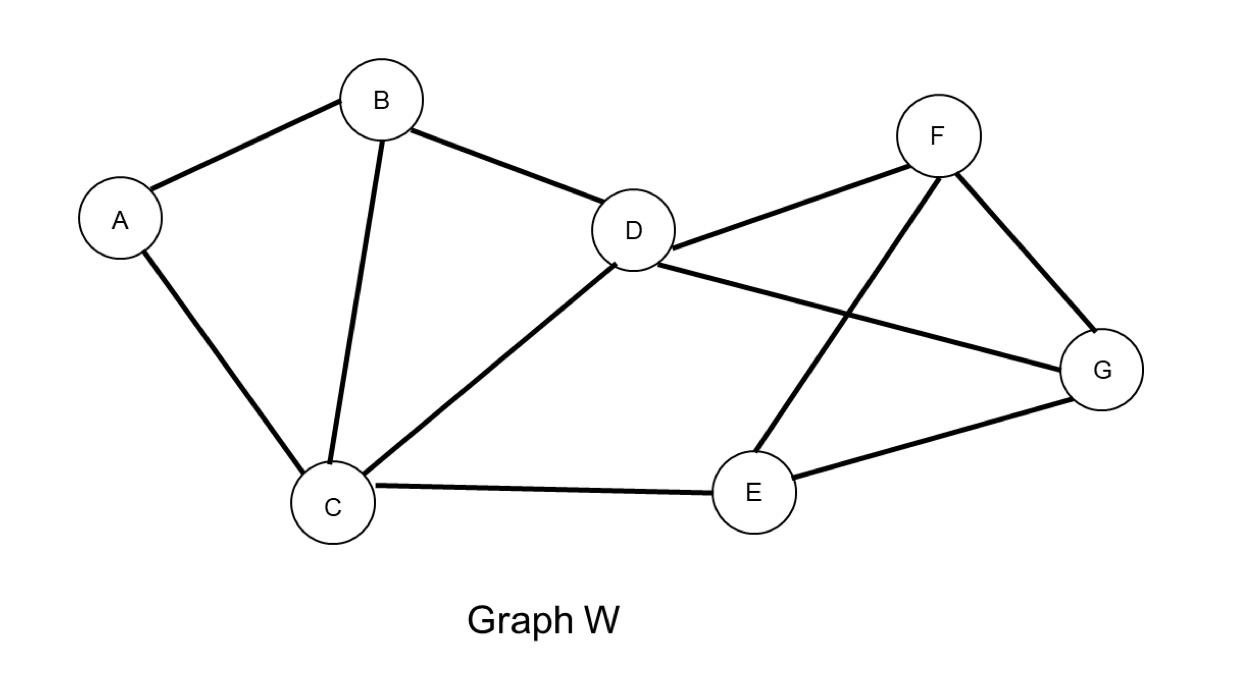
Part 1:

Questions 1 –  8 are based on the social network shown below (Graph W).

You can do the calculations manually or use a software package like Gephi or use R libraries like igraph or sna or tnet.



HW4Q1

True or False: B-D-C-D-B-A is a trail with length 5.

True

False

HW4Q2

What is the maximum number of possible links in Graph W?

7

11

14

21

28

HW4Q3

What is the length of the shortest paths from node A to each of the other nodes?   
A-----[B C D E F G]

1 1 2 2 3 3

0 1 1 2 2 2

0 0 1 2 2 2

1 1 0 1 1 2

2 2 1 0 1 1

HW4Q4

What is the diameter of Graph W?

1

2

3

4

5

HW4Q5

True or False: In a directed graph, the diameter is always defined. Please note that this is a general question and not specific to Graph W.

True

False

HW4Q6

Calculate the node - **degree** centrality of all the nodes.   
In order A B C D E F G  
1 3 3 5 3 3 3

1 2 2 4 2 2 2

2 3 4 4 3 3 3

0.3333333 0.5000000 0.6666667 0.6666667 0.3333333 0.5000000 0.5000000

0.3333333 0.5000000 0.6666667 0.6666667 0.5000000 0.5000000 0.5000000

HW4Q7

Calculate the Node - **closeness** centrality of nodes C, D, and E.   
In order C D E  
0.12500000 0.12500000 0.1000000

0.7500000 0.7500000 0.6666667

0.111 0.143 0.1

0.111 0.143 0.111

Is not defined for undirected graph

HW4Q8

Calculate the Node - **betweeness** centrality of nodes C, D, and E.   
In order C D E  
3 5 3

0.28888889 0.27777778 0.11111111

4.1666667 4.3333333 1.6666667

0.27777778 0.28888889 0.11111111

4.1666667 1.6666667 4.3333333

Part 2:

For the remaining questions, please use the information provide below. We will use the [Review](http://courses.edx.org/asset-v1:GTx+MGT6203x+1T2018+type@asset+block@review.RData)dataset for this part. It should have 1,00,000 rows and 6 columns (variables).

The columns are:

Rev\_ID- The Review ID

productId- Is unique for a movie.

userId – Identifies a user

rating- Is the rating from 1-5 given by the user to the movie

text- Is the User’s review for the given movie

time- Is the time at which the review was ted.

Please remove stop words from the dataset (from column text) and stem the remaining words.

HW4Q9

What is the most common stemmed word in the review? How many times does it occur?

Movie , 12575

Movi, 12575

Film, 13209

Fil, 13209

Movi, 13209

HW4Q10

Plot a 100-word word-cloud. What is the correct size ordering? If it’s not apparent visually, you can check counts to verify.

movi> film >dvd> love> classic> beauti

film> movie>dvd> love> classic> beauti

film> movie > dvd> quot> classic> beauti

movi > film >dvd> special> classic> beauti

movi> dvd > film > flick> classic> life

HW4Q11

Sentiment analysis is:

Understanding the deeper meaning of words expressed in a document

Analyzing the reason behind the usage of certain words in a document

The computational study of opinions, sentiments, and emotions expressed in text

Not very useful in most industries and now considered obsolete.

None of the above

HW4Q12

Perform a sentiment analysis (negative or positive words) and plot a comparison cloud (max words=100). Identify the correct classification.

Positive: thrill, hype, free; Negative: fiction, plot, stun

Positive: thrill, gem, abyss; Negative: fiction, plot, twist

Positive: hell, love, free; Negative: magic, classic, stun

Positive: thrill, gem, free; Negative: fiction, plot, stun

Positive: classic, excel, free; Negative: magic, plot, decent

HW4Q13

The AFINN lexicon assigns words with a score that runs between -5 and 5, with negative scores indicating #negative sentiment and positive scores indicating positive sentiment. Calculate the affin score of all words for each movie, and plot avg score vs avg rating and sum of scores vs avg rating. Is there a linear relationship in the plots?

Yes for avg score vs avg rating, no for sum of scores vs avg rating

Yes for avg score vs avg rating, yes for sum of scores vs avg rating

No for avg score vs avg rating, no for sum of scores vs avg rating

No for avg score vs avg rating, no for sum of scores vs avg rating

Can’t say

HW4Q14

LDA is based on the following guiding principle:

Every topic is a mixture of documents and every topic is a mixture of words

Every document is a mixture of topics and every topic is a mixture of words

Every word is a mixture of topics and every topic is a mixture of documents.

Every topic is a mixture of words and every document is a mixture of words

Each document contains words from several topics

HW4Q15

Topic Modeling using LDA. Perform a 4 topic modeling of the data after converting from tidy form, with one-term-per-document-per-row, to a DocumentTermMatrix. Hint: (Use cast\_dtm). Use seed=1234 for LDA. Obtain the top 5 terms for each of the 4 topics. Which of the following is true?

Movi is the top term for all four topics.

Dvd is top term in two of the topics.

Movi and film are the top two terms in three of the topics.

Movi, film and DVD are the top three terms in three of the topics.

Love is in the top five in all four topics.