TP2: Redes de palabras

M Luz Bavassi, Juan E Kamienkowski
Data Mining en Ciencia y Tecnología

12 de noviembre de 2021

La aplicación del análisis de redes en general, y del análisis de grafos en particular, para el estudio de sistemas complejos es un campo en constante ebullición [1]. La lingüistica, como tantos otros campos científicos, no escapa del alcance de estas herramientas matemáticas. A partir de tareas muy sencillas de asociación semántica y de memoria, se puede inferir la representación mental que tenemos del diccionario de palabras que utilizamos cotidianamente [2]. 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 [3], o las palabras polisemicas resultan ser *hubs* que conectan conceptos alejados [5] dando una estructura de *small-world* a la red de palabras.
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 contruye una red de asociaciones libres, donde cada palabra es un nodo, y las aristas surjen de las conexiones realizadas por los participantes. Hay distintas formas de contruir 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* [4]. 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.

1. Objetivo

Este segundo Trabajo Práctico por objetivo comparar el grafo que surge a partir de los datos recolectados durante el experimento *Small World of Words* con el que surge de la distancia semántica entre las palabras medida con *word2vec*.

[1] `https://smallworldofwords.org/en/project/home`

2. Estructura de los datos:

Dentro del sitio *Small World of Words* encontran distintos *datasets*, en este trabajo utilizaremos **SWOW-EN2018: Preprocessed** que contiene palabras en inglés que ya fueron limpiadas de carácteres raros y no palabras [2].

3. Tarea 1: Construcción de los grafos

3.1. G_{sww}: *Small World of Words*

Los datos que se cargan de la página tienen unas primeras columnas con metadata que en principio no se va a usar, la columna con el **CUE** y las tres respuestas **R1**, **R2**, y **R3**. R1 suele estar completa, no así R2 y R3, la sugerencia es tirar estas últimas dos columnas porque ya se tienen muchos datos para trabajar. Sin embargo puede ser interesante incluirlas, y en el trabajo de referencia lo hacen.

En cuanto a las filas, recomendamos tirar palabras cortas (de una letra por ej.), palabras que no estén como CUE Y R1, pares CUE-R1 que aparezcan menos que una cantidad de veces, o pares CUE-R1 cuya frecuencia sea muy pequeña respecto a la del CUE ($CUE - R1/R1$ pequeño). También pueden tirar si le aparecen NaN y, sugerimos fuertemente, filtrar las palabras que no aparecen en el corpus de *word2vec*.

Así como pueden tomar las respuestas R2 y R3, también pueden realizar otros filtros que consideren más adecuados.

Luego de esto, deben construir un *dataframe* con los pares únicos CUE-R1, y pueden agregar columnas con el número de apariciones del CUE, del par CUE-R1 y del cociente $CUE - R1/R1$. A partir de esta estructura pueden facilmente construir el grafo.

Describir este grafo (¿Número de nodos? ¿Número de aristas? ¿Es dirigido? ¿Es pesado? ¿Tiene loops? Etc) y visualizarlo. También, pueden visusalizar algún subgrafo dentro de la red.

3.2. G_{w2v}: word2vec

Se carga el *word2vec* prentrenado que les proveemos (pero pueden buscar otro), a partir de él es posible definir la matríz de adyacencia, y a partir de esta el grafo.

Filtrar sólo las palabras que aparecen en Gsww. Si es necesario pueden realizar los filtros que consideren más adecuados.

Describir este grafo (¿Número de nodos? ¿Número de aristas? ¿Es dirigido? ¿Es pesado? ¿Tiene loops? Etc) y visualizarlo.

4. Tarea 2: Carácterización de los grafos

Comparar los dos grafos en terminos de medidas de *cliques/clustering*, distintas medidas de centralidad[3], distribución de pesos, distribución de grado, asortatividad, camino mínimo,

[2] `https://smallworldofwords.org/en/project/research`

[3] No es necesario hacer todas, elegir la que consideren adecuada en este caso.

diámetro, etc.

¿Qué pueden decir de cada uno de los grafos? ¿Y qué pueden decir de la comparación entre ellos?

4.1. Opcional 1:

¿Cómo cambian los resultados si se considere G_{sww} como no-dirigido o no-pesado?

5. Tarea 3: Comunidades

Detectar comunidades en G_{sww} y G_{w2v} con el algoritmo que consideren más indicado, justificar por qué. Calcular el indice de modularidad en cada caso ¿Qué pueden decir de las comunidades?

Comparar las comunidades de ambas redes con el índice *rand* (u otra métrica que considere adecuada) ¿Son similares?

Visualizar uno de los grafos, y pintarlo según las comunidades del otro ¿Qué se puede decir?

También se puede analizar la resolución del algoritmo de por el cuál se detectan las comunidades. Para ello, pueden explorar estructuras dentro de alguna de las comunidades.

5.1. Opcional 2:

Utilizar distintos algoritmos y comparar los resultados.

5.2. Opcional 3:

¿Cómo cambian los resultados si se considere G_{sww} como no-dirigido o no-pesado? ¿Y para G_{w2v}?

5.3. Opcional 4:

Utilizar algoritmos de reducción de dimensionalidad como MDS, tSNE, o UMAP para ubicar las palabras en dos dimensiones y pintar según las comunidades identificadas. Se puede elegir un subconjunto de palabras para que sea más claro. También pueden proyectar en el espacio construido con *word2vec* y pintar según las comunidades identificadas en $Gsww$, o viceversa.

6. Tarea 4: *Small-world* y redes prototípicas

Existen trabajos previos que muestran que tanto la red de asociaciones (G_{sww}) como la red semántica (G_{w2v}) tienen estructura de *small-world* [5, 6]. Comparar las características de redes, como el camino mínimo, con distintos modelos de redes, en particular con redes

small-world. Para ello generar N redes *small-world* con el mismo número de nodos y aristas, y evaluar si las características medidas en G_{sww} y G_{w2v} podrían pertenecer a dichas distribuciones.

6.1. Opcional 4:

Además de realizar esta tarea con el *small-world*, repetirlo con las *random* o *scale-free*.

7. Formato

Repetir el formato anterior incorporando los comentarios de la devolución del TP1.

8. Nota final

El TP se realizará en grupos de tres o cuatro personas. El TP consiste de una serie de tareas, que pueden consistir en un análisis o contestar una pregunta. Algunas de estas preguntas o tareas están indicadas como optativas. Realizar estas tareas suma puntos pero no son obligatorias. Se puede usar cualquier herramienta de análisis o combinación de herramientas, debiendo indicarlas en el informe. El lenguaje en el que se desarrolle el TP no es excluyente.

La *fecha límite de entrega* es el día Domingo 05/12/2020 a las 23.55hs a través del campus.

Tabla de puntos: Cantidad máxima de puntos que se pueden obtener por ...

- ... la tarea obligatoria 1: 1.5
- ... la tarea obligatoria 2: 1.5
- ... la tarea obligatoria 3: 2.5
- ... la tarea obligatoria 4: 2.5
- ... la tarea opcional 1: 0.5
- ... la tarea opcional 2: 0.5
- ... la tarea opcional 3: 0.5
- ... la tarea opcional 4: 0.5
- ... la tarea opcional 5: 0.5

Puntaje máximo posible: diez

Referencias

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

[2] Elias Costa, M., Bonomo, F., and Sigman, M. Scale-invariant transition probabilities 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. *Proceedings 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.