«Pagerank»

Pagerank es un algoritmo para puntuar la importancia de una página web utilizando un modelo probabilístico del comportamiento de los usuarios en la web. Primero se modelan las páginas web conocidas mediante un grafo dirigido donde una página web esta unida a otra si hay un enlace dentro de esa página.

El usuario navega por las páginas web aleatoriamente, saltando de una página web a otra en cada unidad de tiempo eligiendo aleatoriamente uno de los enlaces. Además, cada enlace tiene una probabilidad de ser elegido.

En el siguiente ejemplo, vemos una representación gráfica de tres páginas web con sus enlaces.

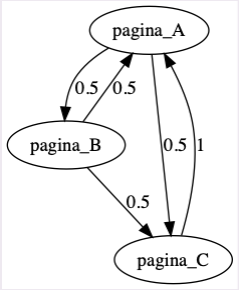

Figura 1:

La distribución de los usuarios esta dada por una distribución de probabilidad que depende del tiempo.

Supongamos que en tiempo cero, la distribución de probabilidad de estar en cada una de las páginas es (0.1, 0.2,0.7). Según el modelo matemático, denotando la probabilidad de estar en la pagina A en el momento t como P(A(t)), tenemos que

$P(A(t)) = P(A(t) \mid B(t-1))P(B(t-1)) + P(A(t) \mid C(t-1))P(C(t-1))$,

$P(B(t)) = P(B(t) \mid A(t-1))P(A(t-1)) + P(B(t) \mid C(t-1))P(C(t-1))$.

$P(C(t)) = P(C(t) \mid A(t-1))P(A(t-1)) + P(C(t) \mid B(t-1))P(B(t-1))$.

El siguiente código demuestra como calcular las diferentes iteraciones para el ejemplo anterior.

```
import numpy as np

M = np.array([[0,0.5,0.5],
              [0.5,0,0.5],
              [1,0,0]
])

v = np.array([0.1,0.1,0.8])
v= np.matmul(v,M)
print(v)
```

```
[0.85 0.05 0.1 ]
```

Modificar el código anterior para calcular que pasará después de que pasen suficientes unidades de tiempo. ¿A qué tienden las probabilidades?

```
import numpy as np

M = np.array([[0,0.5,0.5],
              [0.5,0,0.5],
              [1,0,0]
])
```

```
7
8  v = np.array([0.1,0.2,0.7])
9  print np.matmul(v,M)
```

Estas probabilidades miden la importancia de las páginas web y permiten establecer su importancia.

1) *¿Qué pasa si se cambia la distribución inicial de los usuarios? Haced una prueba con una distribución diferente y comprobad si se tienden a distribuir igual después de varias iteraciones.*

En PageRank los enlaces de las páginas web tienen una distribución uniforme respecto del evento de ser elegidos por el usuario, es decir se suele dar a cada enlace de la página la probabilidad de uno dividido entre el número total de enlaces.

Si una página no tiene enlaces, se supone que los usuarios se quedan indefinidamente en la página.

2) *Calculad la importancia de cada una de estás páginas, independientemente de la distribución de los usuarios.*

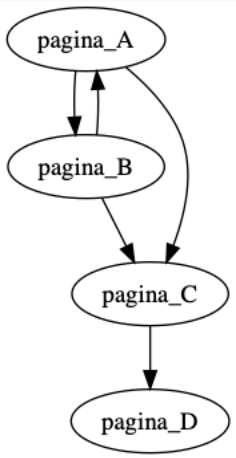

Figura 2:

En los ejemplos anteriores, independientemente de la distribución inicial de los usuarios, la distribución final de los usuarios **cuando el tiempo tiende a infinito** es la misma pero esto no siempre es así.

3) *Poned ejemplos varias páginas web, que dependiendo de la distribución inicial de los usuarios den resultados distintos para la importancia de la página y otras páginas web donde no exista ese límite.*

Nota: Estos ejemplos tienen que estar escritos en lenguaje dot y se debe entregar conjuntamente la imagen en formato png.

Una forma de evitar estos problemas es definir una constante d que representa una probabilidad ir a una página cualquiera. Es decir, con una probabilidad de d el usuario elige una página al azar. En este caso, se puede demostrar que siempre se converge para cualquier de probabilidad inicial.

4) *Se pide dar la importancia de las páginas dadas en el ejercicio 3 para un valor de* $d = 0{,}01$.