Birzeit University

Department of Electrical & Computer Engineering

First Semester, 2020/2021

ENCS313 Linux Laboratory

Shell Scripting Project – Text Summarization using Sentence Centrality

---

Extractive summarization works by choosing a subset of sentences from the original document that contains the main contents. Several techniques presented in the literature to handle extractive text summarization. Centrality concept is one of the most used technique. In this approach the document is tokenized into sentences based on (. ! ?) punctuation marks, then the similarity of each pair of sentences is computed. Finally, top scored sentences are selected based on the summary ratio. Formally, let D denote a document consisting of a sequence of sentences $\{s_1, s_2, ..., s_n\}$, and Sim_{ij} is the similarity score for each pair (s_i , s_j). The degree centrality for sentence s_i can be defined as:

$$Centrality(S_i) = \sum_{j\in\{1,\dots,i-1,i+1,\dots,n\}} Sim_{ij}$$

After obtaining the centrality score for each sentence, sentences are sorted in reverse order and the top ranked ones are included in the summary.

Procedure:

- First: the program ask user to enter name of the text file and summary ratio.
- In the second step, the text file will go into a sequence of text preprocessing including:
  - Sentence tokenization based on (. ! ?) punctuation marks.
  - Convert to small letters
  - Remove Stop words from both sentences. Stop Words are words, which do not contain important information. Use the following list of these words:
    [I, a, an, as, at, the, by, in, for, of, on, that]
  - Remove the duplication of words from both sentences. In other words, each word will appear once per sentence.
- Third: compute similarity between each pair of sentences. The similarity calculated as the size of the intersection of words between the two sentences divided by the size of the union of the two sentences:

$$Sim = \frac{(S1 \cap S2)}{(S1 \cup S2)}$$

A value “0” means the two sentences are completely dissimilar, “1” that they are identical, and values between 0 and 1 representing a degree of similarity.

- Fourth: Compute centrality of each sentence.
- Finally: top ranked sentences are selected based on the summary ratio and then written to a file named summary.txt. Note that the selected sentences are sorted based on their centrality score.