# Project 2

In this project, you need to perform a clustering analysis of the given textual dataset and show your clustering results. Tasks for this project include:

Task1: **Cluster the documents** by their content, **NOT** by document ID. In order to extract some meaningful content from these documents (e.g., person name, location, date, and organization) for clustering, you may consider using the NLTK toolkit (https://www.nltk.org/) or other natural language processing library that you are familiar with. Based on the extracted information, you can use **ANY** clustering algorithms or define your own metrics to perform the clustering.

Task2: **Display the clustering results**. Design and develop some effective visualization(s) to display your clustering results.

You can use **ANY** web-based technology to develop this interactive visualization. D3 are recommended. Since this is a programming project, you should use good software engineering practices. Comment your code, use consistent formatting, use meaningful variable names, etc.

You can work on this project individually or in a team. Each team can have at most two students.

Write a summary document to describe:
1) your method of clustering the given documents,
2) visual encodings (e.g., marks and channels) of your visualization design,
3) is your clustering result meaningful? You need to read the documents and make a judgement. If some documents seem to be mis-clustered, explain why this happens (i.e., the mismatch between your thoughts and your clustering results).

The document should be in MS Word (Times New Roman font, size 12, single line spacing, page limit: 5).