Assignment 5: Text Data Analytics

Dataset: Use your own dataset collected from the assignment 4 (arXiv paper datasets using the search keyword "deep learning" with "Title option)

Create the Rmd file to answer the following questions (use eval = T and echo = T option for all answers)

[Meta data processing]

Q1) Load the exported csv file. How many research papers are there?

Q2) Construct the frequency table with regard to the year of first submission. Then, plot the results with more than two different charts (e.g., bar chart and pie chart).

Q3) What research paper has the largest number of authors? Provide its title and author names.

Q4) What research paper has the longest title? Provide its title.

Q5) Compute the average length of title and abstract, and the average number of authors in each year.

[Text preprocessing & Analysis]

Construct the corpus with the abstracts of the papers. Conduct data preprocessing such as lower case transformation, numbers/punctuation/stopwords removals, and stemming. Beside the preprocessing technique explained in the class, do whatever you think necessary.

Construct the whole term-document matrix for the entire dataset and construct the term-document matrix for each year. For example, if you collect the data for 10 years, you should have one large term-document matrix constructed from the whole dataset (TD-all) and 10 different term-document matrices for each year (TD-year).

Q6) Plot your own graph with the information of non-\sparse entries of each TD-year.

Q7) Provide the top 50 most frequently used words in the TD-all.

Q8) Plot the trends of the words in Q7 with regard to year. What words are more frequently used in recent years? And what words are less frequently used in recent years?

Q9) Construct the word cloud for the TD-all and each TD-year. Interpret the results. There is no single correct answer for this question. Use your own domain knowledge to answer the question.

Q10) Construct the word network for the TD-all and each TD-year. Interpret the results. Do not use the same script to draw the network graph but use your own to make the graph look better.

Q11) Find the association rules for the TD-all and each TD-year. Provide the top 10 rules in terms of the lift value for each result. What insights can you get from the association analysis?