CS 410 Team Project

Team: TwoDH

Selected Theme - [Theme 5: Free Topics]

1. What are the names and NetIDs of all your team members? Who is the captain? The captain will have more administrative duties than team members.
   a. Captain → Donghyun Lee, dl20@illinois.edu
   b. Donghyeon Jeong, djeong20@illinois.edu
2. What is your free topic? Please give a detailed description. What is the task? Why is it important or interesting? What is your planned approach? What tools, systems or datasets are involved? What is the expected outcome? How are you going to evaluate your work?
   a. Topic
      i. Our topic of interest is an "*Article Summarizer*."
   b. Description / Expected Outcome
      i. Our final work will be able to summarize one to two pages of articles like blog posts or newspapers into a few sentences by discarding redundant information and preserving important content.
      ii. We expect articles to be summarized significantly when an article contains long sentences (contains numerous words) and redundant information. Also, we assume the summarized article cannot be additionally summarized.
   c. Importance / Interest
      i. Nowadays, people are flooded with information; whereas people do not invest as much time to consume the knowledge. Hence, we would like to enable people to effectively understand and take new information from news and other articles using the summaries without losing key information.
   d. Approach
      i. So far, we have searched for prevailing approaches, key concepts, and differences between existing methods. In result, we have learned and

analyzed *Extractive Text Summarization* and *Abstractive Text Summarization*.

1. ***Extractive Text Summarization*** – attempts to identify significant sentences and then adds them to the summary, which will contain exact sentences from the original text.
2. ***Abstractive Text Summarization*** – attempts to identify important sections, interpret the context and intelligently generate a summary.

ii. Now, we are going to find specific implementation methods for each summarization method and implement either one or both of the methods.

e. Datasets

  i. Cornell Newsroom: https://lil.nlp.cornell.edu/newsroom/index.html

  ii. CNN: https://cs.nyu.edu/~kcho/DMQA/

f. Evaluation

  i. The summarizer will be evaluated by a human user. Multiple people including the team members will read the original article and summarized version and rate the wholeness.

  ii. In case we have spare time and implement both types of summarizer, we will also compare one summarized version to another.

3. Which programming language do you plan to use?

  a. Python 3

4. Please justify that the workload of your topic is at least 20*N hours, N being the total number of students in your team. You may list the main tasks to be completed, and the estimated time cost for each task.

  a. Find text source (1 hr.)

  b. Extractive Text Summarization (15 hr.s)

    i. Parsing

      1. Text Cleaning

    ii. Ranking (Choose most adequate ranking function)

      1. Calculate the TF-IDF
      2. Threshold tuning

  c. Evaluation of *Extractive Text Summarizer* (3 hr.s)

  d. Abstractive Text Summarization (30 hr.s)

    i. This will be performed if time allows