Project 2

In this project, you will add a clustering feature to the tool that you have built in Project 1. All 3 projects of this course are correlated with each other. Tasks for this project include:

Task1: **Cluster the documents** by their content, **NOT** by document ID. In order to extract 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/)](http://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** in your visual analytics tool. In the document ID list, you should organize the IDs of documents in the same cluster near each other (i.e., neighboring elements in the list). If your clustering method only assigns one cluster label for each document, the document ID list should visually reveal *N* blocks (*N* is the total number of clusters), and each block holds document IDs that belong to this cluster. If you clustering methods allow assigning the same document into multiple clusters, you may want to duplicate document IDs or develop an optimized way to organize document IDs in the list. One possible optimization is to minimize the number of clusters whose document IDs are not neighboring each other in the list.

Task3: **Direct manipulation on the *N* blocks**. You should design interactive features on your visually revealed *N* blocks. The interactions should allow users performing the following two actions. Also, these newly added interactions should **NOT** break the ones you have developed in Project 1.

* Ordering the blocks.
* Opening all documents in a cluster at once in the workspace.

You can use **ANY** web-based technology to develop this interactive visualization. D3 and jQuery are recommended.

You are encouraged to work on the project in a group (at most 3 team members). If you want to do those projects by yourself, it would be fine. In the following up projects, we will implement more functions to this visual analytics tool.

Write a summary document to describe your method of clustering documents, your design of showing clusters in the document ID list and interactions on them, and how to run your code. The document should be in MS Word (Times New Roman font, size 12, single line spacing, page limit: 3).

Since this is a programming project, you should use good software engineering practices. Comment your code, use consistent formatting, use meaningful variable names, etc.

The deadline of the first project is **April 2 (Friday) 11:59pm**.