```
--------------------------------------------------------------------------------
CS230 Assign7 Task 2 Answers | Shirley Lei | 4/25/19

Note: I computed Step 1, Step 2, and Step 3 on both the subset dataset and the full dataset. The
answers for the subset are in the first section, and those for the full dataset are below this.


--------------------------------------------------------------------------------

---------------------------Answers for SUBSET (1000) DATASET----------------------------

--------------------------------------------------------------------------------

Reading in Pantheon dataset:
==> Reading in vertices...
==> Reading in edges...
==> All done!

----------------------------STEP  1: Getting to know your data----------------------------

(1) The highest out-degree person(s) is/are:[Anthony Hopkins]

(2a) The number of people that have out-degree 0 is: 203
(2b) The number of people that have out-degree 1 is: 134

(3) The person with highest in-degree is: Bill Clinton

(4a) The woman with highest in-degree is: Angela Lansbury, with 30 in-degrees.

(4b) The number of men with higher in-degrees than her is: 7

(5) Question of my choice: Total number of connections within each country.

I was interested in discovering the number of connections within each country, since this number
can represent how 'well connected' that particular country is, relative to other countries -- and
I think there is a possible positive correlation between the level of connectedness of people
within a country and the degree to which knowledge, information, and ideas sharing happens within
that country.

OUTPUT STATS on the number of connections within each country:
-->{AZERBAIJAN=1, CAMEROON=1, SRI LANKA=1, Syria=2, JAPAN=6, TOGO=1, China=6, ICELAND=2,
Brazil=3, Bosnia and Herzegovina=1, CHINA=2, Canada=1, BELARUS=1, BRAZIL=10, GERMANY=20, Italy=8,
Ireland=2, CANADA=9, PAKISTAN=1, ITALY=46, ROMANIA=3, Poland=7, Czech Republic=3, NEW ZEALAND=1,
Cape Verde=1, FINLAND=10, ETHIOPIA=1, Chile=1, Montenegro=1, POLAND=9, Denmark=1, Belgium=5,
CHILE=2, Israel=7, UNITED ARAB EMIRATES=1, SLOVAKIA=7, SENEGAL=1, UNITED STATES=138, Trinidad and
Tobago=1, ISRAEL=1, Romania=3, Kyrgyzstan=1, Switzerland=1, Somalia=1, Finland=1, Belarus=1, Isle
of Man=1, Germany=41, Sweden=2, NICARAGUA=1, Nigeria=2, United Kingdom=27, Turkey=14, SWEDEN=6,
AUSTRIA=4, Libya=3, Greece=9, LUXEMBOURG=1, TURKEY=4, VENEZUELA=1, GREECE=3, Lebanon=1, South
Africa=2, Saudi Arabia=2, Kazakhstan=1, Paraguay=1, Uganda=1, Cameroon=1, Ivory Coast=3, United
States=36, SLOVENIA=1, Pakistan=1, Kosovo=1, ARGENTINA=3, LATVIA=1, PORTUGAL=11, UZBEKISTAN=1,
Mexico=2, India=11, Azerbaijan=2, Austria=4, BANGLADESH=2, IRAQ=1, Ethiopia=1, ECUADOR=1,
Panama=1, LIBERIA=1, INDIA=1, CROATIA=4, Estonia=2, Iraq=5, SWITZERLAND=8, NEW CALEDONIA=1,
MEXICO=8, Iran=5, CZECH REPUBLIC=11, Georgia=2, COLOMBIA=1, INDONESIA=1, TAJIKISTAN=1, Tunisia=2,
Unknown=37, Vietnam=2, France=43, Spain=7, UKRAINE=5, Argentina=4, JAMAICA=1, CYPRUS=3, KENYA=1,
SPAIN=27, Norway=1, CUBA=1, UNITED KINGDOM=83, US Virgin Islands=1, Russia=33, NETHERLANDS=9,
FRANCE=36, AUSTRALIA=11, MOZAMBIQUE=1, PHILIPPINES=2, Croatia=1, HUNGARY=2, Cuba=1, Costa Rica=1,
NORWAY=6, Australia=3, BULGARIA=2, Zambia=1, South Korea=7, ALBANIA=1, URUGUAY=4, YEMEN=1, SOUTH
AFRICA=3, Hungary=3, SAUDI ARABIA=2, Bhutan=1, Egypt=8, BERMUDA=1, IRELAND=2, MALAYSIA=1,
Kuwait=1, Palestine=1, Ukraine=5, EGYPT=3, MOROCCO=1, Serbia=3, NAURU=1, KUWAIT=1, HONDURAS=1,
Albania=1, DENMARK=10, Japan=2, BELGIUM=7, Portugal=3, SERBIA=4}

PARTICULAR INSIGHTS on specific countries that I am interested in:
-->United States has 174 connections
-->United Kingdom has 110 connections
-->China has 8 connections
-->Uganda has 1 connection

*Note: this dataset of people connections is most likely biased towards Western countries since
it features Wikipedia, a Western source.
Thus Western countries are expected to have a skewed-higher number of connections relative to the
rest of countries.


-------------------------STEP  2: Finding shortest paths between nodes-------------------

(1a)The shortest path between Madeleine Albright'59 and J.R.R. Tolkien is:
[Madeleine Albright, Bill Clinton, Donald Tusk, Frank-Walter Steinmeier, Margrethe II of Denmark,
J. R. R. Tolkien].This path is 5 steps.

(1b) Pair of my choice is Bill Gates and Stephen King. The shortest path between Bill Gates and
Stephen King is:
[Bill Gates, Dirk Nowitzki, Larry Bird, Bill Murray, Morgan Freeman, Stephen King].This path is 5
steps.
```

(2)The farthest person(s) from Madeleine Albright'59 is/are:
[João Moutinho, Roger Guerreiro, Alex Oxlade-Chamberlain, Ezequiel Garay]

----------STEP 3: (Extra Credit) Bias in the connectivity of Wikipedia pages----------

(1) Fractions of nodes of each gender
Female: 0.132
Male: 0.868

(2) Interpretation of the representation index:

The representation index defines how many times more/less likely a page is to link in-group than the page is to link to any random article.

If representation index = 1, it means a page is just as likely to link to a woman as the fraction of women articles that exist in the dataset (same goes for to men articles).
This would mean that a woman article is just as likely to link to a woman article as to a random article.

If the representation index of a certain group > 1, for example if it equals 2, then a page in that group is *twice* as likely to link in-group than to link to a random article (i.e. MORE likely to link in-group).

If the representation index of a certain group < 1, for example if it equals 0.75, then a page in that group is *3/4* times as likely to link in-group than to link to a random article (i.e. LESS likely to link in-group).

(3a) Average in-group link fractions for each gender:
Female: 0.40286728069641436
Male: 0.7448353752366303

(3b) Representation indices for each gender:
Female: 3.052024853760715
Male: 0.8581052710099428

(4a) Fractions of nodes of each domain
INSTITUTIONS: 0.28
EXPLORATION: 0.01
ARTS: 0.25
SCIENCE & TECHNOLOGY: 0.124
SPORTS: 0.162
BUSINESS & LAW: 0.012
HUMANITIES: 0.129
PUBLIC FIGURE: 0.033

(4b) Average in-group link fractions for each domain
INSTITUTIONS: 0.7342718371184014
EXPLORATION: 0.25
ARTS: 0.6557417976713197
SCIENCE & TECHNOLOGY: 0.6283572021519076
SPORTS: 0.7743165784832452
BUSINESS & LAW: 0.0
HUMANITIES: 0.49015285077852566
PUBLIC FIGURE: 0.1267038517038517

(4c) Representation indices for each domain
INSTITUTIONS: 2.6223994182800046
EXPLORATION: 25.0
ARTS: 2.6229671906852787
SCIENCE & TECHNOLOGY: 5.067396791547642
SPORTS: 4.779731965945958
BUSINESS & LAW: 0.0
HUMANITIES: 3.7996345021591136
PUBLIC FIGURE: 3.839510657692476

(5) Interpretation and commentary of results on SUBSET:

Looking at the subset dataset, first of all, it is saddening but also not surprising women are so underrepresented in Wikipedia articles (only 13% of the articles are of women).
However, it is very striking that women articles have a 3x representation index, meaning that women are 3 times as likely to link to other women than to a random article.
This might mean that women are highly connected with other women, relative to how connected they are with men. This is an interesting insight that powerful women
are well connected with other powerful women -- potentially there is a reverse causation there that powerful women become powerful due to their connections with other powerful women, but obviously there is no enough evidence to draw this conclusion and therefore this is just a provoking thought.

It is also interesting that men articles (around 0.8 representation index) are less likely to link to other men than to a random article. This is a surprising fact, and it may be due to Wikipedia’s efforts to include more links to women within existing men’s pages.

Within domains, the representation index of 'exploration'is an extreme outlier: people within

exploration domain are 25x more likely to link in-group (whereas the average for other groups is around 4x). There is not enough information on why this is, but it would be interesting to learn about the characteristics of this “exploration” field which makes the field so much more tightly connected.

----------Answers for FULL DATASET----------

Reading in Pantheon dataset:
==> Reading in vertices...
==> Reading in edges...
==> All done!

----------STEP 1: Getting to know your data----------

(1) The highest out-degree person(s) is/are:[Pope John Paul II]

(2a) The number of people that have out-degree 0 is: 135
(2b) The number of people that have out-degree 1 is: 178

(3) The person with highest in-degree is: Barack Obama

(4a) The woman with highest in-degree is: Meryl Streep, with 429 in-degrees.
(4b) The number of men with higher in-degrees than her is: 8

(5) Question of my choice: Total number of connections within each country.

I was interested in discovering the number of connections within each country, since this number can represent how 'well connected' that particular country is, relative to other countries -- and I think there is a possible positive correlation between the level of connectedness of people within a country and the degree to which knowledge, information, and ideas sharing happens within that country.

OUTPUT STATS on the number of connections within each country:
-->{KAZAKHSTAN=5, Saudi Arabia=15, South Sudan=2, AFGHANISTAN=16, NAURU=1, ALBANIA=11, SENEGAL=8, Liberia=1, FINLAND=56, Latvia=4, Hungary=30, Kenya=4, GUINEA=3, BAHRAIN=1, ARMENIA=5, SERBIA=36, NORWAY=53, BOLIVIA=2, MONACO=1, Greenland=1, ROMANIA=37, MAURITIUS=1, Gibraltar=1, PUERTO RICO=4, GEORGIA=13, Mozambique=1, TONGA=1, Republic of Macedonia=10, UGANDA=1, The Netherlands=1, KUWAIT=2, Turkey=146, Timor-Leste=1, Yemen=5, CANADA=87, CHINA=15, INDIA=13, LIBERIA=4, Cape Verde=3, United Kingdom=325, GUINEA-BISSAU=3, MOZAMBIQUE=6, IRAQ=6, Ecuador=2, SLOVENIA=16, BERMUDA=1, Cambodia=4, TRINIDAD AND TOBAGO=3, Spain=52, SAUDI ARABIA=21, JERSEY=1, Singapore=7, Palestine=14, Nigeria=24, PARAGUAY=7, TURKMENISTAN=4, SAINT KITTS AND NEVIS=1, BHUTAN=1, CHILE=26, FRANCE=463, Afghanistan=6, Estonia=6, Taiwan=8, PANAMA=3, Mali=2, BRUNEI DARUSSALAM=1, Cyprus=6, BOSNIA AND HERZEGOVINA=21, Albania=5, CONGO=2, Finland=4, ECUADOR=3, SRI LANKA=5, CROATIA=38, Antigua and Barbuda=1, Democratic Republic of Congo=2, Armenia=5, GUATEMALA=3, Germany=485, JAPAN=108, NICARAGUA=5, EL SALVADOR=2, Romania=20, BULGARIA=19, SLOVAKIA=18, UNITED ARAB EMIRATES=4, Kosovo=9, Nepal=4, NEW ZEALAND=15, CAMBODIA=1, Benin=2, SAINT LUCIA=2, Chad=1, Denmark=26, ALGERIA=9, Syria=19, Slovenia=4, MEXICO=42, MALTA=3, Central African Republic=1, MALAYSIA=7, ARGENTINA=53, Poland=65, Egypt=60, Paraguay=7, BURKINA FASO=1, South Africa=15, Madagascar=1, Italy=149, Montenegro=3, GERMANY=262, Malawi=1, Israel=81, OMAN=2, Puerto Rico=3, Somalia=2, MOROCCO=13, HONDURAS=4, SAO TOME AND PRINCIPE=1, Cameroon=3, SWEDEN=81, KENYA=6, DENMARK=75, Russia=374, CUBA=9, TOGO=3, HAITI=7, Venezuela=5, GREENLAND=1, Greece=104, Peru=4, New Zealand=1, Bulgaria=12, Slovakia=8, JORDAN=3, UNITED STATES=1675, Croatia=15, ANDORRA=1, Kiribati=1, Trinidad and Tobago=3, Moldova=8, YEMEN=2, SOMALIA=7, VENEZUELA=10, Uzbekistan=2, US Virgin Islands=2, BRAZIL=115, Malaysia=1, Ivory Coast=14, SOUTH AFRICA=25, COSTA RICA=2, Guinea=2, SPAIN=244, Serbia=23, Norway=6, Isle of Man=4, TAJIKISTAN=2, Monaco=3, Algeria=9, Samoa=1, Sudan=2, CAMEROON=8, LIBYAN ARAB JAMAHIRIYA=4, Azerbaijan=8, ANGOLA=4, Ghana=4, UZBEKISTAN=7, ZAMBIA=3, Uganda=4, Costa Rica=3, Myanmar [Burma]=2, Kuwait=1, DOMINICAN REPUBLIC=1, CENTRAL AFRICAN REPUBLIC=1, Hong Kong=1, Canada=30, Micronesia=1, Dominican Republic=1, Morocco=1, Namibia=2, LATVIA=15, Bangladesh=3, Belarus=16, THAILAND=4, ETHIOPIA=6, NEPAL=1, LESOTHO=1, Ukraine=44, BENIN=2, INDONESIA=8, NETHERLANDS=141, Sri Lanka=1, Bhutan=3, TUNISIA=9, MALI=6, SIERRA LEONE=1, Libya=10, France=400, Zimbabwe=2, MAURITANIA=1, Luxembourg=7, BANGLADESH=5, TURKEY=57, Panama=3, Ireland=47, El Salvador=1, EGYPT=25, Vietnam=12, ITALY=659, BELARUS=18, Iraq=26, Iran=69, Lebanon=2, SWITZERLAND=86, Burkina Faso=1, MADAGASCAR=2, LUXEMBOURG=6, Australia=27, Colombia=5, NEW CALEDONIA=1, Liechtenstein=1, Saint Kitts and Nevis=1, BARBADOS=1, SEYCHELLES=1, BOTSWANA=3, CHAD=1, UKRAINE=75, SURINAME=5, Tonga=2, CAPE VERDE=4, CZECH REPUBLIC=85, QATAR=1, IRELAND=22, Mexico=16, CYPRUS=6, Thailand=3, AUSTRALIA=72, China=93, Ethiopia=5, Austria=42, PORTUGAL=66, India=123, ZIMBABWE=6, PAKISTAN=16, Iceland=6, Jamaica=6, Mongolia=6, Lithuania=7, Equatorial Guinea=1, Netherlands=19, LEBANON=12, PHILIPPINES=12, Sweden=55, São Tomé and Príncipe=1, HUNGARY=53, ERITREA=1, Tunisia=9, Belgium=23, PERU=19, Tanzania=3, Uruguay=1, LITHUANIA=25, SUDAN=3, Chile=3, COLOMBIA=11, Rwanda=1, Unknown=435, GHANA=13, Czech Republic=26, POLAND=108, AUSTRIA=98, Bosnia and Herzegovina=9, Djibouti=1, Switzerland=13, Jordan=3, ICELAND=10, AZERBAIJAN=9, JAMAICA=8, Suriname=1, Kyrgyzstan=4, Guatemala=2, MALAWI=3, Swaziland=1, ISRAEL=3, BURUNDI=1, Laos=1, Japan=32, Senegal=3, Brazil=21, South Korea=37, Kazakhstan=3, GAMBIA=1, MONGOLIA=3, BELGIUM=84, Portugal=21, URUGUAY=21, MALDIVES=3, Pakistan=8, United States=488, Aruba=1, KYRGYZSTAN=3, SWAZILAND=1, VANUATU=1, Guyana=1, GREECE=33, Georgia=12, Argentina=47, Cuba=8, Togo=2, Angola=1, UNITED KINGDOM=814, Zambia=1, Philippines=7, ESTONIA=10, North Korea=6, =33}

PARTICULAR INSIGHTS on specific countries that I am interested in:

-->United States has 2163 connections
-->United Kingdom has 1139 connections
-->China has 108 connections
-->Uganda has 4 connection

*Note: this dataset of people connections is most likely biased towards Western countries since it features Wikipedia, a Western source.
Thus Western countries are expected to have a skewed-higher number of connections relative to the rest of countries.

----------STEP 2: Finding shortest paths between nodes----------

(1a)The shortest path between Madeleine Albright'59 and J.R.R. Tolkien is:
[Madeleine Albright, Helen Hayes, Derek Jacobi, J. R. R. Tolkien]. This path is 3 steps.

(1b) Pair of my choice is Bill Gates and Stephen King. The shortest path between Bill Gates and Stephen King is:
[Bill Gates, Arthur Ashe, Kareem Abdul-Jabbar, Stephen King]. This path is 3 steps.

(2)The farthest person(s) from Madeleine Albright'59 is/are:
[Sunny Leone]

----------STEP 3: (Extra Credit) Bias in the connectivity of Wikipedia pages---------
(1) Fractions of nodes of each gender
Female: 0.13160685875905956
Male: 0.8683931412409405

(2) Interpretation of the representation index:

The representation index defines how many times more/less likely a page is to link in-group than the page is to link to any random article.

If representation index = 1, it means a page is just as likely to link to a woman as the fraction of women articles that exist in the dataset (same goes for to men articles).
This would mean that a woman article is just as likely to link to a woman article as to a random article.

If the representation index of a certain group > 1, for example if it equals 2, then a page in that group is *twice* as likely to link in-group than to link to a random article (i.e. MORE likely to link in-group).

If the representation index of a certain group < 1, for example if it equals 0.75, then a page in that group is *3/4* times as likely to link in-group than to link to a random article (i.e. LESS likely to link in-group).

(3a) Average in-group link fractions for each gender:
Female: 0.4538121146885095
Male: 0.9285380552796435

(3b) Representation indices for each gender:
Female: 3.44824060818388
Male: 1.0692600058456883

(4a) Fractions of nodes of each domain
INSTITUTIONS: 0.3048435566554711
EXPLORATION: 0.008926993105886512
ARTS: 0.2526073890754817
SCIENCE & TECHNOLOGY: 0.12064698603500089
SPORTS: 0.15494078133286193
BUSINESS & LAW: 0.009457309528018385
HUMANITIES: 0.1171115432207884
PUBLIC FIGURE: 0.03146544104649107

(4b) Average in-group link fractions for each domain
INSTITUTIONS: 0.8536930139188333
EXPLORATION: 0.44864886474382915
ARTS: 0.8252903219051543
SCIENCE & TECHNOLOGY: 0.7676736882816151
SPORTS: 0.9513549723064749
BUSINESS & LAW: 0.11704938901447716
HUMANITIES: 0.5871424999220365
PUBLIC FIGURE: 0.174169361997137

(4c) Representation indices for each domain
INSTITUTIONS: 2.800429909967434
EXPLORATION: 50.257556987244385
ARTS: 3.267087019606339
SCIENCE & TECHNOLOGY: 6.3629744389876866
SPORTS: 6.140119884013382
BUSINESS & LAW: 12.376605488876585
HUMANITIES: 5.0135322597116385

PUBLIC FIGURE: 5.535258880998899

(5) Interpretation and commentary of results on FULL dataset:

Looking at the results from the full dataset, it seems that the subset (1000 rows) correctly represented both the fraction of each gender represented in Wikipedia (again, unfortunately only 13% of the articles are of women), and the representation index for women (around 3).

However, the representation index for men is higher, at 1.069. In contrast to the representation index from the subset, this number means that an article about a male is just as likely to link to another male article as to any random article. This is a surprisingly good turn out, since I would have expected that males are more likely to link to males.

Domain wise, we also see that the fractions of nodes in each domain from the subset pretty well represent those of the full dataset. However, the average in-group links for most domains is higher than that of the subset dataset, and this makes sense since more links for any given individual are reflected in the larger dataset.