Assignment-2: n-gram TF-IDF and document similarity

Tag: `MDS555-2023-Assignment-2`

Objective
- Learn Term Frequency
- Compute document similarity

Tasks

1) Task 1: Dataset Preparation: Prepare the Nepali news dataset *(hint: you can obtain text from news websites, at least 20 different news of 2/3 different categories)*.
   Host the dataset in the public git repository.
   In your notebook data should be downloaded from git or some other public places.
   No additional step should be done to get the notebook working.
   You can reuse the dataset of Assignment 1 as well.

2) Task 2: Prepare one-gram, bi-gram, tri-gram vocabulary
3) Task 3: Compute TF-IDF vectors for each vocabulary
4) Task 4: Compute document similarity matrix (if your document size = N , this will result in the NxN matrix) for each vocab list.
5) Task 5: Write your interpretation on the result of Task 4.

Deliverable

The deliverable should be in ipython notebook format. Use the Assignment 1 template and do necessary changes.