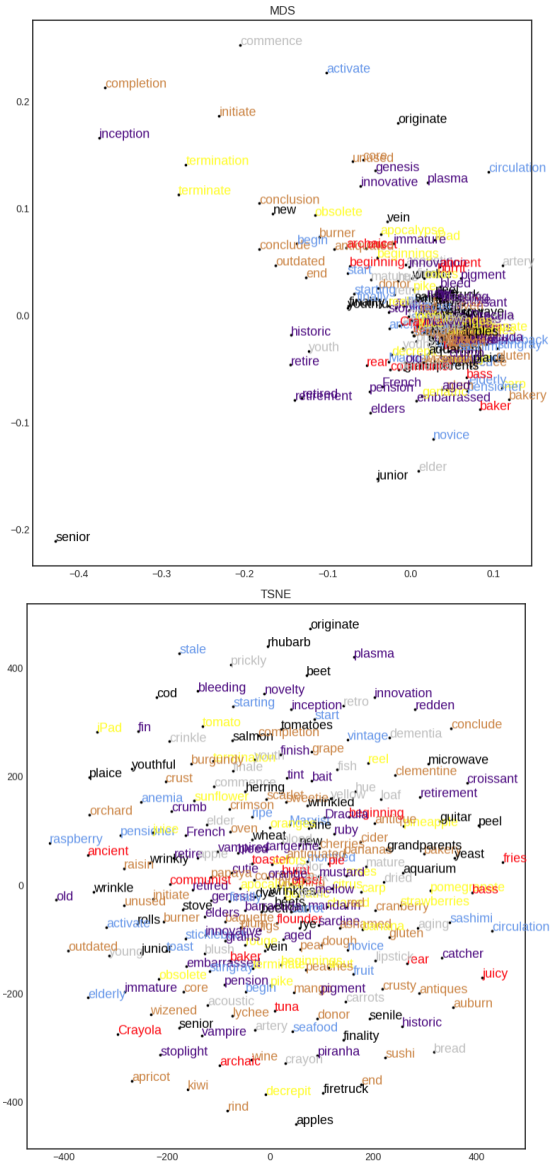

Figura 7: *Gráfico de palabras de la red Gw2v, agrupadas según distancias a través de MDS (arriba) o TSNE (abajo). Cada color representa una comunidad.*

Como consecuencia de la gran cantidad de datos, existe superposición de los términos y se dificulta su interpretación. Sin embargo, se puede observar que las comunidades (representadas con distintos colores) muestran una lógica dentro de su identidad. En la representación de MDS, se observa la presencia de una comunidad con los elementos ”retired”, ”retirement”, ”elders”, ”pension”, ”retire”. Este agrupamiento tiene coherencia por representar elementos relacionados con la jubilación o el retiro laboral. Los mismos se encuentran cer-

canos en la representación de las dos primeras componentes de MDS. Lo mismo ocurre para los términos ”conclude”, ”begin”, que una misma comunidad y se relacionan semánticamente con la duración de actividades.

La representación de los datos en las componentes de TSNE (Figura 7) no muestra un claro agrupamiento espacial de los términos.

Tarea 4: Small-world y redes prototípicas.

Por último, se compararon las medidas de clustering y camino promedio contra redes prototípicas definidas según Barabasi Albert, Erdos Rényi o Newman, Watts Strogatz. Para ello se generaron 1000 redes prototípicas con el mismo número de nodos y aristas que las redes de este trabajo, a fin de evaluar si las características medidas en Gsww y Gw2v podrían pertenecer a dichas distribuciones. Los resultados se indican en las Figuras 8, 9 y 10.

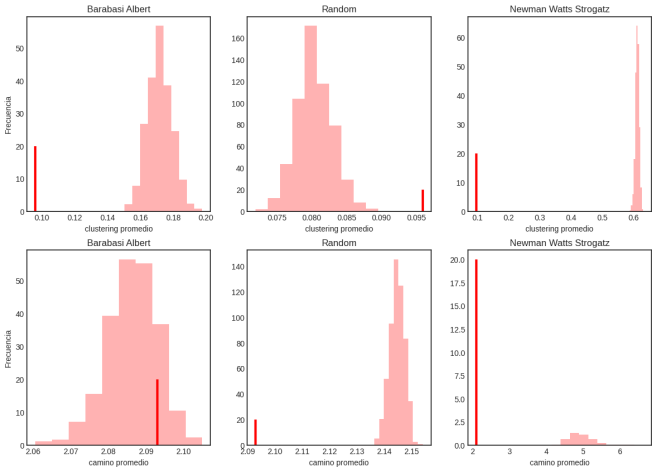

Figura 8: *Distribución de clustering y camino promedio de 1000 redes prototípicas de Barabasi-Albert, Random (Erdos-Rényi) o Newman-Watts-Strogatz. La línea sólida roja representa el clustering y camino promedio de la red Gw2v.*

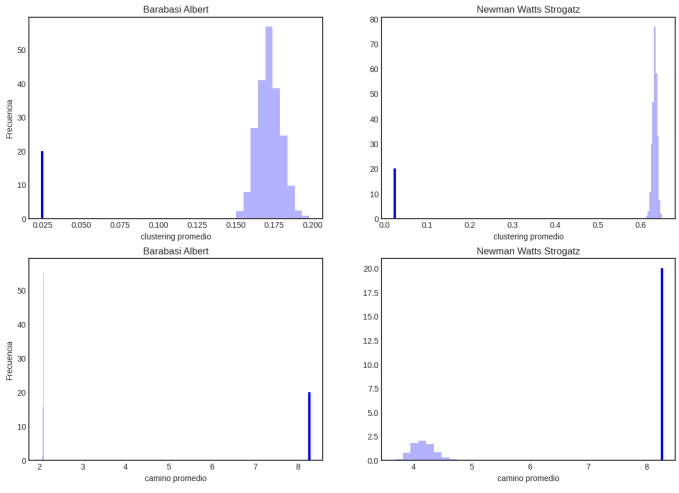

Figura 9: *Distribución de clustering y camino promedio de 1000 redes prototípicas de Barabasi-Albert, Random (Erdos-Rényi) o Newman-Watts-Strogatz. La línea sólida azul representa el clustering y camino promedio de la red Gsww.*

Para las redes Gw2v y Gsww podemos observar que sólo la medida de camino promedio de la red Gw2v es comparable a una de las redes prototípicas (Barabasi-Albert - red libre de escalas).

Por otro lado, dado que las redes de small world se caracterizan por ser redes de bajos caminos promedio y altos clustering promedios se puede llegar a concluir que si comparamos Gw2v contra la red Random (Erdos-Rényi), este concepto se estaría cumpliendo, y no así si se compara la misma contra la red libre de escala de Barabasi Albert y la red Newman Strograz. En cuanto a la red de asociación Gsww, la misma no tiene un camino promedio más pequeño ni un clustering más elevado que las redes de Barabasi Albert y la red Newman Strograz.

Estos resultados se condensaron en la Figura 10. Se representaron el camino promedio y el clustering promedio de las 1000 redes generadas para cada red prototípica.

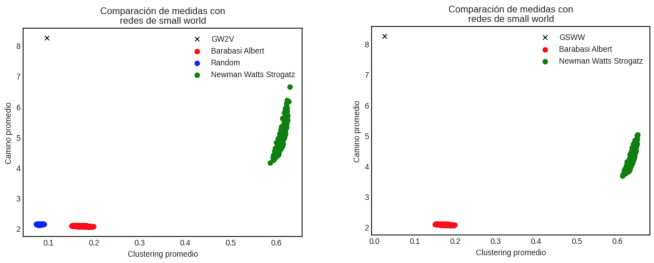

Figura 10: *Distribución de clustering promedio y camino promedio de 1000 redes prototípicas de Barabasi-Albert, Random (Erdos-Rényi) o Newman-Watts-Strogatz. La cruz representa las redes Gw2v y Gsww en los gráficos de izquierda y derecha, respectivamente.*

El análisis integral de la Figura 10 indica que las medidas de las redes Gsww y Gw2v no son compatibles con las redes prototípicas evaluadas.

Discusión

Durante el desarrollo de este trabajo se utilizaron técnicas para familiarizarse con el manejo de las redes. Particularmente, se exploró en forma individual y comparativa dos redes de palabras. Por un lado, se generó a una red de conexiones de términos relacionados por palabras estímulo y respuestas pertenecientes al proyecto Small Word of Words. En contraposición, la otra red fue generada a partir de distancias semánticas entre de palabras provistas por Google (el embedding que se utilizó es word2vec).

La visualización de ambas redes, puso en evidencia la presencia de nodos sobre cuya conexión se sostiene gran parte de los elementos que componen el grafo Gsww. De forma diferencial, la distribución de los elementos y las conexiones en el grafo Gw2v es más azarosa y no permite detectar nodos de mayor importancia.

Esta observación, permitiría explicar las diferencias entre los diámetros calculados para las redes. La longitud diametral de Gsww resulta mayor que el de Gw2v. En este escenario, es

plausible pensar que existen múltiples caminos de conexiones en la segunda red que acortan la distancia entre sus nodos más distantes. En el caso del grafo de Gw2v, existirían nodos "hubs" que generan atajos, permitiendo acortar las distancias entre los centros. En cambio, para Gsww, como resultado de estos nodos centrales, existe la necesidad de atravesar estos centros para moverse entre los extremos de toda la red.

De forma análoga, se puede argumentar que la densidad calculada es cualitativamente inversa a la diferencia en el diámetro. Mientras que para la red de Gsww, la densidad es mayor, debido a la gran cantidad de ramificaciones, la Gw2v es menor debido su estructura más compacta y de mayor cantidad de conexiones.

Adicionalmente, el camino mínimo promedio calculado para ambas redes, refuerza las observaciones realizadas para el diámetro y la densidad. El camino mínimo que presenta Gsww es superior al de Gw2v. Nuevamente, la falta de conexiones entre los elementos que la componen, permitiría explicar estas diferencias.

Continuando con la caracterización de las redes, la tendencia al clustering resultó mayor para la red de Gw2v, mientras que fue menor para la Gsww. Para el coeficiente de clustering, Gw2v presenta vecinos más conectados entre sí. En comparación, en Gsww, los nodos cercanos tienen menor cantidad de aristas que los unen. La distribución de este último, se asemeja a cluster jerarquizado en donde hay sólo conexión con un nodo central y con los puntos terminales.

El índice de asortatividad, otra medida de caracterización de la red, fue negativa para las dos grafos. Este resultado nos indica que no existe una tendencia inherente al agrupamientos entre los nodos de una misma clase. La medida obtenida señala que aunque en el grafo de Gsww, existan nodos de mayor grado, las conexiones entre éstos no serían directas. En línea con los mencionado anteriormente, es probable que se trate de palabras polisémicas, que permiten conectar conceptos disímiles.

Finalmente, de acuerdo al análisis realizado, podemos afirmar que la red de palabras de Gsww presenta una estructura lógica intrínseca y sugiere la presencia de una relación asociativa. Por el contrario, las relaciones entre las palabras obtenidas a través del embedding de Google presentan una estructura más azarosa y menos organizada.

La organización de la red de palabras para la Gsww genera una estructura más susceptible a fallas, en donde la pérdida de una conexión puede aislar grandes porciones del grafo. Sin existir múltiples conexiones que faciliten el desplazamiento entre los elementos, la incapacidad para acceder a una palabra compromete ampliamente la integridad de la red.

En un sentido biológico, la caracterización de estas redes cognitivas, podría aportar información, y permitir el uso de estrategias frente en enfermedades neurodegenerativas en donde se pierde progresivamente elementos nodos de palabras y recuerdos.

En conclusión, el estudio y caracterización de las redes en general, permite comprender el comportamiento y las relaciones de sistemas complejos. En este trabajo práctico, a partir de conexiones entre palabras, pudimos caracterizar las estructuras de agrupamiento y especular sobre los procesos subyacentes que las generan.

Referencias

Albert-Laszlo Barabasí. Network Science, http://networksciencebook.com/

Igual, L. Segui, S. (2017). Introduction to data science. In Introduction to Data Science

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

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.

Sigman, M., and Cecchi, G. A. Global organization of the wordnet lexicon. Proceedings of the National Academy of Sciences 99, 3 (2002), 1742–1747.

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