Trabajo Práctico 2
Tirate un qué, tirate un ranking...

Métodos Numéricos

Segundo cuatrimestre - 2014

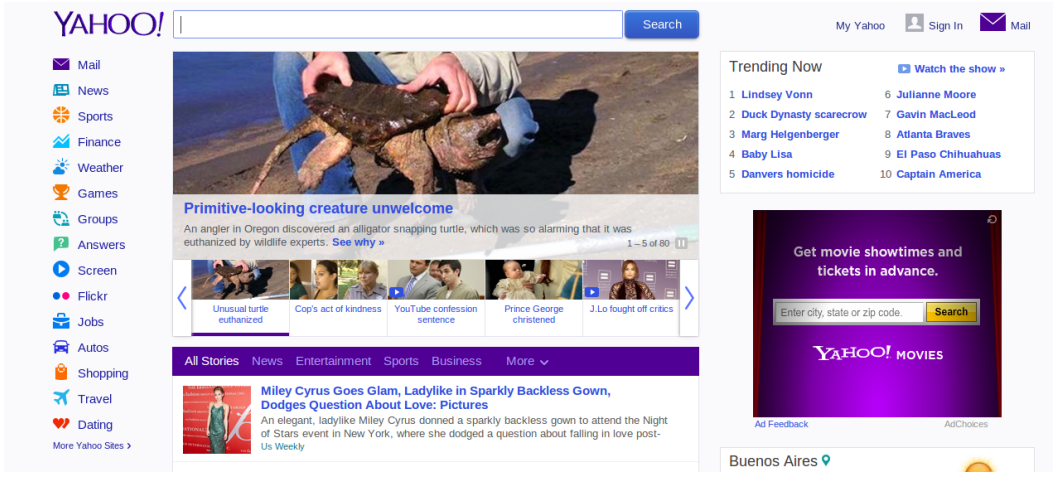
Hasta ahora

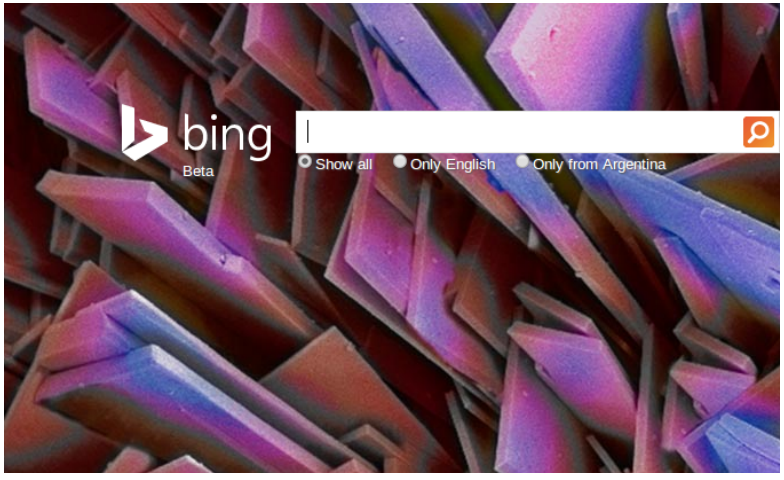
- TP 1: Difusión de calor en placa metálica.
- Taller 1: Eliminación de ruido en imágenes.

Objetivo

Seguir viendo aplicaciones reales de MN.

Motores de búsqueda





Motores de búsqueda

- Explorar la red e identificar todas las páginas con acceso público.
- Almacenar la información obtenida, para realizar búsquedas eficientemente.

Motores de búsqueda

- Explorar la red e identificar todas las páginas con acceso público.
- Almacenar la información obtenida, para realizar búsquedas eficientemente.
- Determinar un orden de las páginas según su importancia, para presentar la información con un orden de relevancia.

Que características son deseables para un *ranking*?

Outline

Contexto TP2

Cadenas de Markov

Algoritmo PageRank

Algoritmo HITS

Enunciado

Cadenas de Markov

Definición

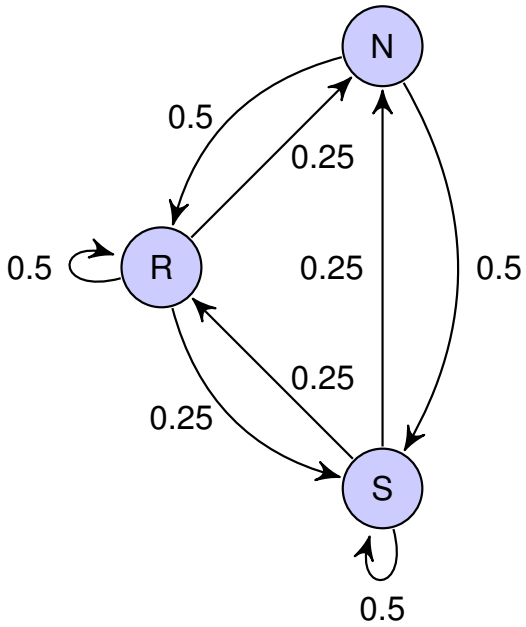
Consideramos un conjunto de estados $S = \{s_1, s_2, \ldots, s_r\}$. El proceso empieza en alguno de estos estados y se mueve de un estado a otro. A cada movimiento se lo denomina *paso*. Si la cadena se encuentra actualmente en el estado s_i, en el siguiente paso se mueve al estado s_j con probabilidad p_{ij}. Esta probabilidad no depende de los estados anteriores a s_i en los que se haya encontrado el proceso.

Cadenas de Markov

Ejemplo: Cambio de clima

- Tres posibilidades: Bueno (N), Lluvioso (R), Nieve (S).
- p_{ij} es la probabilidad de que si en un determinado día estamos en un estado i (i.e., N, R ó S) al día siguiente estemos en el estado j.
- Particularidad: no pueden haber dos días buenos (N) seguidos.

Grafo de transiciones:



Matriz de transiciones:

$$P = \begin{array}{c} \\ \mathrm{R} \\ \mathrm{N} \\ \mathrm{S} \end{array} \begin{array}{c} \begin{array}{ccc} \mathrm{R} & \mathrm{N} & \mathrm{S} \end{array} \\ \begin{pmatrix} 0.5 & 0.25 & 0.25 \\ 0.5 & 0 & 0.5 \\ 0.25 & 0.25 & 0.5 \end{pmatrix} \end{array}$$

- Filas: Estado acutal.
- Columnas: Estado al que podemos movernos.
- Matriz estocástica por filas.

Cadenas de Markov

Mirando más allá de un día

Nuevo problema

Queremos saber cuál es la probabilidad qué, si hoy está lluvioso, nieve dentro de dos días. Llamamos a esta probabilidad $p_{RS}^{(2)}$.

Esto es la unión disjunta de los siguientes eventos:

1. Lluvioso (R) mañana y nieve (S) pasado.
2. Bueno (N) mañana y nieve (S) pasado.
3. Nieve (S) mañana y nieve (S) pasado.

$$P = \begin{array}{c} \\ \mathrm{R} \\ \mathrm{N} \\ \mathrm{S} \end{array} \begin{array}{c} \begin{array}{ccc} \mathrm{R} & \mathrm{N} & \mathrm{S} \end{array} \\ \begin{pmatrix} 0.5 & 0.25 & 0.25 \\ 0.5 & 0 & 0.5 \\ 0.25 & 0.25 & 0.5 \end{pmatrix} \end{array}$$

$$p_{RS}^{(2)} = \underbrace{p_{11}p_{13}}_{1.} + \underbrace{p_{12}p_{23}}_{2.} + \underbrace{p_{13}p_{33}}_{3.}$$

Cadenas de Markov

En general

En el caso anterior,

$$p_{ij}^{(2)} = \sum_{k=1}^{r} p_{ik} p_{kj} = (P^2)_{ij}.$$

Propiedad

El resultado de multiplicar dos matrices estocásticas por filas es una matriz estocástica por filas.

Teorema

Sea P la matriz de transición de una cadena de Markov. El elemento $p_{ij}^{(k)}$ de la matriz P^k es la probabilidad de que la cadena de Markov, empezando en el estado i, se encuentre en el estado j después de k pasos.

Cadenas de Markov

Y si no conocemos el estado actual?

Hasta ahora, supusimos que conocemos el estado actual. Qué pasa si la cadena se encuentra en algún estado con una probabilidad?

Definición: vector de probabilidades

$x \in \mathbb{R}^k$ es un vector (fila) de probabilidades si $x_i \geq 0$ y $\sum_{i=1}^{k} x_i = 1$.

Teorema

Sea P la matriz de transición de una cadena de Markov, y sea u el vector que representa la distribución inicial. Entonces, la probabilidad de que la cadena se encuentre en el estado s_i luego de k pasos es la componente $i-$ésima del vector

$$u^{(k)} = uP^k$$

Cadenas de Markov

Estado estacionario: qué pasa en el largo plazo

Qué sucede con el sistema si consideramos

$$\lim_{n\to\infty} P^n?$$

Definición: Matriz Regular

Una matriz de transiciones P se dice regular si P^k tiene solamente entradas positivas para algún entero k.

Teorema

Sea P una matriz de transiciones regular. Entonces:

- $\lim_{n\to\infty} P^n = W$, donde todas las filas de W son un mismo vector w.
- $wP = w$, y todos los vectores que cumplan $vP = v$ son un múltiplo de w.
- $xP^n \to w$ con $n \to \infty$.

Cadenas de Markov

Estado estacionario: qué pasa en el largo plazo

Si P es una matriz de transiciones regular, entonces:

- 1 es un autovalor de P.
- Hay un único vector de probabilidades que es el autovector asociado al autovalor 1, y es w.
- Se demuestra que los demás autovalores cumplen $1 = \lambda_1 > |\lambda_2| \geq \cdots \geq |\lambda_r|$.

Interpretación

Al vector de probabilidades w se lo denomina *estado estacionario*. La componente w_i representa la proporción de tiempo que la cadena se encuentra, en el largo plazo, en el estado s_i.

En la práctica

Como $wP = w$, entonces $P^t w^t = w^t$. Podemos intentar usar el método de la potencia para calcular w^t.

PageRank

Problema

- Tenemos un conjunto de páginas $Web = \{1, \ldots, n\}$.
- El objetivo es asignar a cada una de ellas un puntaje que determine la importancia relativa de la página respecto de las demás.
- Vamos a trabajar directamente sobre la matriz traspuesta.
- Si definimos una cadena de Markov regular, entonces el estado estacionario nos dará la proporción de tiempo que el navegante aleatorio pasará en cada página.

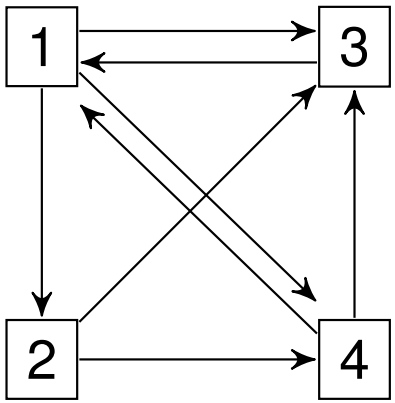
Modelo mediante cadenas de Markov

Consideramos el modelo del *navegante aleatorio*, que comienza en una página cualquiera del conjunto y va navegando a través de sus links.

- Cada página representa un estado de la cadena.
- Podemos pasar de una página j a otra i si hay un link de i a j. Definimos $W \in \{0,1\}^{n \times n}$ como $w_{ij} = 1$ si hay un link de j a i, y $w_{ij} = 0$ en caso contrario.
- $n_j = \sum_{i=1}^{n} w_{ij}$ es el grado de la página j (cantidad de links salientes).
- Definimos $P \in \mathbb{R}^{n \times n}$ como $P_{ij} = 1/n_j$ como la probabilidad de ir de la página j a la i, dado que existe un link de j a i.

PageRank

Ejemplo (Bryan y Leise)



$$n_1 = 3,\ n_2 = 2,\ n_3 = 1,\ n_4 = 2$$

$$P = \begin{bmatrix} 0 & 0 & 1 & \frac{1}{2} \\ \frac{1}{3} & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{3} & \frac{1}{2} & 0 & 0 \end{bmatrix}$$

P es estocástica por columnas.

Pregunta:

Qué pasa si una página i no tiene links salientes (i.e., $n_i = 0$, denominado *dangling node*)?

PageRank

Solución a *dangling nodes*

Definimos:

- $v \in \mathbb{R}^n$, $v_i = 1/n$.
- $d \in \{0,1\}^n$, $d_i = 1$ si $n_i = 0$, $d_i = 0$ en caso contrario.
- $D = vd^t$
- $P_1 = P + D$

Idea

Si estamos en una página sin links salientes, entonces con probabilidad uniforme $1/n$ el navegante pasa a cualquiera de las páginas en *Web*.

Pregunta:

Ahora la matriz es estocástica por columnas. Es regular?

PageRank

Asegurando regularidad

Depende del grafo de conectividad. Sin embargo, podemos extender la idea anterior en general a todas las páginas. A este fenómeno se lo denomina *teletransportación*.

- $\bar{1} = (1, \ldots, 1) \in \mathbb{R}^n$.
- $E = v\bar{1}^t$.
- $P_2 = cP_1 + (1 - c)E$, $c \in (0, 1)$.
- P_2 es estocástica por columnas y $(P_2)_{ij} > 0$, $1 \leq i, j \leq n$.

Finalmente

Tenemos una cadena de Markov que modela el problema y cumple todas las condiciones. Para generar el ranking de las páginas, buscamos un autovector w asociado al autovalor 1 de P_2, tal que $P_2 w = w$, y w sea un vector de probabilidades.

Algoritmo HITS (Kleinberg [5])

Descripción general y objetivos

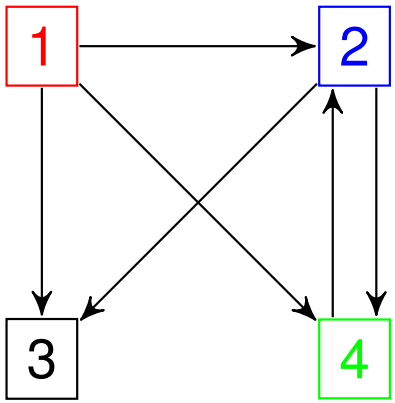
- Una noción de *autoridad* se transfiere de una página a otra mediante los links que las relacionan.
- Dada una búsqueda concreta, retornar un subconjunto acotado de páginas relevantes.
- Se considera que existen páginas que cumplen un rol de *autoridad* sobre un tema específico y se busca modelar la relación entre estas páginas y aquellas que apuntan a varias de estas autoridades, denominadas *hubs*.

Intuición

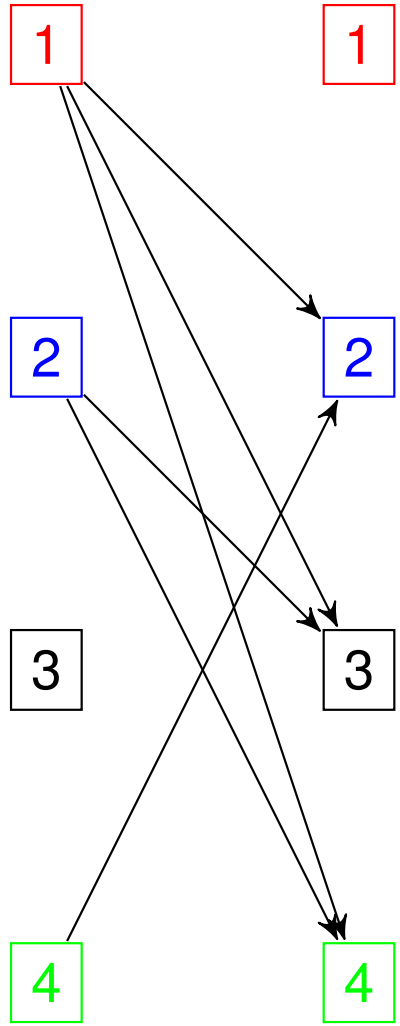
- suele existir una especie de equilibrio en la relación entre hubs y autoridades.
- Un buen *hub* es una página que apunta a muchas autoridades, y una buena *autoridad* es una página que es apuntada por muchos *hubs*.

Algoritmo HITS

Ejemplo



$$A = \begin{bmatrix} 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}$$



Algoritmo HITS

Formalización

Definición

Consideramos la matriz de adyacencia $A \in \{0,1\}^{n \times n}$ donde $a_{ij} = 1$ si existe un link de la página i a la página j. (Notar que $A = W^t$)

Definición

Para cada página $i \in Web$ se considera el *peso de autoridad* x_i y el *peso de hub* y_i.

$$x_j = \sum_{i:i \to j} y_i. \tag{1}$$

$$y_i = \sum_{j:i \to j} x_j. \tag{2}$$

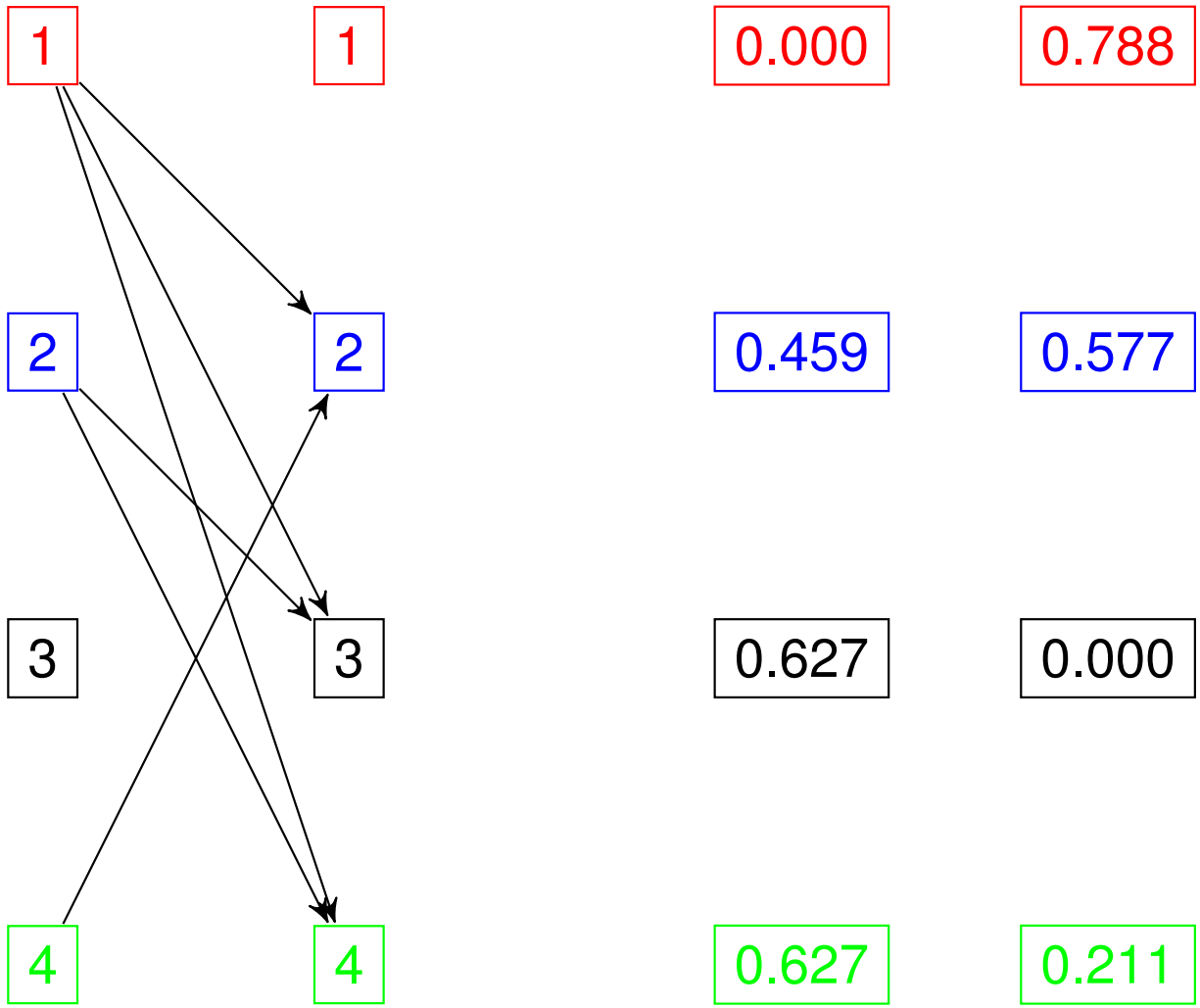
Matricialmente:

$$x = A^t y \tag{3}$$

$$y = Ax, \tag{4}$$

Algoritmo HITS

Ejemplo: a cuál puntaje corresponde cada uno?



TP2

Objetivos generales

- Trabajar sobre una aplicación real, implementando prototipos de algoritmos relevantes utilizados en la práctica.
- Simular un trabajo de investigación:
  - Relevamiento de literatura (qué hay hecho).
  - Desarrollo de algoritmos para el problema.
  - Decisiones de implementación.
  - Experimentación, comparación y análisis de resultados.

TP2

Enunciado: Investigación y revisión de literatura

1. Considerar el trabajo de Kleinberg [5] con los detalles sobre HITS, en particular las secciones 1, 2 y 3.
2. Considerar el trabajo de Bryan y Leise [3] donde se explica la intución y algunos detalles técnicos respecto a PageRank. Además, en Kamvar et al. [4] se propone una mejora del mismo. Si bien esta mejora queda fuera de los alcances del trabajo, en la Sección 1 se presenta una buena formulación del algoritmo. En base a su definición, P_2 no es una matriz esparsa. Sin embargo, en Kamvar et al. [4, Algoritmo 1] se propone una forma alternativa para computar $x^{(k+1)} = P_2 x^{(k)}$. Este resultado puede ser utilizado para mejorar el almacenamiento de los datos.
3. (Opcional) Completar la demostración del Teorema 3.1 de Kleinberg [5], incluyendo el detalle de los puntos que el autor asume como triviales.

TP2

Enunciado: Implementación

1. Se pide utilizar alguna de las siguientes estructuras de datos para la representación de las matrices esparsas:
   - *Dictionary of Keys* (dok),
   - *Compressed Sparse Row* (CSR)
   - *Compressed Sparse Column* (CSC).

   Se deberá incluir una justificación respecto a la elección que consdiere el contexto de aplicación. Una vez definida la estructura a utilizar, se deberá implementar el algoritmo HITS utilizando las ecuaciones (3) y (4). Para el caso de PageRank, se debe implementar el método de la potencia para calcular el autovector principal.

TP2

Enunciado: Experimentación sobre los métodos

1. Estudiar la convergencia de PageRank, analizando la evolución de la norma Manhattan (norma L_1) entre dos iteraciones sucesivas. Comparar los resultados obtenidos para al menos dos instancias de tamaño mediano-grande, variando el valor de c.
   Opcional: Establecer una relación con la proporción entre $\lambda_1 = 1$ y $|\lambda_2|$.
2. Estudiar la convergencia de los vectores de peso x e y para HITS de forma similar al punto anterior.
3. Estudiar el tiempo de cómputo requerido por PageRank y HITS. Analizar como impacta el factor temporal en este sentido.
4. Estudiar cualitativamente los rankings obtenidos por los tres métodos. Analizar los resultados individualmente en una primera etapa, y luego realizar un análisis comparativo entre los tres rankings obtenidos.

TP2

Enunciado: Resolución del problema y competencia

1. Para cada algoritmo, proponer ejemplos de tamaño pequeño que ilustren el comportamiento esperado (puede ser utilizando las instancias provistas por la cátedra o generadas por el grupo).
2. Supongamos que tenemos los pesos (ranking) asignados por uno de los algoritmos estudiados. ¿Cuál sería la estrategia que le sugiere al cliente para mejorar su correspondiente ranking?

TP2

Instancias y datasets

- Conjunto provisto en SNAP [2], con redes de tamaño grande obtenidos a partir de datos reales.
- Instancias propuestas en [1]. Estas instancias son de tamaño mediano, obtenidas también en base a datos reales, y corresponden a redes temáticas obtenidas a partir de una búsqueda particular. Si bien algunas de las URL ya no son válidas, la descripción permite tener algo más de información para realizar un análisis cualitativo.

Material extra (optativo)

Para generar las instancias, se adjunta un codigo Python que, dada una lista de direcciones de páginas web, parsea el código html de cada una de ellas y genera el grafo de conectividad.

Algunas aclaraciones

- Se restringe a links entre las páginas de la lista. El resto de los links son descartados.
- El chequeo para decidir si un link es o no a una página de la lista es básico (ejemplo: www.example.com, ó example.com, ó example.com.ar son considerados links distintos)
- Links que aparezcan dos o más veces son contados una única vez.
- Pueden tomar este código y modificarlo según sus necesidades.
- Si encuentran algún error en el código, por favor contacten a los docentes.

Utilización

El comando

```
python webparser.py weblist.in graph.out
```

toma como entrada la lista de páginas y genera el grafo, con el formato indicado en el enunciado del trabajo, en el archivo graph.out.

TP3

Recomendaciones

- Viernes 26/09: Lectura papers, comprensión métodos, implementación matriz esparsa, implementación HITS.
- Viernes 03/10: Método de la potencia, primeros experimentos (L_1 vs. iteraciones, tiempo de cómputo, etc)
- Viernes 10/10: Set final de experimentos cualitativos, casos de éxito/fracaso para cada algoritmo, respuestas al cliente.
- Sábado 11/10: Entrega TP2.

Trabajo Práctico

Fecha de entrega

- **Formato Electrónico:** Sábado 11 de Octubre de 2014, hasta las 23:59 hs, enviando el trabajo (informe + código) a la dirección **metnum.lab@gmail.com**. El subject del email debe comenzar con el texto **[TP3]** seguido de la lista de apellidos de los integrantes del grupo.
- **Formato físico:** Miercoles 15 de Octubre de 2014, 17 hs., en la clase teórica.

Importante

El horario es estricto. Los correos recibidos después de la hora indicada serán considerados re-entrega.

Bibliografía

http://www.cs.toronto.edu/~tsap/experiments/datasets/.

Stanford large network dataset collection.

http://snap.stanford.edu/data/#web.

Kurt Bryan and Tanya Leise.

The linear algebra behind google.

SIAM Review, 48(3):569–581, 2006.

Sepandar D. Kamvar, Taher H. Haveliwala, Christopher D. Manning, and Gene H. Golub.

Extrapolation methods for accelerating pagerank computations.

In *Proceedings of the 12th international conference on World Wide Web*, WWW '03, pages 261–270, New York, NY, USA, 2003. ACM.

Jon M. Kleinberg.

Authoritative sources in a hyperlinked environment.

J. ACM, 46(5):604–632, September 1999.

Por último...

Pregunta

En qué minuto(s) del video oficial de *Tirate un paso* podemos ver un momento IVAN EHT NIOJ?
Agregarlos en un apéndice del informe, indicando la situación y cantidad de veces.