DATATHÓN UNAM 2020

Presentación del Problema

M.C. Victor M. Corza
M.C. Sinuhé D. Hernández

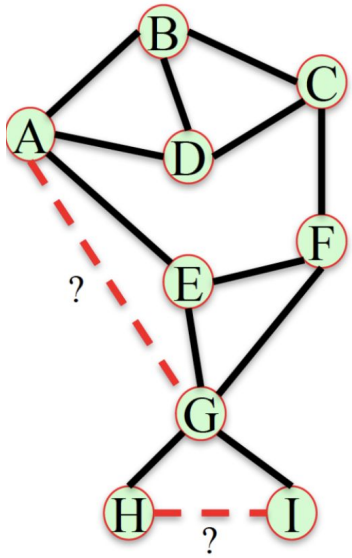
Índice

- Link Prediction Problem
- Coautorías de Artículos Científicos
- Grafo DBLP
- Evaluación

Problemática

Predecir la probabilidad de una futura asociación entre dos nodos, sabiendo que no existe asociación entre ellos en el estado actual del grafo.

Problema de Predicción de Links.



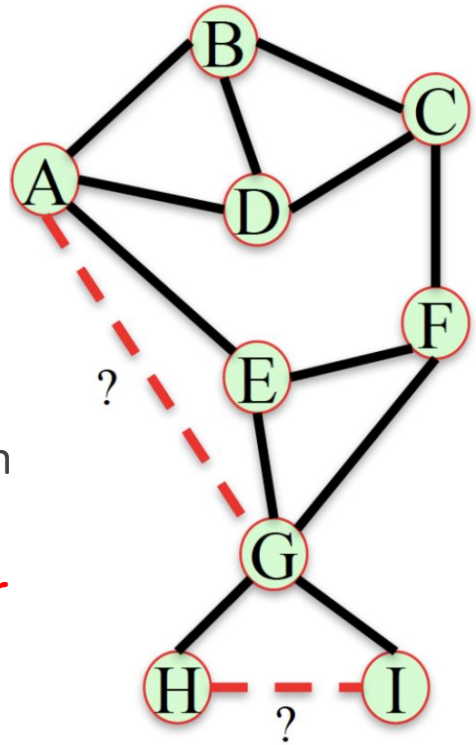
Predicción de Coautorías

Predecir nuevas Coautorías, requiere predecir nuevos enlaces a partir de un conjunto de datos dado.

Qué tengo qué hacer?

Con los datos recibidos deberás generar el grafo que represente las Coautorías existentes en el rango de 1990 al 2000.

En la imagen las coautorías existentes están representadas por enlaces en color **negro** y las coautorías a predecir aparecen con enlaces en **color** rojo con una línea discontinua.



Set de datos

DBLP es una base de datos de publicaciones en temas de Ciencia de la Computación.

Componentes del grafo

Nodes: Represented by Authors.

Edges: Co-authorship between two Authors.

Cómo participar en el Datathon y no morir en el intento

- Tu equipo debe ser organizado.
- Colocar en la carpeta correspondiente a tu equipo las evidencias del trabajo que están realizando: Archivos de entrada, código final y archivos de salida por cada paso.
- En la carpeta correspondiente a tu equipo encontrarás una estructura de carpetas que les permitirá organizar adecuadamente sus avances.

Estructura del Sistema de Archivos

Step_1_XML_processing

Se procesa el archivo XML que se encuentra en la ruta *Step_1_XML_processing/Input_Step_1* para generar un archivo CSV con los campos "**id_article**", "**author**". Deberán extraer únicamente las publicaciones con etiqueta: “**article**” e “**inproceeding**”
Propuesta para el nombre del archivo: "authorships.csv"

Step_2_nodes_catalogue

Deberás generar un catálogo de nodos donde debes descartar autores con menos de **3 autorías.**
Generar un archivo CSV con los **nombres de autores** sin duplicados.

Propuesta para el nombre del archivo: "**nodes_catalogue.csv**"

Step_3_filtering_authorships

En este paso deberás filtrar las autorías para únicamente conservar aquellas cuyos autores aparecen en tu catálogo de autores.
El archivo CSV de salida debe contener los campos "**id_article**", "**author**".
Propuesta para el nombre del archivo:
"filtered_authorships.csv"

Step_4_edges

Con la lista de autorías filtrada deberás generar las aristas correspondientes y eliminar aristas duplicadas.

El archivo CSV de salida debe contener los campos "**source**", "**target**".

Propuesta para el nombre del archivo: "edges.csv"

```
source,target
A. A. Abouelsoud,Mohamed A. Sultan
A. A. Abouelsoud,Mohamed Fahim Hassan
Aapo Hyvärinen,Patrik O. Hoyer
Aapo Hyvärinen,Petteri Pajunen
A. A. Post,Andreas Daffertshofer
A. A. Post,C. (Lieke) E. Peper
A. A. Post,Peter J. Beek
A. Ardeshir Goshtasby,Jayaram K. Udupa
A. Ardeshir Goshtasby,Milan Sonka
Aarne Mämmelä,Veli-Pekka Kaasila
Aaron D. Benally,Reza Zoughi
Aaron D. Wyner,Abraham J. Wyner
Aaron F. Bobick,Andrew D. Wilson
Aaron F. Bobick,Claudio S. Pinhanez
Aaron F. Bobick,James W. Davis
Aaron F. Bobick,John Liu
Aaron F. Bobick,Stephen S. Intille
Aaron F. Bobick,Yuri A. Ivanov
Aaron Marcus,Andries van Dam
Aaron Marcus,Manfred Tscheligi
Aaron Meyerowitz,Elbert A. Walker
Aart Blokhuis,Ákos Seress
Aart Blokhuis,Christine M. O'Keefe
Aart Blokhuis,Henny A. Wilbrink
Aart Blokhuis,Klaus Metsch
Aart Blokhuis,Leo Storme
Aart Blokhuis,Sergei L. Bezrukov
```

Step_5_edges_samplig

En la ruta Step_5_edges_sampling/Input_Step_5 encontrarán un set de aristas etiquetadas con “**P**” para aristas Positivas y “**N**” para aristas Negativas.

El objetivo será predecirlas y verificar el **rendimiento** de tu solución.

```
source,target,prediction
Curt H. Davis,A. A. Soliman,N
Colin Studholme,A. Ardeshir Goshtasby,P
Lawrence H. Staib,A. Ardeshir Goshtasby,N
Ivan Marsic,A. Ardeshir Goshtasby,P
Ertem Tuncel,A. Ardeshir Goshtasby,N
Lee M. Garth,A. C. Cem Say,N
Ion Petre,A. C. Cem Say,N
Gagan L. Choudhury,A. C. Cem Say,N
Donia Scott,A. C. Cem Say,P
Alessio Carullo,A. C. Cem Say,P
Wei Wang 0010,A. Chockalingam,N
Jesús Navarro-Moreno,A. Chockalingam,P
Eldon Y. Li,A. Chockalingam,P
Kaushal Chari,A. Chockalingam,P
Paul L. Rosin,A. David Marshall,N
Zhili Sun,A. David Marshall,N
```

Step_6_final_results

Descargar del servidor el archivo **final_results.csv**, el cual contiene el un conjunto de aristas que deberás comparar con tu propuesta de solución.

Importante!!
Deberán subir el archivo con tu solución a la carpeta **Step_6_final_results/Output_Step_6**

El nombre del archivo será **final_results_xx.csv,** donde xx representa el número de tu equipo en 2 dígitos.

Evaluación

El archivo final_results_xx.csv debe respetar el mismo formato que el archivo “final_results.csv” que recibieron originalmente, esto quiere decir, debe ser un archivo csv, con un campo llamado **source** y otro llamado **target**, sólo deben agregar un campo nuevo a la derecha que se llame “**prediction**” en el que colocarán los resultados de su predicción de cada arista.

Los concursantes **no deben modificar el nombre de ninguno de los autores** en el archivo final_results_xx.csv, si lo hacen, no podrá generarse una evaluación de sus resultados.

```
source,target,prediction
Curt H. Davis,A. A. Soliman,N
Colin Studholme,A. Ardeshir Goshtasby,P
Lawrence H. Staib,A. Ardeshir Goshtasby,N
Ivan Marsic,A. Ardeshir Goshtasby,P
Ertem Tuncel,A. Ardeshir Goshtasby,N
Lee M. Garth,A. C. Cem Say,N
Ion Petre,A. C. Cem Say,N
Gagan L. Choudhury,A. C. Cem Say,N
Donia Scott,A. C. Cem Say,P
Alessio Carullo,A. C. Cem Say,P
Wei Wang 0010,A. Chockalingam,N
Jesús Navarro-Moreno,A. Chockalingam,P
Eldon Y. Li,A. Chockalingam,P
Kaushal Chari,A. Chockalingam,P
Paul L. Rosin,A. David Marshall,N
Zhili Sun,A. David Marshall,N
```

Ganadores

El Jurado evaluará el archivo enviado por equipo para posicionar al ganador del Datathón.

Happy Coding!!