Arturo Fornés Arvayo
A01227071

Práctica 2.1: Datos de *Likes*

Planteamiento

En este documento se buscará ver si los datos de *likes* entre páginas en Facebook, que se obtuvieron utilizando la herramienta *Netvizz*, son tangibles, es decir que cumplan con lo siguiente: validez, precisión, con registros completos, consistencia, uniformidad.

Además se encontrarán dentro de los mismos datos, cuales son los *outliers*, es decir qué datos se encuentran numéricamente distantes del resto.

Procedimiento

Se obtuvieron los datos de los *likes* realizados por la página de Facebook *The National* usando *Netvizz*, pidiendo un grafo de profundidad dos, es decir, realizar de nuevo la búsqueda por cada conexión que *The National* tenía.

Se utilizó la herramienta de visualización de grafos *Gephi* para obtener los documentos en formato *CSV* (Comma-separated values) de la tabla de aristas y de nodos que se pueden observar en la vista de *Data Laboratory* dentro de *Gephi*.

Modelo matemático del filtrado

Los datos se pueden filtrar usando un filtro gaussiano utilizando la siguiente fórmula:

$$\frac{1}{\sqrt{2\pi\sigma^2}}\ e^{-\frac{(\mu-x)^2}{2\sigma^2}}$$

Donde la letra griega *mu* se refiere a la media de una distribución y la letra griega *sigma* hace referencia a la desviación estándar de la misma. El filtro gaussiano sirve para quitar datos que hacen ruido en un set de datos con referencia a la media *mu*; la fórmula mencionada arriba obtiene la probabilidad de *x* en una distribución centrada en la media *mu* y con desviación estándar *sigma*. Con esta fórmula se define un filtrado según los datos con menor probabilidad.

Desarrollo

El procedimiento está documentado en el código incluido en *explore.py*. La métrica a comparar fue la popularidad de una página de Facebook en el círculo de *likes* de la página *The National*. La idea fue encontrar las páginas con popularidad no tan alejada de *The National* dentro del grafo, para esto se conservó la desviación estándar original de la distribución de *likes*, pero se cambió la media por la cantidad de likes que tenía la página *The National*, esto se hizo para centrar los datos en esta página.

Se definió como probabilidad de filtro la probabilidad de que *x* fuera igual a 4, esto sustituído en la fórmula descrita en el modelo matemática de filtrado se utilizó para comparar contra la probabilidad de los *likes* cada vértice.

Resultados

```
python/bases_avanzadas/practica2-1
➜ python explore.py
Mean: 31 StdDev: 7.829941943334298
Outlier limit: 0.00013338311528125846
Deleted 479 items from the data.
Original data contained 1141 items.
Dictionary is now of length: 662

       ==========================================================================================================================================================

[('The Black Keys', 58), ('Rolling Stone', 56), ('Radiohead', 53), ('My Morning Jacket', 52), ('Mumford and Sons', 52), ('iTunes', 44), ('The Rolling Stones', 43), ('The Avet
t Brothers', 43), ('Wilco', 42), ('Relix Magazine', 37), ('Artists Den', 36), ('Adele', 36), ('Willie Nelson', 36), ('PBS', 33), ('Alabama Shakes', 32), ('Bon Iver', 32), ('G
rateful Dead', 32), ('The National', 31), ('Iron & Wine', 30), ('Pearl Jam', 30), ('Jason Isbell', 30), ('Red Hot Chili Peppers', 28), ('Tame Impala', 26), ('St. Vincent', 26
), ('Outside Lands Music Festival', 26), ('Dave Matthews Band', 26), ('Grizzly Bear', 25), ('Bob Marley', 25), ('Old Crow Medicine Show', 25), ('Widespread Panic', 25), ('Dea
th Cab for Cutie', 24), ('Ray LaMontagne', 24), ('Band of Horses', 24), ('Lightning 100', 24), ('Florence + The Machine', 24), ('Tedeschi Trucks Band', 24), ('The Lumineers',
 24), ('David Bowie', 24), ('Stevie Wonder', 23), ('Dr. Dog', 23), ('Lettuce', 22), ('Dawes', 22), ('Edward Sharpe and the Magnetic Zeros', 22), ("Umphrey's McGee", 22), ('Ph
ish', 22), ('Galactic', 22), ('The Head and The Heart', 22), ('Grace Potter', 21), ('Spotify', 21), ('The Civil Wars', 21), ('Tom Petty and the Heartbreakers', 21), ('Warren
Haynes', 21), ('Pretty Lights', 21), ('Punch Brothers', 20), ('Arcade Fire', 20), ('Ben Harper', 19), ('Andrew Bird', 19), ('Fitz & The Tantrums', 19), ('Dumpstaphunk', 19),
('JAY-Z', 19), ('BrooklynVegan', 18), ('SBTRKT', 18), ('Gary Clark Jr', 18), ('STS9', 18), ('Jack White', 18), ("Gov't Mule", 18), ('Big Gigantic', 18), ('Phoenix', 18), ('Fl
aming Lips', 18), ('Local Natives', 17), ('Brandi Carlile', 17), ('Vevo', 17), ('Houndmouth', 17), ('tune-yards', 17), ('Young the Giant', 17), ('The Roots', 17), ('The Shins
', 17), ('Beyoncé', 17), ('Sharon Van Etten', 17), ('The Del McCoury Band', 17), ('The xx', 17), ('Norah Jones', 16), ('Drive-By Truckers', 16), ('The Bluegrass Situation', 1
6), ('M83', 16), ('Moon Taxi', 16), ('Flying Lotus', 16), ('Bassnectar', 16), ('Beach House', 16), ('Animal Collective', 16), ('Paul McCartney', 16), ('KOPECKY', 16), ('G. Lo
ve & Special Sauce', 16), ('AC Entertainment', 16), ('Deer Tick', 16), ('Justin Townes Earle', 16), ('The Decemberists', 16), ('Troy ""Trombone Shorty"" Andrews & Orleans Ave
nue', 16), ('Zac Brown Band', 16), ('Allen Stone', 15), ('Regina Spektor', 15), ('Yonder Mountain String Band', 15), ('Arctic Monkeys', 15), ('Little Dragon', 15), ('Delta Sp
irit', 15), ('Laura Marling', 15), ('Feist', 15), ('Daft Punk', 15), ('Railroad Earth', 15), ('The String Cheese Incident', 15), ('Eminem', 15), ('Michael Franti and Spearhea
d', 15), ('Carolina Chocolate Drops', 15), ('Brassland', 14), ('Sufjan Stevens', 14), ('Cage The Elephant', 14), ('Led Zeppelin', 14), ("Karl Denson's Tiny Universe", 14), ('
Trampled by Turtles', 14), ('Nicki Bluhm & The Gramblers', 14), ('Sam Bush', 14), ('Jam In the Van', 14), ('The Black Lips', 14), ('The War On Drugs', 14), ('Foster The Peopl
e', 14), ('Sublime', 14), ('Eric Clapton', 14), ('Beastie Boys', 14), ('The Apache Relay', 14), ('Vampire Weekend', 13), ('Spoon', 13), ('alt-J', 13), ('Ed Sheeran', 13), ('A
mos Lee', 13), ('Thirteen WNET New York', 13), ('Kurt Vile', 13), ('Macklemore', 13), ('Hurray for the Riff Raff', 13), ('Kendrick Lamar', 13), ('Superfly', 13), ('Shovels an
d Rope', 13), ('James Blake', 13), ('Glen Hansard', 13), ('Passion Pit', 13), ('Of Monsters and Men', 13), ('Reggie Watts', 13), ('Metallica', 13), ('J Roddy Walston and The
Business', 13), ('Ryan Bingham', 13), ('Beats Antique', 13), ('Robyn', 13), ('Primus', 13), ('Chromeo', 13), ('Lotus', 13), ('Weezer', 13), ('Buke and Gase', 12), ('CHVRCHES'
, 12), ('The Antlers', 12), ('The Current', 12), ('Two Door Cinema Club', 12), ('The Killers', 12), ('Mayer Hawthorne', 12), ('Robert Plant', 12), ('World Cafe Live', 12), ('
Sara Watkins', 12), ('Papadosio', 12), ('Jamie xx', 12), ('Béla Fleck', 12), ('Warpaint', 12), ('HAIM', 12), ('Father John Misty', 12), ('Dirty Projectors', 12), ('Preservati
on Hall Jazz Band', 12), ("The Dirty Guv'nahs", 12), ('Ben Howard', 12), ('The Soul Rebels', 12), ('Sarah Jarosz', 12), ('Major Lazer', 12), ('WALK THE MOON', 12), ('Bruce Ho
rnsby', 12), ('Portugal. The Man', 12), ('Mavis Staples', 12), ('Gregg Allman', 12), ('VH1', 11), ('Elvis Costello', 11), ('The Metropolitan Museum of Art, New York', 11), ('
Brett Dennen', 11), ('GRiZ', 11), ('Unknown Mortal Orchestra', 11), ('Shakey Graves', 11), ('Sturgill Simpson', 11), ('Ben Folds', 11), ('Break Science', 11), ('First Aid Kit
 (SE) OFFICIAL', 11), ('John Butler Trio', 11), ('Slightly Stoopid', 11), ('Lucius', 11), ('Purity Ring', 11), ('Paper Diamond', 11), ('NAS', 11), ('Björk', 11), ('Daughter',
 11), ('Cherub', 11), ('The Devil Makes Three', 11), ('Skrillex', 11), ('Nine Inch Nails', 11), ('Modest Mouse', 11), ('GIVERS', 11), ('Best Coast', 11), ('Cold War Kids', 11
), ('Wiz Khalifa', 11), ('Lil Wayne', 11), ('Steve Martin', 11), ('Stereogum', 10), ('Foals', 10), ('ELVIS PRESLEY', 10), ('Calexico', 10), ('KQED', 10), ('Patty Griffin', 10
), ('Ringo Starr', 10), ('The London Souls', 10), ('Snarky Puppy', 10), ('91.9 WFPK Independent Louisville', 10), ('Bahamas', 10), ('Gramatik', 10), ('The Lone Bellow', 10),
('The Wood Brothers', 10), ('Elton John', 10), ('Music on Facebook', 10), ('The Revivalists', 10), ('Kacey Musgraves', 10), ('Jonny Fritz', 10), ('Drew Holcomb and the Neighb
ors', 10), ('Frank Turner', 10), ('The Tallest Man On Earth', 10), ('ASAP Rocky', 10), ('Jim James', 10), ('Wu-Tang Clan', 10), ('NEEDTOBREATHE', 10), ('Phantogram', 10), ('G
reensky Bluegrass', 10), ('Hayes Carll', 10), ('The Low Anthem', 10), ('Band of Skulls', 10), ('Béla Fleck and the Flecktones', 10), ('Big Boi', 10), ('Gogol Bordello', 10),
('Lucero', 10), ('The Disco Biscuits', 10), ('Blitzen Trapper', 10), ('Manchester Orchestra', 10), ('Bryce Dessner', 9), ('Phosphorescent', 9), ('The Secret Sisters', 9), ('I
ngrid Michaelson', 9), ('The Black Crowes', 9), ('Steep Canyon Rangers', 9), ('Caribou', 9), ('St. Paul and The Broken Bones', 9), ('Lake Street Dive', 9), ('Jonathan Wilson'
```

Usando el código en *explore.py* sobre los datos en *depth_2_the_national_Edges.csv*, que contiene las aristas y *depth_2_the_national_Nodes.csv* que contiene los vértices, con la métrica definida en la sección de desarrollo.

Este procedimiento resultó en la eliminación de los vértices que tenían una cantidad de *likes* menores a 4 y un vértice que tenía 123 *likes*. Con esto se obtuvieron, de los 1141 vértices originales, 662 vértices.

Conclusiones

Tras un filtrado gaussiano se cuenta ahora con un dataset que tiene únicamente las entradas que se encuentran cercanas a la página *The National* en cuanto a popularidad, definida por *likes*, dentro de este círculo. Si la página de Facebook, *The National*, fuera a hacer una fiesta y no quisiera ser ni el centro de atención, pero tampoco ser opacada por las páginas más populares podría invitar a este set de páginas.

En cuanto a la tangibilidad de los datos, en este proceso obtuve que cumple o no cumple con los siguientes:

- Son válidos: todos los datos siguen el mismo esquema, esto se puede ver en los archivos *CSV*.
- Son precisos: ya que los datos son unitarios, se tiene un *like* o no se tiene de una página a otra, por lo tanto son precisos.
- Se cuenta con los registros completos: se tienen todos los datos necesarios para el procedimiento, es un subset del universo de datos, pero para el procedimiento que se realizó se puede decir que fueron completos.
- Son consistentes: teniendo 1141 páginas analizadas, la cantidad de *likes* de toda página es entendible dentro del dataset, por lo tanto hay consistencia.
- Tienen uniformidad: las mediciones son unitarias y se usa la misma unidad para medirlo, *se dio like o no*.

Referencias

Gaussian Filtering. (s.f.). De The University of Auckland. Recuperado de: https://www.cs.auckland.ac.nz/courses/compsci373s1c/PatricesLectures/Gaussian%20Filtering_1up.pdf