UNIVERSIDAD DE BUENOS AIRES
FACULTAD DE CIENCIAS EXACTAS Y NATURALES
DEPARTAMENTO DE COMPUTACIÓN

Trabajo Práctico 2
Impacto de noticias falsas en las redes sociales

Bases de Datos

29 de septiembre 2017

1. Introducción y objetivos

La introducción de las redes sociales en estos últimos años, ha llevado a que la dinámica social cambie y se adapte a las nuevas tecnologías. Tanto es así, que las nuevas maneras de adquirir información cambiaron. Prender el noticiero o leer el diario son actividades que comenzaron a quedar en desuso. Cada vez es más usual el informarse a través de las redes sociales como *Facebook* o *Twitter*. Según el *Pew Research Center*[1] cerca del 64 % de la población estadounidense utiliza las redes sociales y de ese porcentaje un 50 % lo hace para informarse de las noticias. Este es un fenómeno que viene creciendo según el diario *Telegraph*[2].

La información y las noticias pueden tomar distintos formatos y narrativa, podemos etiquetar la información que consumimos en cierta, engañosa o simplemente falsa. Detrás de estos dos últimos tipos de información se encuentran intereses difíciles de comprender. El *World Economic Forum* ha declarado a las fuentes falsas de información como uno de los mayores riesgos globales actuales, junto a la escasez del agua, el terrorismo, etc.[3]

Los fenómenos de esparcimiento de noticias falsas son complejos, pero han sido observados comportamientos polarizadores en los cuales comunidades enteras se segregan [1] o, en el caso de campañas electorales donde se han detectado abusos[3].

En este trabajo práctico analizaremos el impacto de las noticias falsas en una comunidad. Las bases de datos orientadas a grafos utilizan fuertemente estructura y pueden ser explotas con ese fin (ver 1). Utilizaremos para esto la base de datos orientada a grafos *neo4j*[4] y nos basaremos en los datos generados por el grupo de investigación de ciencias de las redes, *IUNI* [5] en su plataforma *Hoaxy*[6].

2. La base de datos propuesta

Para la elaboración de este trabajo práctico, se utilizará el motor de la base de datos *neo4j*[7] en su versión comunitaria. Se utilizará la consola de *neo4j* que es web e interactiva. En la figura 2 se puede ver como es.

Para comenzar a familiarizarse con el ambiente recomendamos comenzar con el tutorial de películas para entrar en calor con el lenguaje *Cypher*[8] que utilizaremos para este trabajo práctico.

3. Los datos

Los datos utilizados serán los recabados por el sistema *Hoaxy*[9]. Estos vienen en formato JSON por lo cual será importante transformarlos en un formato apto para ser consultados en *neo4j*.

Veamos un ejemplo de los datos:

[1] http://www.pewresearch.org/fact-tank/2014/09/24/how-social-media-is-reshaping-news/
[2] http://www.telegraph.co.uk/technology/0/fake-news-origins-grew-2016/
[3] http://reports.weforum.org/global-risks-2013/risk-case-1/digital-wildfires-in-a-hyperconnected-world/
[4] https://neo4j.com/
[5] http://iuni.iu.edu/
[6] http://hoaxy.iuni.iu.edu/
[7] https://neo4j.com/
[8] https://neo4j.com/docs/cypher-refcard/current/
[9] http://hoaxy.iuni.iu.edu/

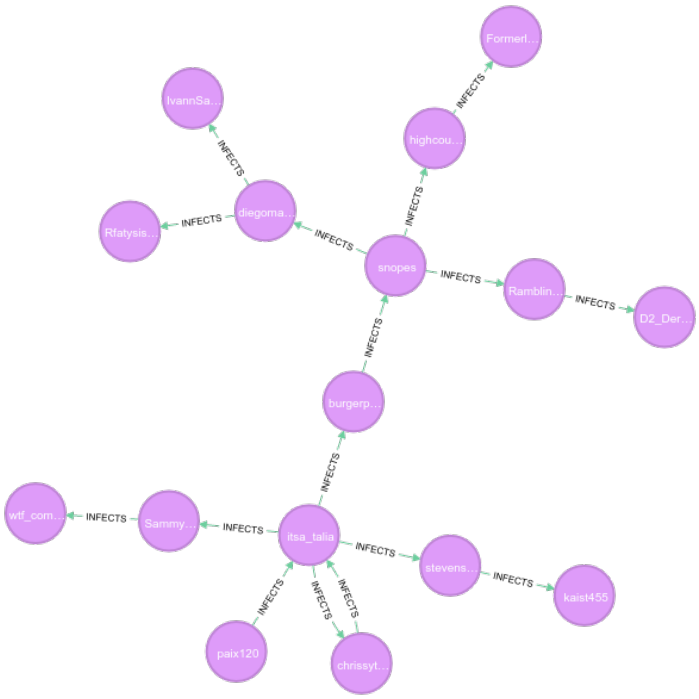

Figura 1: Ejemplo de nodos y relaciones en *neo4j*

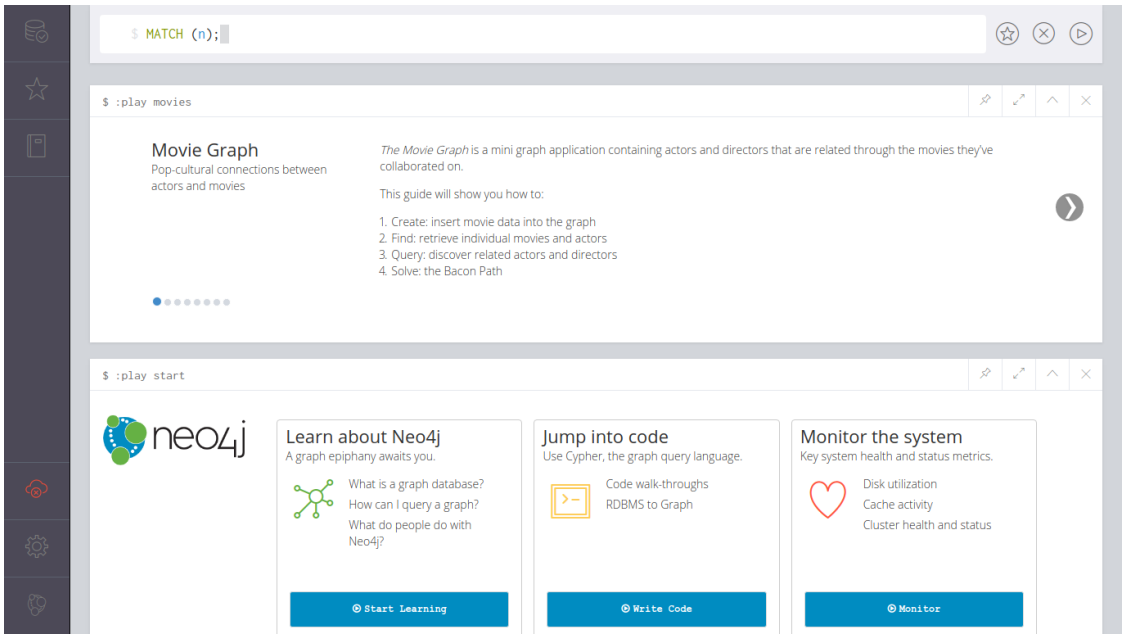

Figura 2: Consola de trabajo de *neo4j*. Instanciada con el comando bin/neo4j console

```
1   {
2     # La URL de la noticia
3     "canonical_url": "http://www.dcclothesline.com/2016/09/09/
4                       hillary-clinton-wore-an-earpiece-at-commander-in-chief-
                            forum/",
5     # Fecha de la noticia
6     "date_published": "2016-09-09T09:00:40.000Z",
7     # El dominio del sitio de noticias
8     "domain": "dcclothesline.com",
9     # El nombre del usuario que compartio la noticia
10    "from_user_screen_name": "BigSkyGuy57",
11    # Si hizo un 'mention' al usuario destino
12    "is_mention": true,
13    # Si el sitio es de tipo 'clain' o 'fact_checking'
14    "site_type": "claim",
```

```
15        # El titulo de la noticia
16        "title": "Hillary Clinton wore an earpiece at Commander-in-Chief Forum",
17        # El nombre de usuario al que fue dirigida la noticia
18         "to_user_screen_name": "DCClothesline",
19     }
```

4. Modelado

Se desea modelar los datos presentados en la sección anterior para representar dos grafos con distintas características:

- *El grafo bipartito* que tiene los nodos de noticias relacionados con los usuarios que se relacionaron con esta pieza de información.
- *El grafo de infección* que relaciona los usuarios que compartieron información falsa con los usuarios que la recibieron.

Para hacer esto, se deberá transformar la información a un formato apropiado para que *neo4j* pueda reconocerlo. Veamos detenidamente este documento de tipo JSON. Contiene:

- La url, el título, el dominio y la fecha de publicación de la **noticia**.
- Para cada usuario que participó en la transmisión de este artículo su id de *Twitter*, su nombre de usuario (*screen name*).
- Para el tweet en el cual se compartió, tenemos qué tipo de interacción fue (retweet o mention por nombrar algunas), el id de el tweet, la fecha de creación (que no es la misma que la de la noticia).

4.1. Ejercicios de modelado

1. Siguiendo la guía de importación de datos de *neo4j*[10] prepare un archivo por cada modelo. Los lotes deberán contener la información transformada del archivo *noticias.json* en CSV para ambos modelos con la mínima cantidad de campos necesaria. Analizar herramientas para la traducción, por ejemplo: *jq*.
2. Escribir en el lenguaje *Cypher* los scripts de importación para ambos modelos. Describir los nodos y relaciones con la información dada en el archivo. ¿Qué atributos agregó a los nodos?¿Y a las relaciones?¿Por qué?

Importante: Para ambos modelos garantice unicidad en los nodos de usuarios y noticias. Es decir que si dos usuarios comparten su nombre o id, deben tener solo una instancia. Análogamente para las noticias.

4.2. Preguntas a responder

Escriba en el lenguaje *Cypher* las consultas que respondan estas preguntas. Utilice herramientas de graficación como histogramas para presentar la información de manera adecuada. Para las preguntas que devuelvan un sub-grafo, utilice las herramientas que provee *neo4j* para exportar la representación gráfica de dichos grafos.

1. Enumere las noticias que han impactado en más del 25 % de la comunidad.
2. Genere el sub-grafo de usuarios que consumen las mismas noticias.
3. ¿Existen usuarios de *Twitter* que han estado en contacto con más del 20 % del lote de noticias?
4. ¿Cómo es la distribución de los grados de entrada y salida de los nodos? Presente la información en un histograma.
5. Llamaremos root-influencers a los nodos raíces del grafo de infección. Escriba una consulta que dado un nodo de usuario en el grafo de infección diga si es root-influencer o no. ¿Qué proporción hay de root-influencers? Muestre la información apropiadamente.
6. Calcule el grado de la infección para un root-influencer dado. El grado de infección está dado por el camino más largo que se puede alcanzar desde un root-influencer.
7. Pode el grafo quitando todos los root-influencers y muestre gráficamente como queda el grafo resultante. Si la información es muy grande, recorte apropiadamente.
8. Considere la introducción de índices a los modelos. Evalúe la *performance* de las consultas implementadas con y sin utilización de índices.

[10] `https://neo4j.com/developer/guide-import-csv/`

5. Condiciones de entrega

La entrega deberá constar, como mínimo, de la siguiente documentación:

- Introducción y explicación del problema a resolver.
- Detalle del diseño de los modelos de grafo con las imágenes de *neo4j*.
- Análisis de las consultas implementadas en *Cypher*.
- Código y explicación de las consultas así como también de cada decisión tomada.
- Conclusiones.

Fecha de entrega: 10 de Noviembre de 2017
Las entregas y consultas deben ser enviadas al tutor y **no** a bddoc.

Referencias

[1] Michela Del Vicario, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H Eugene Stanley, and Walter Quattrociocchi. The spreading of misinformation online. *Proceedings of the National Academy of Sciences*, 113(3):554–559, 2016.

[2] Flaviano Morone, Byungjoon Min, Lin Bo, Romain Mari, and Hernán A Makse. Collective influence algorithm to find influencers via optimal percolation in massively large social media. *Scientific reports*, 6, 2016.

[3] Jacob Ratkiewicz, Michael Conover, Mark R Meiss, Bruno Gonçalves, Alessandro Flammini, and Filippo Menczer. Detecting and tracking political abuse in social media. *ICWSM*, 11:297–304, 2011.