

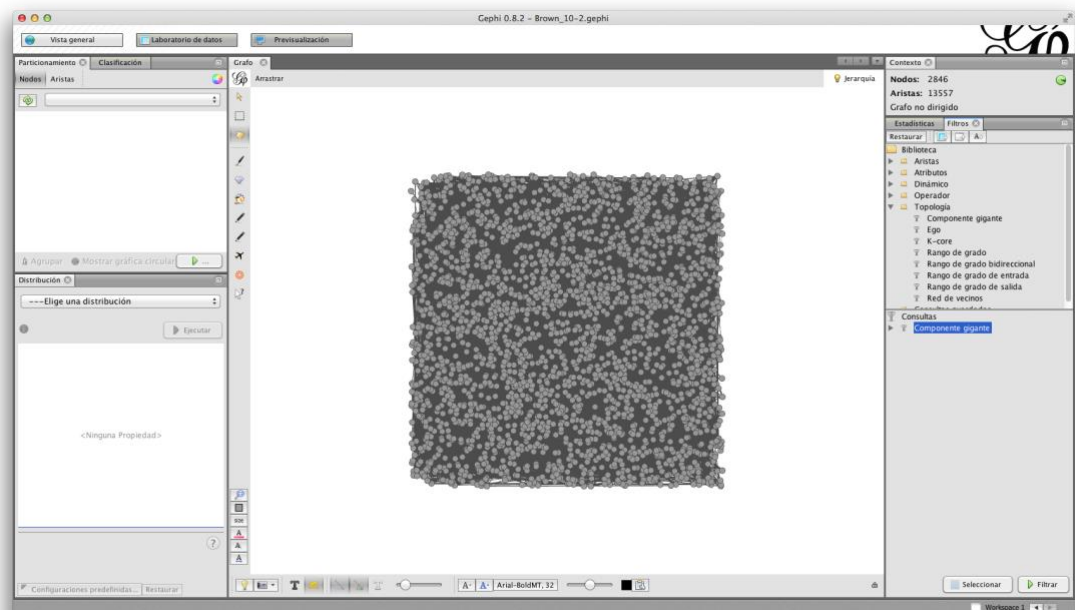
Técnicas Basadas en Grafos Aplicadas al Procesamiento del Lenguaje Natural

Tema 3

En esta tarea vamos a introducirnos en el análisis de redes sociales (Social Network Analysis - SNA) para analizar el comportamiento de las relaciones entre términos dentro de un conjunto de documentos. Para ello utilizaremos un grafo construido a partir de los dos elementos básicos: los nodos serán los términos presentes en un conjunto de documentos y los enlaces se establecerán según la co-aparición de dichos términos en un mismo documento. El objetivo de esta tarea es analizar las relaciones subyacentes que se producen entre estos términos.

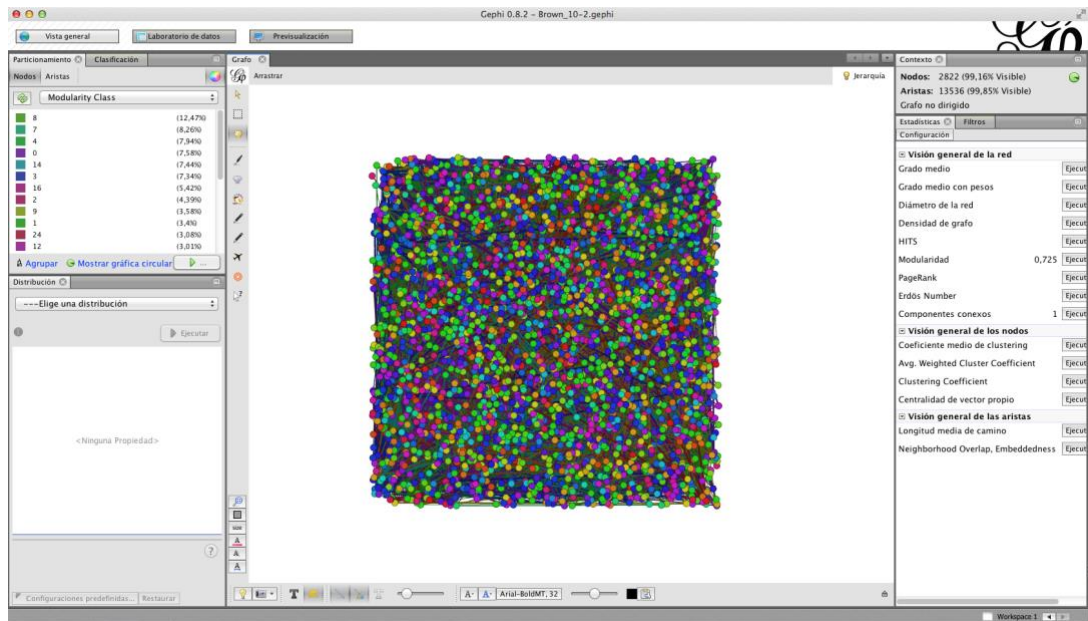
A continuación se indican los pasos de los que se compone la tarea:

1. La tarea comienza con la apertura del fichero que contiene el grafo “Brown_10-2.graphML” (disponible en el curso virtual).
2. Posteriormente nos iremos a la pestaña Filtros, y arrastraremos el filtro “Topología->Componente gigante” a la ventana inferior de “Consultas”. Posteriormente aplicaremos el filtro mediante la opción “Filtrar”.

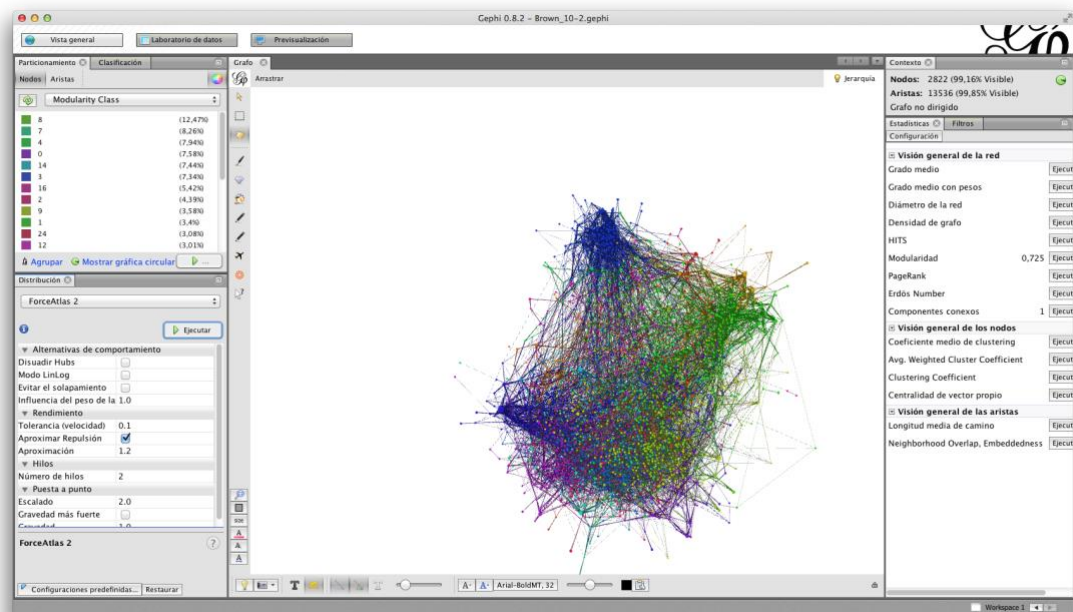


3. En la pestaña “Estadísticas” comprobaremos que el número de “Componentes Conexos” es 1.
4. Posteriormente en “Estadísticas” ejecutaremos la opción “Modularidad” (usando pesos). El informe nos debería devolver un grado de Modularidad alrededor del 0,725 y un número de comunidades alrededor de 25. El algoritmo ejecutado en este punto tiene una cierta aleatoriedad, por este motivo es posible que entre diferentes ejecuciones pueda mostrar cambios.

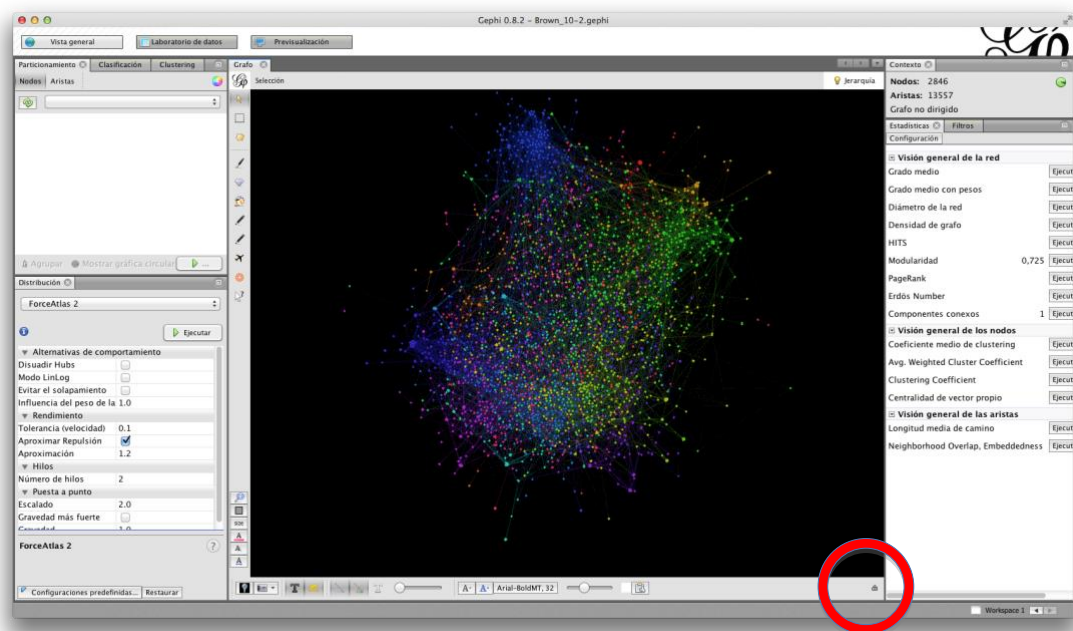
5. Ahora calcula el “Grado Medio” y el “Grado Medio con Pesos” ya que posteriormente los utilizaremos.
6. En la pestaña “Particionamiento” pulsaremos en refrescar para actualizar los parámetros disponibles. En este punto seleccionaremos “Nodos->Modularity Class” y ejecutaremos el coloreado.



7. Una vez realizado el particionamiento y coloreado el grafo, sería interesante poder visualizar el grafo de una manera más apropiada para poder diferenciar las diferentes comunidades. Para ello en la pestaña “Distribución” seleccionaremos la opción “ForceAtlas 2”. Algunos algoritmos de distribución deben pararse por el usuario en un determinado punto. Este algoritmo en concreto podremos pararlo después de 30 segundos aproximadamente para poder tener un resultado optimo.

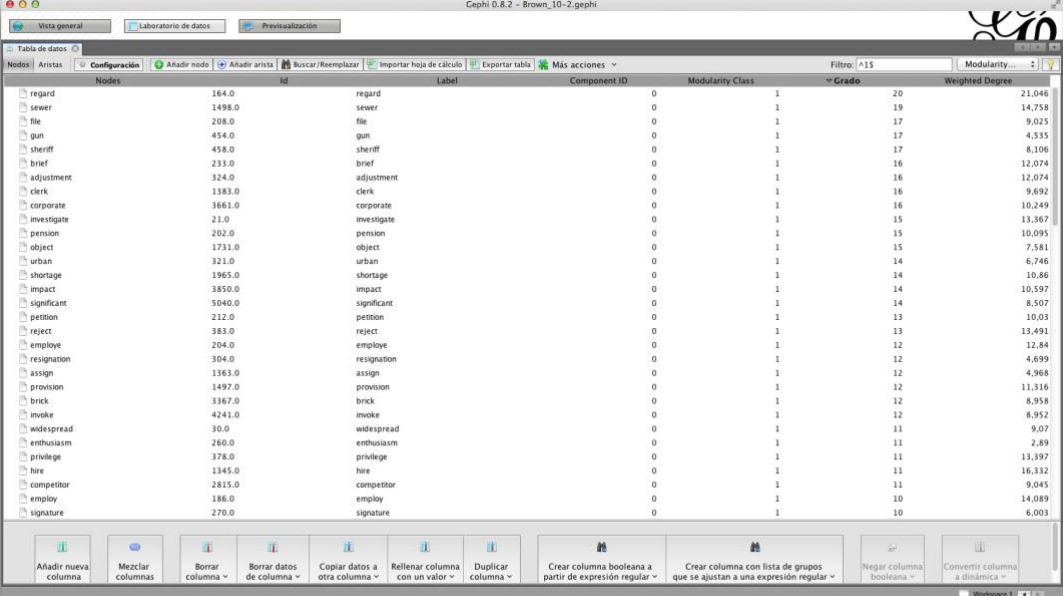


8. Por último en cuanto a la visualización, justo debajo de la imagen del grafo puedes editar algunas opciones.



9. A continuación nos adentraremos en el “Laboratorio de Datos”, donde podremos analizar tanto los nodos como las aristas de una manera más eficiente. Como el objetivo de esta tarea es analizar las comunidades del grafo considerado, podemos seleccionar de manera individual cada comunidad de la siguiente manera. En la parte superior derecha podemos seleccionar el filtro “Modularity” donde podremos introducir el identificador asignado a cada comunidad. El filtro recibe expresiones regulares, por tanto si queremos

analizar la comunidad “1”, podemos introducir una expresión regular del tipo “^1\$” que indique que el identificador debe comenzar y finalizar por 1.



Nodes	Id	Label	Component ID	Modularity Class	Grado	Weighted Degree
regard	164.0	regard	0	1	20	21.046
sewer	1498.0	sewer	0	1	19	14.758
file	208.0	file	0	1	17	9.025
gun	454.0	gun	0	1	17	4.535
sheriff	458.0	sheriff	0	1	17	8.106
brief	233.0	brief	0	1	16	12.074
adjustment	324.0	adjustment	0	1	16	12.074
clerk	1388.0	clerk	0	1	16	9.692
corporate	3661.0	corporate	0	1	16	10.249
investigate	21.0	investigate	0	1	15	13.367
pension	202.0	pension	0	1	15	10.095
object	1731.0	object	0	1	15	7.581
urban	321.0	urban	0	1	14	6.746
shortage	1965.0	shortage	0	1	14	10.86
impact	3850.0	impact	0	1	14	10.597
significant	5040.0	significant	0	1	14	8.507
petition	212.0	petition	0	1	13	10.03
reject	383.0	reject	0	1	13	13.491
employe	204.0	employe	0	1	12	12.84
resignation	304.0	resignation	0	1	12	4.699
assign	1363.0	assign	0	1	12	4.968
provision	1497.0	provision	0	1	12	11.316
brick	3367.0	brick	0	1	12	8.958
invade	4241.0	invade	0	1	12	8.952
widespread	30.0	widespread	0	1	11	9.07
enthusiasm	260.0	enthusiasm	0	1	11	2.89
privilege	378.0	privilege	0	1	11	13.397
hire	1345.0	hire	0	1	11	16.332
competitor	2815.0	competitor	0	1	11	9.045
employ	186.0	employ	0	1	10	14.089
signature	270.0	signature	0	1	10	6.003

Una vez terminado este ejemplo guiado, **se pide un documento PDF en el que se incluya:**

- Realizar un informe en el que por cada comunidad, se ordenen los nodos por “Grado” y “Grado con pesos” y se proporcionen los 10 nodos con una mayor valor correspondiente a cada medida y comunidad para ambos casos.
- Desarrollar una breve reflexión de las diferencias entre los nodos contenidos en cada comunidad en función de la medida de grado utilizada (Grado vs Grado con pesos)
- Por cada comunidad y dados los 10 nodos con un mayor grado sin pesos, proporcionar un término que represente por su significado a esa comunidad. Es decir, en una comunidad donde los 5 primeros nodos fuesen [pelota, arbitro, gol, pie, césped] un ejemplo de termino que podría representarla sería “futbol” (no se trata de un ejemplo real sino ilustrativo). Se valorará positivamente que el término forme parte del grafo.

Tarea Opcional:

Esta tarea opcional consiste en estudiar en profundidad un algoritmo basado en grafos y analizar su repercusión en el Procesamiento del Lenguaje Natural (PLN). El trabajo consiste en seleccionar un algoritmo basado en grafos (Shortest Path, Random Walks, Spreading Activation, Minimum Spanning Trees, etc.) dentro del temario del curso (Capítulo 2 – Graph-Based Algorithms – bibliografía básica). Se deberá hacer una breve memoria en la que se explique su funcionamiento con ilustraciones de ejemplo. Se buscará un artículo científico (por ejemplo en [Google Scholar](#)) que trate de resolver un problema de PLN (Word Sense Disambiguation, Question Answering, Keyword Extraction, etc.) mediante el algoritmo seleccionado y se deberá explicar cómo ha sido aplicado este algoritmo en dicho artículo y cuales han sido las ventajas de utilizar el algoritmo seleccionado frente a otras posibilidades. Aparte de los problemas de PLN mencionados anteriormente, pueden encontrarse otros ejemplos en el capítulo 9 de la bibliografía básica en la que se citan ciertas aplicaciones del uso de grafos en el área del PLN.

Se pide un documento PDF en el que se incluya el informe realizado.