**Projects related to text summarization**

**Project 28: Automatic Text Summarization 1**

This project aims to implement new approaches for automatic text summarization and evaluate their performances on small sample dataset. The Rouge-N metric is the standard in evaluating the

1. First, study the open text summarization available in <https://github.com/jaijuneja/PyTLDR> It uses an extractive based summarization method where the sentences are scored and the highly scored sentences are included in the summarizer. Three scoring techniques have been implemented on this package. One is based on TextRank algorithm (it uses PageRank) and the second is based on Latent Semantic Analysis (You can also check for another PageRank summarizer at <https://github.com/davidadamojr/TextRank>), while the third one uses relevance sentence scoring using cosine similarity, see details on the link. Check that the programs correctly when using either html documents or text documents as input. Demonstrate this finding through an example of your own original document and comment on the summarizer outputted by TextRank, Latent Semantic and Relevance sentence scoring algorithms.
2. Design a simple GUI where the user can input a link or source file of the document to be summarized and output the summarizer using each of the three above methods. You may inspire from the Summarizer toolkit provided in the Moodle section of the course.
3. We would like to evaluate the performance of the three summarizers using a standard evaluation metric. ROUGE-2, ROUGE-3 are commonly employed to evaluate the extent of overlapping between an automatically generated abstract and a set of manually generated summaries. Consider the CNN/Dailymail dataset that you can download from <https://github.com/morningmoni/FAR>. You need a simple python script that allows you to quantify ROUGE-2 and ROUGE-3, you can inspire from numerous implementations available online of automatic summarizers. Your task is to assess the performance of each of three summarizers on CNN/Dailymail dataset using ROUGE-2 and ROUGE-3 metrics, You should . Comment on the performance and limitations of the tested algorithms.
4. We want to extend the above summarization by incorporating coherence of text with respect to named-entity. For this purpose, first use SpaCy named-entity tagger and identify person or organization named-entity. Suggest a simple heuristic that enables whenever a sentence outputted by a given algorithm contains a person or an organization named-entity, then other sentences in the original document that contain the same named-entity, if not outputted by the underlined algorithm, will also be included in the summarizer up to a certain threshold (that you can discuss and tune up). Run the newly designed algorithm on the same CNN/Dailymail dataset, and report the ROUGE-2 and ROUGE-3 performances.
5. Use ChatGPT API (you may inspire from implementation of Summarizer toolkit provided in Moodle section of the course) as a basis to generate 10 different summarizers of the same original text that you used previously. Then use these summaries as golden summarizes to quantify the summarizer generated in 4) using ROUGE-2 and ROUGE-3.
6. Consider the Opinosis dataset available at <https://kavita-ganesan.com/opinosis-opinion-dataset/#.YVw6J5ozY2x>, which contains sentences extracted from user’s reviews on a approximately 51 topics, each having around 100 sentences on average, and includes gold standard summaries. Test the performance of TextRank, Latent Semantic and Relevance sentence scoring on this dataset in terms of Rouge-1 and Rouge-2.
7. Study an implementation of Edmundson summarization system, which uses basic features (word frequency, position, cue words, document structure) available in [Edmundson Heuristic Method for text summarization (opengenus.org)](https://iq.opengenus.org/edmundson-heuristic-method-for-text-summarization/). Test the program in terms of Rouge-1, Rouge-2 score for Opinosis dataset.
8. Now we want to modify the implementation in 6) to account for topic of document in light of the structure of Opinosis. Suggest an approach how to achieve this goal and script that implements your approach. Test the result of the summarizer on Opinosis dataset and CNN/dailymail dataset.
9. Identify relevant literature that allows you to comment on the methodology and results of your implementation.
10. Suggest a GUI where the user can input his own text to be summarized (or a link / pointer to the location of the original document) and output the summary according to each of the aforementioned methodologies.