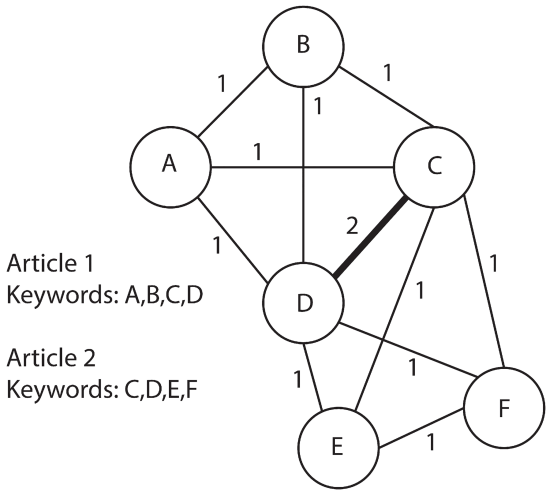
Project 3

The objective of project 3 is to perform keyword network analysis and word frequency analysis

Task 1

1. Download the dataset (https://docs.google.com/spreadsheets/d/1GTwv07i98vL7S-J9eeP8NV1fJVnymm1eJ31RDyt4Mxw/edit?usp=sharing)
2. Write a Python code to extract keyword data from the above file and convert it to a weighted adjacency matrix. See the figure below to understand the process



3. Read the adjacency matrix and convert it into a weighted network
4. Compute node degree and strength
5. Show the top 10 nodes by degree and top 10 nodes by strength
6. Show the top 10 node pairs by weight
7. Plot average strength on y-axis and degree on x-axis

Task 2

The link provides the twitter data of Elon Musk from 2010-2022. For analysis consider the years 2017-2022. Each year has thousands of tweets. Assume each year to be a document (all the tweets in one year will be considered as a document).

1. Compute word frequencies for each year. Exclude the stop words
2. Show top 10 words (for each year) by the highest value of word frequency
3. Plot histogram of word frequencies for each year
4. Use Zipf's law and plot log-log plots of word frequencies and rank for each year
5. Create bigram network graphs for each year

Submission Format

1. Submit all the solutions as a Python notebook
2. Students can also create their own custom functions if necessary
3. This is a group effort
4. Only one member from each group needs to submit the solution
5. Submit the solution by Dec 15