Universidad de Los Andes
Mérida - Venezuela

seminario de tesis

Acercamiento al lenguaje usado en Twitter mediante Redes Complejas

Br. Jeinfferson Bernal G.

Tutor: prof. Kay Tucci

Centro de Física Fundamental
Grupo de Caos y Sistemas Complejos
http://www.ciens.ula.ve/cff/caoticos

Sobre la obtención de datos de Twitter…

Twitter: 300 millones de usuarios

Fig 1. Logo de Twitter [1]

- Plataforma social creada por Jack Dorsey (2006).
- Mensajes cortos de 280 caracteres (tweet).
- Compartir experiencias en tiempo real.

Características

- Sencillo (utilización).
- Breve (cortos).
- Universal (25 idiomas).
- Gratuito.
- Asimétrico (información).
- Accesible (cuenta).
- Multiformato (inclusión).

Una de las redes más populares en la actualidad!

[1] Wosu public media. Juguetes parlantes, actualizacion de twitter y anuncios de facebook. Publicado 3 de octubre de 2017[En linea]. Consultado(09 de febreo de 2019). Disponible: http://radio.wosu.org/post/tech-tuesday-talking-toys-twitter-update-and-facebook-ads

Obtención de Tweets

API de Twitter → **Interface de Programación de Aplicaciones**

- API de Anuncios.
- Filltrado de tweets en tiempo real.
- Búsqueda de Tweets **(Search tweets)**.
- API de Mensajes directos.
- API de Índice de Referencia.

Ofrece 3 niveles de APIs de búsqueda

Estándar | **Premium** | **Empresarial**

Pasos para comenzar

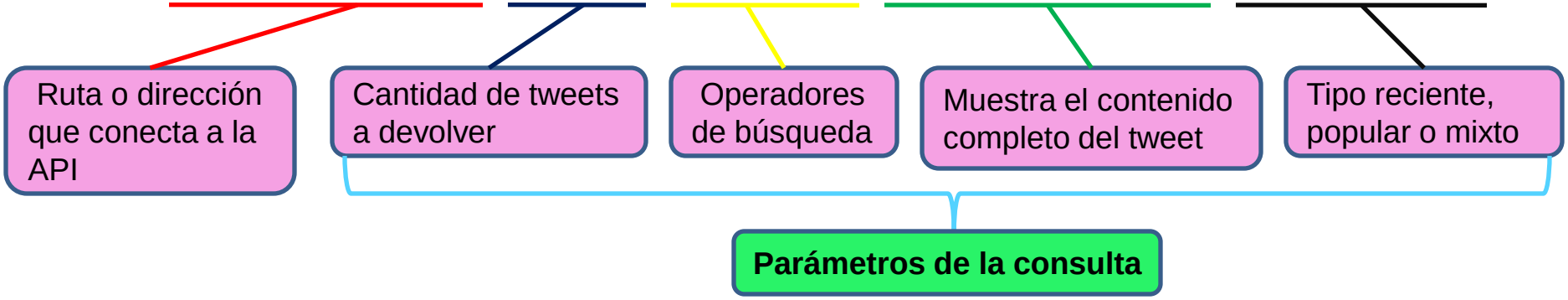
- Tener una cuenta de Twitter
- Registrarse en el portal de desarrolladores https://developer.twitter.com
- Crear una aplicación.
- Utilizar Twurl https://www.youtube.com/watch?v=SF22GlA_nZU

- Búsqueda de Tweets recientes (últimos 7 días).
- Gratuita.
- 180 peticiones / 15 minutos.
- Máximo de tweets por petición: 100.
- Formato de respuesta**: json**.
- Consulta se realiza mediante **parámetros** y **operadores**

Peticiones a la Search Tweets mediante Twurl

Desde la terminal

$ twurl “/1.1/search/tweets.json?count=100&q=**operadores**&tweets_mode=extended&result_type=recent”



Operadores de búsqueda

Operador	Descripción
#Noticias	Contiene el hashtag Noticias
from:Twitter	Enviado desde la cuenta Twitter
to:NASA	Tweets en respuesta a la cuenta NASA
Bota since:2018-12-24	Contiene ‘Bota’ y enviado desde la fecha dada
Bota until:2018-12-24	Contiene ‘Bota’ y enviado antes de la fecha dada
‘hora feliz’	Tweets con la frase exacta ‘hora feliz’

Formato de los Tweets

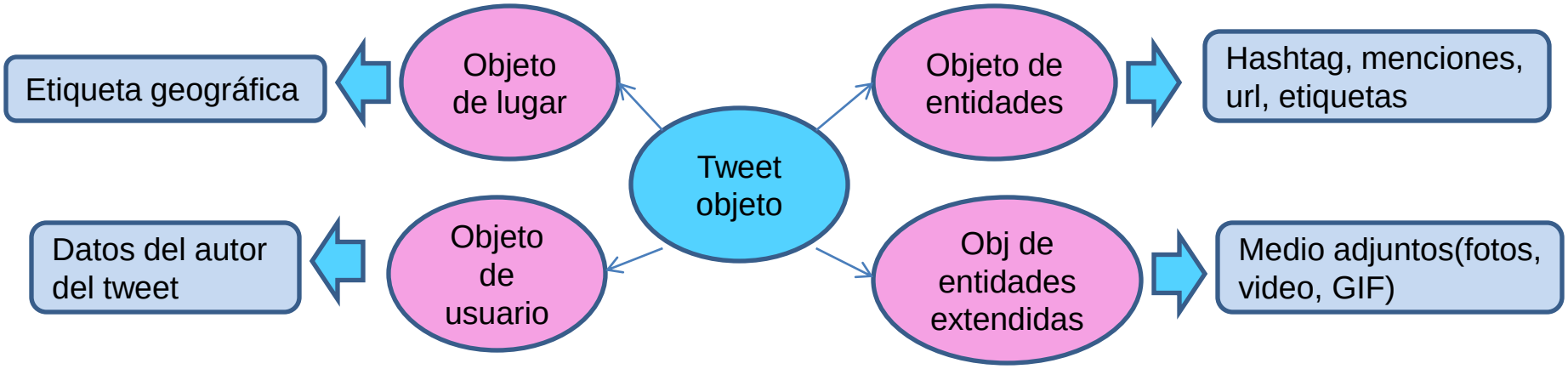
Formato JSON

Java**S**cript **O**bject **N**otation
(Notación de objetos de JavaScript)
- Formato ligero de intercambio de datos.
- Es independiente del lenguaje utilizado.
- Constituido por dos estructuras:

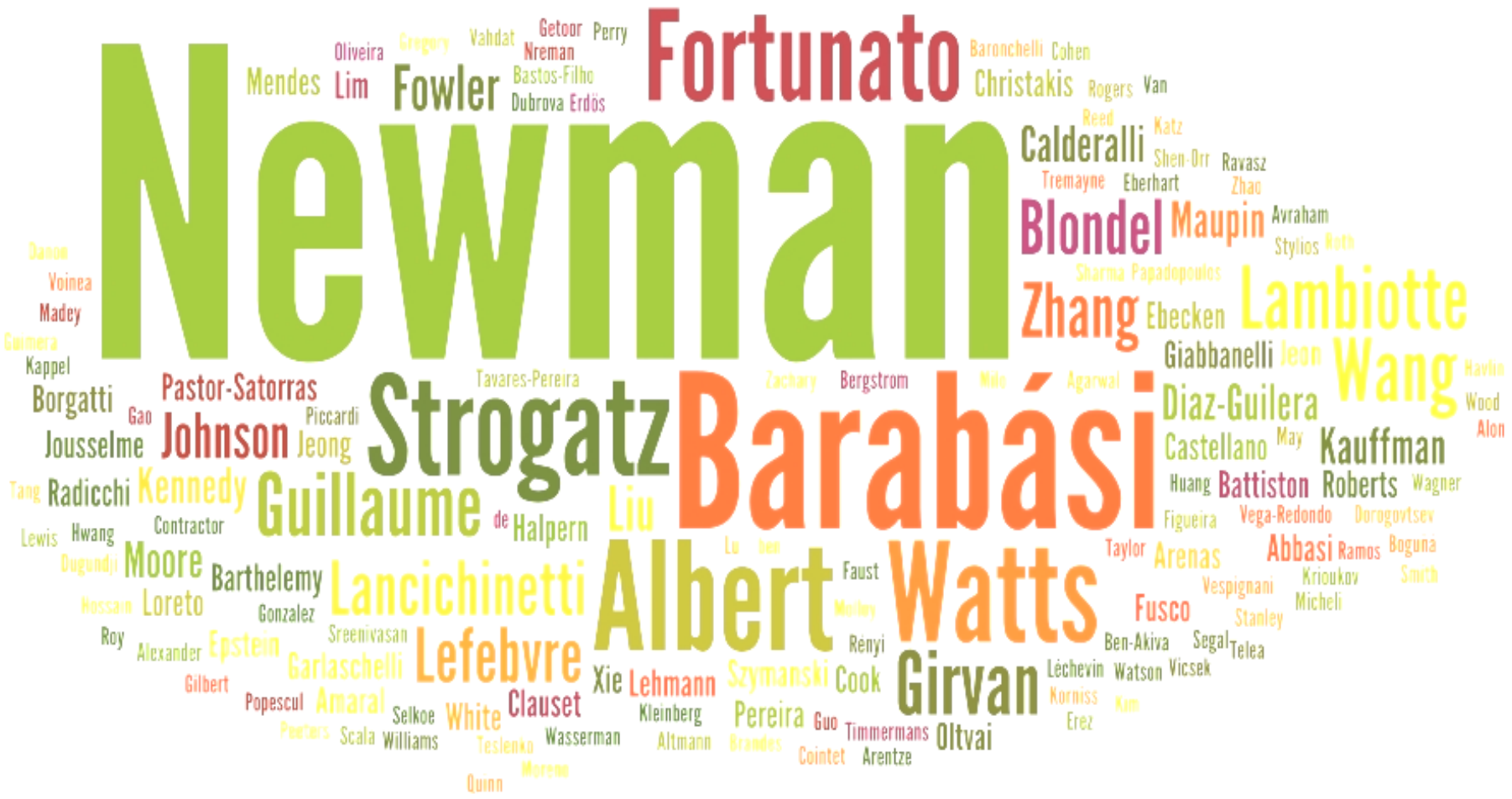
1.-Colección de pares nombre/valor. (objetos)
2.-Lista ordenada de valores.(arreglos)

```json
{
    "Nombre":"Pepito Pérez",
    "DNI": "5167778E",
    "Edad":22,
    "Asignaturas": {
        "Obligatorias":{
            "Sistemas Operativos",
            "Compiladores",
            "Bases de Datos"
        },
        "Optativas":{
            "Bases de Datos NoSQL",
            "Minería de Datos",
            "Programación Lógica"
        },
        "LibreElección":{
            "Ajedrez",
            "Música Clásica"
        }
    }
}
```

Un tweet objeto puede contener alrededor de 150 atributos



Sobre Redes complejas…



Teoría de Grafos

Comienzos…

El problema de los 7 puentes de Königsberg (1736)

Dado el mapa de Königsberg, con el río Pregel dividiendo el plano en cuatro regiones distintas, que están unidas a través de los siete puentes, ¿es posible dar un paseo comenzando desde cualquiera de estas regiones, pasando por todos los puentes, recorriendo sólo una vez cada uno, y regresando al mismo punto de partida?

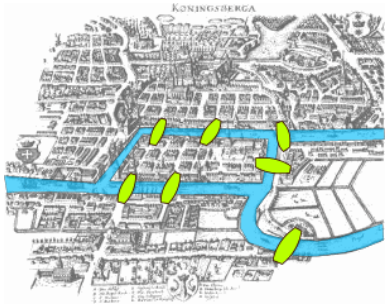


Fig 2. Bosquejo de Königsberg [2]

Solución propuesta por Leonard Euler →

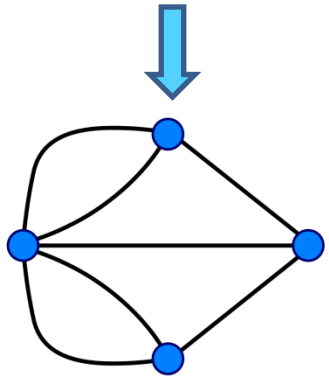


Fig 3. Grafo de los puentes [3]

"Como los 4 puntos en el diagrama poseen un número impar de líneas incidentes, se concluye que es imposible definir un camino con las características buscadas."

"Estar en una red" → **Característica** → Sistemas Complejos

[2] [3] Wikipedia. Problema de los puentes de Königsberg .Actualizado 6 de Octubre 2018.[En linea]. Consultado(30 de Mayo de 2019). Disponible: https://es.wikipedia.org/wiki/Problema_de_los_puentes_de_K%C3%B6nigsberg

¿Qué es una red?

"Es una estructura constituida por entidades que son representadas por nodos y sus relaciones o interacciones por enlaces."

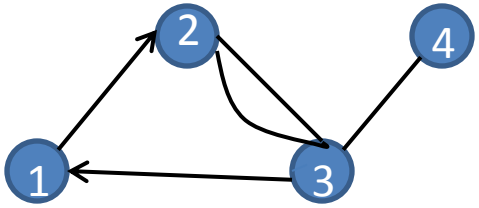


Fig 4. Red Mixta n = 4

Una red puede ser expresada en forma matemática mediante matrices.

Matriz de Adyacencia de la Red

$$A = \begin{array}{c} \begin{matrix} 1 & 2 & 3 & 4 \end{matrix} \\ \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 1 & 2 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix} \end{array} \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \end{matrix}$$

Contiene toda la información característica de la red**.**

Red Dirigida: Matriz de adyacencia es asimétrica.
Red no dirigida: Matriz de adyacencia simétrica.

Algunas medidas que caracterizan una Red

- **Coeficiente de Clustering global**
- **Coeficiente de Clustering local**.
- **Centralidad**.
- **Modularidad**.
- **Distribución de grados**.
- Distancia media.
- Diametro de la red.

Medidas de la Red

Coeficiente de Clustering Local

Propuesto por Watts y Strogatz (1998) .
"Proporción de triángulos asociados a un nodo con el número de triángulos que soporta según su grado".

$$C_i = \frac{2t_i}{K_i(K_i - 1)}$$

t_i: número de triángulos adjuntos al nodo i.

K_i: grado del nodo i.

Definiciones.

Grado: número de enlaces que posee un nodo.
Camino: ruta en la que todos los nodos son distintos.
Ciclo: camino cerrado en la red.
Triángulo: ciclo de longitud 3.

Coeficiente de Clustering Global

Propuesto por J. Newman (2002). "Mide la agrupación o transitividad de la red ".

$$C = \frac{3T}{|P_2|}$$

$$|P_2| = \sum_{i=1}^{n} \frac{K_i(K_i - 1)}{2}$$

T: número total de triángulos de la red.

P_2: número de caminos de longitud 2 en la red.

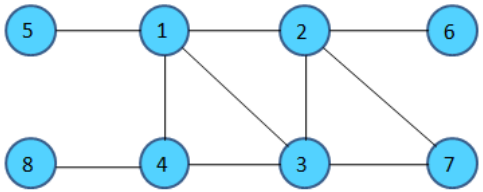


Fig 5. Red no dirigida n = 8

Coeficiente de Clustering local

$$C_1 = \frac{2(2)}{4(3)} = \frac{1}{3} = 0.3 = C_2 = C_4 \qquad C_7 = \frac{2(1)}{2(1)} = 1 \qquad C_3 = \frac{2(3)}{4(3)} = 0.5 \qquad C_5 = C_6 = C_8 = 0$$

Coeficiente de Clustering global

$$|P_2| = 22 \qquad C = \frac{3(3)}{22} = 0.41$$

Medidas de la Red

Centralidad.

Propuestas por L.Freeman (1978).
- Grado.
- Cercanía.
- Intermediación.

Propuesta por J. Newman (2003).
- Cercanía Armónica.

Propuesta por Bonacich (1987).
- Centralidad de Autovector.

Centralidad de Grado

"Mide la influencia o prestigio de un nodo según su grado (in-degree , out-degree) ". Medida hecha en función de los vecinos más cercanos.

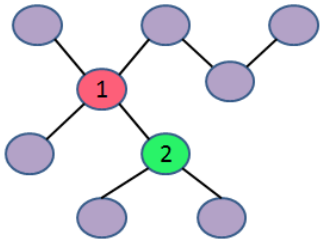


Fig 6. Red no dirigida n = 9

Grado nodo i $K_i = \sum_{j=1}^{n} a_{ij} = (A\boldsymbol{e})_i$

Grado entrada i $K_i^{in} = \sum_{i=1}^{n} a_{ji} = (\boldsymbol{e}^T A)_i$

Grado salida i $K_i^{out} = \sum_{j=1}^{n} a_{ij} = (A\boldsymbol{e})_i$

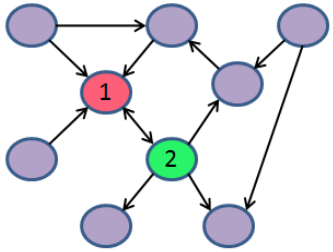


Fig 7. Red dirigida n = 9

nodo mayor centralidad: $C_D(1) = K_1 = 4$

Segundo nodo de mayor centralidad: $C_D(2) = K_2 = 3$

nodo más influyente: $C_D^{out}(2) = K_2^{out} = 4$ $C_D^{in}(2) = K_2^{in} = 1$

nodo con más prestigio: $C_D^{out}(1) = K_1^{out} = 1$ $C_D^{in}(1) = K_1^{in} = 4$

Medidas de la Red

Centralidad de Cercanía

"Mide qué tan cercano esta un nodo al resto de los nodos de la red". Medida que indica la cercanía del nodo al centro de la red.

$$C_c(i) = \frac{n-1}{s(i)} \qquad s(i) = \sum_{j \epsilon V(G)} d(i,j) = (De)_i$$

$s(i)$: Suma de la distancias del nodo i.

$d(i,j)$: Distancia del camino más corto entre j e i.

D: Matriz de distancia de la red

Camino más corto: nº mínimo de enlaces para ir del nodo A al nodo B

Matriz de distancia de la red

$$D = \begin{bmatrix} 0 & 1 & 1 & 1 & 2 & 1 & 1 & 2 & 3 & 3 \\ 1 & 0 & 1 & 2 & 3 & 2 & 2 & 3 & 4 & 4 \\ 1 & 1 & 0 & 1 & 2 & 2 & 2 & 2 & 3 & 3 \\ 1 & 2 & 1 & 0 & 1 & 2 & 2 & 1 & 2 & 2 \\ 2 & 3 & 2 & 1 & 0 & 1 & 2 & 2 & 3 & 3 \\ 1 & 2 & 2 & 2 & 1 & 0 & 1 & 3 & 4 & 4 \\ 1 & 2 & 2 & 2 & 2 & 1 & 0 & 3 & 4 & 4 \\ 2 & 3 & 2 & 1 & 2 & 3 & 3 & 0 & 1 & 1 \\ 3 & 4 & 3 & 2 & 3 & 4 & 4 & 1 & 0 & 2 \\ 3 & 4 & 3 & 2 & 3 & 4 & 4 & 1 & 2 & 0 \end{bmatrix}$$

Vector Suma de distancia de cada nodo

$$\mathbf{S} = \begin{bmatrix} 15 & 22 & 17 & 14 & 19 & 20 & 21 & 18 & 26 & 26 \end{bmatrix}^T$$

Vector de todas las centralidades de cercanía

$$C_C = \begin{bmatrix} 0.600 & 0.409 & 0.529 & 0.643 & 0.474 & 0.450 & 0.428 & 500 & 0.346 & 0.346 \end{bmatrix}^T$$

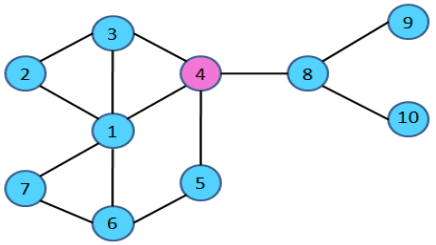


Fig 8. Red no dirigida n = 10

Nodo más central según Cercanía : 4

Nodo más central según Grado : 1

Para redes dirigidas

- **Cercanía de entrada**: cómo se acerca un nodo al centro de la red tras recibir información del resto de los nodos.
- **Cercanía de salida**: cómo se acerca un nodo al centro de la red tras enviar información al resto de los nodos.

Medidas de la Red

Centralidad de Intermediación

"Mide la relevancia que tiene un nodo para la comunicación entre otro par de nodos".
La información viaja desde un nodo a otro a través del camino más corto que los conecta.

$$C_B(i) = \sum_i \sum_k \frac{\rho_{jk}(i)}{\rho_{jk}}, i \neq j \neq k$$

Redes no dirigidas

- ρ_{jk}: nº de caminos más cortos que conectan a los nodos j y k.
- $\rho_{jk}(i)$: nº de caminos más cortos que pasan por i y conectan a los nodos j y k.

Redes dirigidas

- ρ_{jk}: nº total de caminos dirigidos que conectan a los nodos j y k.
- $\rho_{jk}(i)$: nº de caminos dirigidos que pasan por i y conectan a los nodos j y k.

C_B sin normalizar

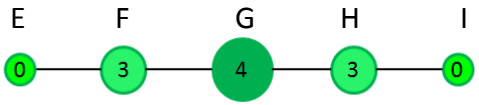


Fig 9. Red no dirigida n = 5

Mayor Intermediación : G

C_B normalizado

$$C_B(i) = \frac{nº\ caminos\ mas\ cortos\ que\ pasan\ por\ i}{nº\ caminos\ mas\ cortos\ de\ la\ red}$$

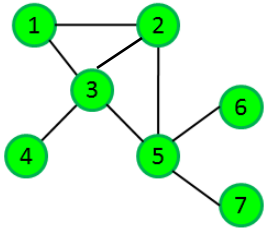


Fig 10. Red no dirigida n = 7

Mayor Intermediación : 5

$$C_B(1) = C_B(4) = C_B(7) = C_B(8) = 0$$

$$C_B(2) = \frac{1}{6} = 0.17$$

$$C_B(3) = \frac{8}{11} = 0.72$$

$$C_B(5) = 1$$

Medidas de la Red

Centralidad de Cercanía Armónica

"Mide qué tan conectado esta un nodo a la red".
Permite detectar diferentes componentes dentro de la red.

$$C_H(i) = \frac{1}{n-1}\sum_{j \neq i}\frac{1}{d(i,j)}$$

$d(i,j)$: Distancia de camino más corto entre los nodos j e i.

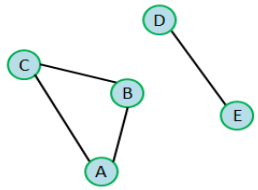


Fig 11. Red no dirigida n = 5

Matriz de distancia de la red

$$D = \begin{array}{c} \\ A \\ B \\ C \\ D \\ E \end{array} \begin{array}{c} \begin{array}{ccccc} A & B & C & D & E \end{array} \\ \begin{bmatrix} 0 & 1 & 1 & \infty & \infty \\ 1 & 0 & 1 & \infty & \infty \\ 1 & 1 & 0 & \infty & \infty \\ \infty & \infty & \infty & 0 & 1 \\ \infty & \infty & \infty & 1 & 0 \end{bmatrix} \end{array}$$

$$C_H(A) = \frac{1}{4}\left(\frac{1}{1}+\frac{1}{1}+\frac{1}{\infty}+\frac{1}{\infty}\right) = \frac{1}{2}$$

$$C_H(D) = \frac{1}{4}\left(\frac{1}{\infty}+\frac{1}{\infty}+\frac{1}{\infty}+\frac{1}{1}\right) = \frac{1}{4}$$

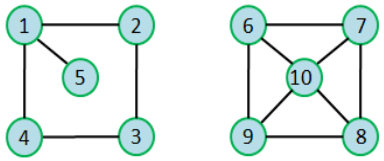


Fig 12. Red no dirigida n = 10

$C_H(1) = 0.27$ $C_H(6) = 0.38$

$C_H(5) = 0.26$ $C_H(10) = 0.44$

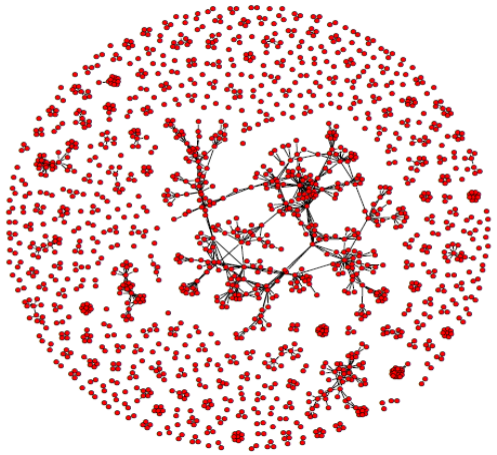


Fig 13. Red de científicos n = 1589 m = 2742 [4]

Nodos bien conectados C. Armónica mayor.
Nodos de una misma componente , valor de C. Armónica similar.

[4] Yannick Rochat. Closeness Centrality Extended To Unconnected Graphs :The Harmonic Centrality Index. Publicado por ASNA 2019 Zürich. Consultado(30 de Mayo de 2019).

Medidas de la Red

Centralidad de Autovector

"Mide la relevancia de un nodo según la relevancia que tiene sus vecinos más cercanos". Basada en el **Teorema de Perron-Frobenius**.

Definiciones…

Matriz irreducible

•Para una red dirigida, la matriz es irreducible si sus nodos son conexo por camino.

• Para una red no dirigida, dicha red esta completamente conectada.

"El teorema de Perron-Frobenius nos asegura de que existe un autovalor mayor, en modulo, al resto de los autovalores de A y que a tal autovalor le corresponde un único autovector."

$$C_A = \left(\lambda_1 \sum_{i=1}^{n} q_1(i)\right) \boldsymbol{q_1}$$

λ_1: Autovalor más grande de la Matriz de adyacencia A.

$\boldsymbol{q_1}$: Autovector asociado al autovalor más grande λ_1 .

$q_1(i)$: Componente i de $\boldsymbol{q_1}$.

Ecuación característica de A.

$$A\boldsymbol{q} = \lambda \boldsymbol{q}$$

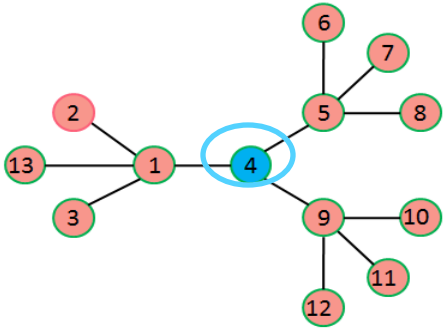


Fig 14. Red no dirigida n =13

$$\boldsymbol{q_1} = \begin{matrix}1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 13\end{matrix}$$
$$= [0.408 \quad 0.167 \quad 0.167 \quad 0.500 \quad 0.408 \quad 0.167 \quad 0.167 \quad 0.167 \quad 0.408 \quad 0.167 \quad 0.167 \quad 0.167 \quad 0.167]^T$$

Nodo mayor centralidad según autovector : 4
Nodo mayor centralidad según grado : 1

Un pequeño grupo de nodos relevantes conectados a un nodo hace que el nodo sea más central que si se conecta un grupo grande de nodos con poca relevancia.

Medidas de la Red

Centralidad de Autovector

Red dirigida

Se utiliza los autovectores principales derecho e izquierdo de la matriz de adyacencia como autovectores de centralidad.

$$AX = \lambda_1 X$$

X: Autovector principal derecho de la matriz de adyacencia A.

Relevancia que un nodo adquiere al dirigir sus enlaces a otros nodos.

$$A^T Y = \lambda_1 Y$$

Y: Autovector principal izquierdo de la matriz de adyacencia A.

Relevancia que un nodo adquiere cuando otros nodos dirigen sus enlaces hacia éste.

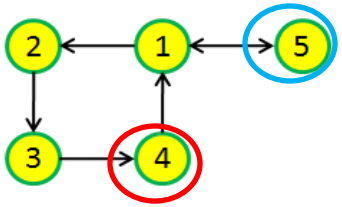


Fig 15 Red dirigida n =5

$$X = \begin{matrix} 1 & 2 & 3 & 4 & 5 \\ [0.592 & 0.298 & 0.366 & 0.465 & 0.465 \end{matrix}]^T$$

→ Privilegio

$$Y = [0.592 \quad 0.465 \quad 0.366 \quad 0.298 \quad 0.465]^T$$

→ Influencia

Medidas de la Red

Modularidad

Algunas definiciones…

Clúster: conjunto de nodos conectados internamente para los cuales la densidad interna es mayor que la externa .
Partición: división de la red que da paso a la formación de comunidades.

$$\delta(G) = \frac{2m}{n(n-1)}$$

$\delta(G)$: Densidad de la red.

$$\delta_{int}(C) = \frac{2m_c}{n_c(n_c-1)} = \frac{\sum_{i \epsilon C} k_i^{int}}{n_c(n_c-1)}$$

$\delta_{int}(C)$: Densidad interna al clúster.

m_c: Enlaces del clúster.

n_c: Nodos del clúster.

$$\delta_{ext}(C) = \frac{\sum_{i \epsilon C} k_i^{ext}}{n_c(n-n_c)}$$

$\delta_{ext}(C)$: Densidad externa al clúster.

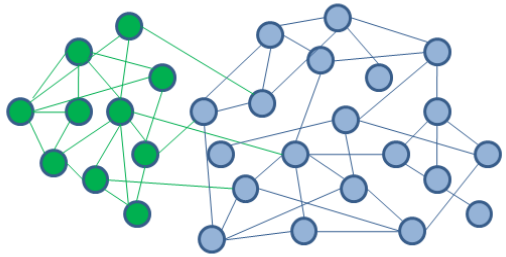


Fig 16. Red no dirigida n = 30, m =48

$$\delta_{int}(G) = \frac{96}{30(30-1)} = 0.11$$

$$\delta_{ext}(C_1) = \frac{4}{10(30-10)} = 0.02$$

$$\delta_{int}(C_1) = \frac{32}{10(10-1)} = 0.35$$

$$\delta_{ext}(C_2) = \frac{4}{20(30-20)} = 0.02$$

$$\delta_{int}(C_2) = \frac{60}{20(20-1)} = 0.16$$

Como $\delta_{int} > \delta_{ext}$ se forma una comunidad

Medidas de la Red

Modularidad

Propuesto por Girvan y Newman.
"Mide la calidad de los clusters en la red".

"Sumas de todas las particiones de la diferencia entre la fracción de enlaces dentro cada partición y la fracción esperada, al considerar una red aleatoria con el mismo grado para cada nodo."

$$Q = \sum_{k=1}^{n_c} \left[\frac{|E_k|}{m} - \frac{1}{4m} \left(\sum_{j \in V_k} k_j \right)^2 \right]$$

n_c: Número de cluster en que se divide la red.

$|E_k|$: Número de enlaces entre nodos de la k-esima partición de la red.

m: Número de enlaces en la red.

k_j: Grado del nodo j.

- Si el número de enlaces internos al clúster no es mas grande que el valor esperado , Q = 0
- Máximo valor de modularidad Q = 1.
- A mayor valor de modularidad, más evidentes son las estructuras de comunidades.

Problema de resolución
Pequeñas comunidades bien definidas son adyacentes a comunidades grandes.

Solución
Uso de otra medidas de calidad para determinar particiones optimas.

$$|E_{c_1}| = 7 \qquad \sum_{j \in V_1} K_j = 16$$

$$|E_{c_2}| = 9 \qquad \sum_{j \in V_2} K_j = 20$$

$$Q = \frac{7}{18} - \left(\frac{16}{36}\right)^2 + \frac{9}{18} - \left(\frac{20}{36}\right)^2 = 0.383$$

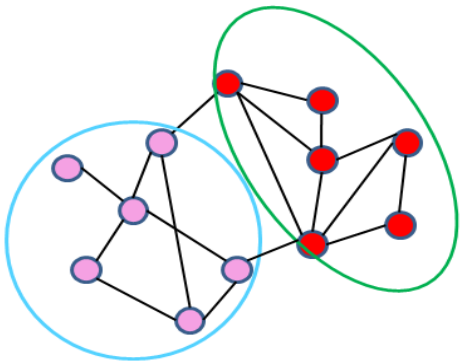


Fig 17. Red no dirigida n = 12, m =17

Medidas de la Red

Distribución de Grados

"Muestra cómo están distribuidos los nodos en la red."

Función de distribución: probabilidad de que un nodo seleccionado al azar tenga k enlaces.

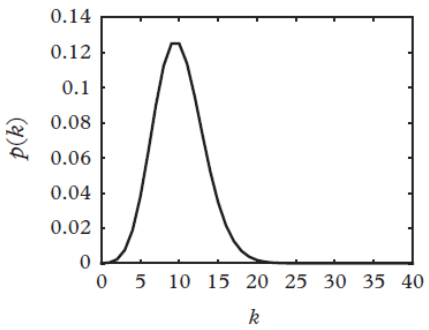
$$p(k) = \frac{n(k)}{n}$$

$p(k)$: Probabilidad de que un nodo seleccionado aleatoriamente sea de grado k.

$n(k)$: numero de nodos de grado k.

Grafica p(k) vs k

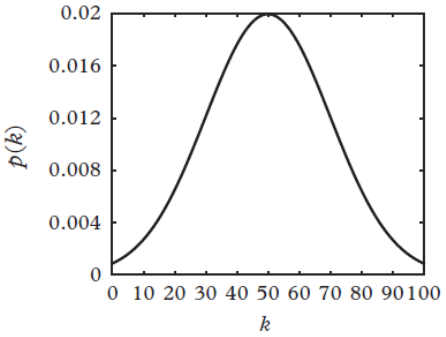
Red Aleatoria



(a) $p(k) = \frac{e^{-k}\overline{k}^{k}}{k!}$

D. Poisson

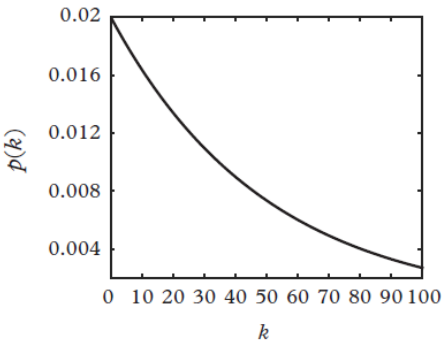
Red Aleatoria con gran cantidad de nodos



(b) $p(k) = \frac{1}{\sqrt{2\pi\sigma_k}}\, e^{-(k-\overline{k})^2/(2\sigma_k^2)}$

D. Normal

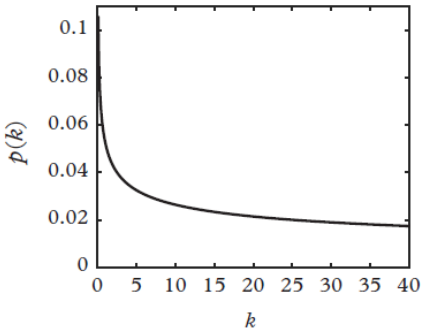
Pequeño Mundo



(c) $p(k) = Ae^{-k/\overline{k}}$

D. Exponencial

Libre de escala



(d) $p(k) = Bk^{-\gamma}$

D. Ley de potencia

D. Ley de potencia ⟶ Redes libre de escala

"Fenómenos bajo estudio se reproducen a si mismo en diferentes escalas de tiempo y espacio" (Autosimilaridad)

$$p(k, C) = B(CK)^{-\gamma} = C^{-\gamma}p(k)$$

Sobre algunas leyes lingüísticas…

Ley de Zipf

Propuesta por George Zipf (1940).
"Describe la distribución de las palabras en el texto mediante su frecuencia de aparición".

Principio de mínima acción

Si las palabras se ordenan según su frecuencia de aparición (de la más frecuentes a la menos) r = 1,2,3….V, la frecuencia viene dada por:

$$f(r) = \frac{k}{r^{\gamma}}$$

$f(r)$: Frecuencia en función de la posición de la palabra r.

k: Constante que depende del tamaño del texto.

γ: Parámetro de distribución.

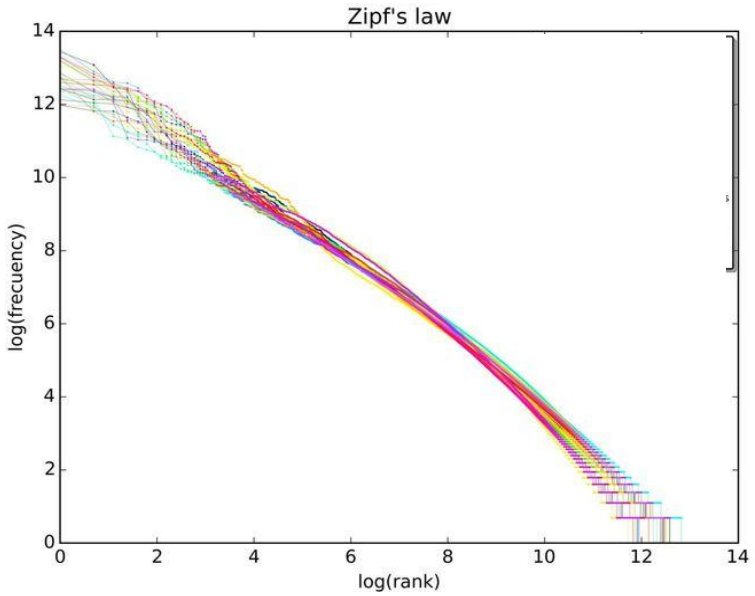


Fig 18 . Ley de Zifp para diferentes idiomas [5]

[5] Wikipedia. Ley de Zipf .Actualizado 9 de Marzo 2019.[En linea]. Consultado (30 de Mayo de 2019). Disponible: https://es.wikipedia.org/wiki/Ley_de_Zipf

Ley de Heaps

Propuesta por Gustav Herdan (1960).
"Describe el número de palabras distintas en un texto en función de la longitud del texto".

$$V = KN^{\beta}$$

V: Número de palabras únicas.

k: Constante que depende del tamaño del texto (10 < k < 100).

β: Constante que depende del tamaño del texto y del idioma.

β es una medida de la riqueza del vocabulario de un texto

A medida que se agrega nuevo texto al documento, habrá rendimientos decrecientes en cuanto al hallazgo de nuevas palabras del vocabulario.
Para el idioma inglés: $0.5 < \beta < 0.6$

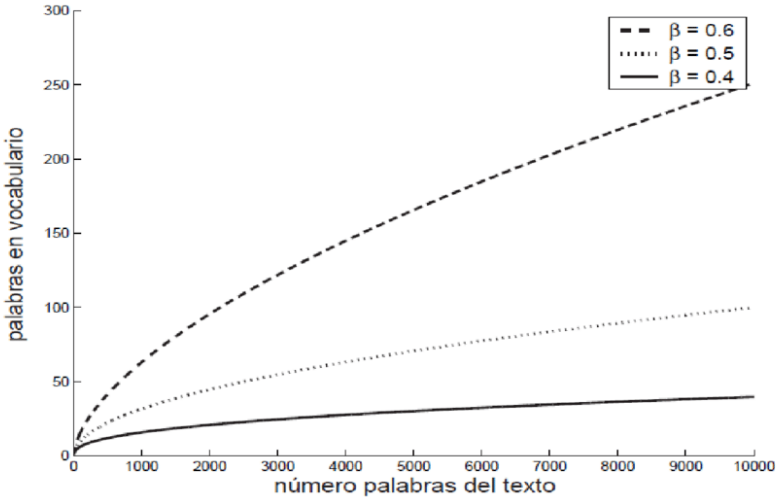


Fig 19 . Ley de Heaps para textos en inglés [6]

[6] Miguel Betanzos Atienza. Aportación a un corpus para el estudio de la variación en árabe. Universidad de Granada(Julio 2015). Consultado (30 de Mayo de 2019).

Nuestra propuesta

Analizar el discurso de usuarios influyentes que utilizan con frecuencia la red social Twitter, por medio de medidas utilizadas en redes complejas. Para ello, se realizan los siguientes paso:

- ❑Tomar los Tweets (sin incluir Retweets) del usuario como discurso y descartar palabras pocos relevantes (proposiciones, artículos, conjunciones), números y símbolos.

- ❑ Construir la red del discurso tomando las palabras como nodos y la relación entre una palabra y la palabra posterior como enlaces.

- ❑ Investigar sobre la posibilidad de caracterizar el discurso mediante la aplicación de las medidas utilizadas en redes complejas.

- ❑ Analizar cómo evoluciona el discurso ante eventos relevantes e investigar sobre la detección de eventos a partir de cambios en las medidas de la red.

- ❑ Aplicar la ley de Zipf y la ley de Heaps al discurso e investigar sobre cambios en los parámetros de las leyes antes y después de un evento relevante.

Lo que publica en redes sociales dice mucho más de Ud. de lo que piensa!

Gracias!