CS 571: Natural Language Processing **Fall 2023**

Homework 3

Lecturer: Dr. Fei Liu *Due: Monday, 11/6 11:59PM EST*

3.1 Summarization Evaluation

The goal of this assignment is for you to become familiar with a summarization evaluation toolkit: ROUGE. Automatic text summarization aims to generate a concise textual summary from a collection of documents. It has widespread applications in areas such as question answering, search engines, and text analytics.

In this assignment, you are provided with two sets of summaries: `Human_Summaries` contain the goldstandard summaries created by highly qualified human annotators. `System_Summaries` contain a set of folders. Each folder is named by a particular summarization system. In the following paper, you will have the opportunity to learn more about these systems.

`http://www.lrec-conf.org/proceedings/lrec2014/pdf/1093_Paper.pdf`

The input document collection contains 50 topics (indexed from `D30001` to `D31050`). Each topic is associated with 10 news articles collected from major news agencies. An automatic summarization system will generate a summary of 100 words or less (about 5 sentences) for each topic using the 10 news documents as input. The original documents are `not` provided to you for the purpose of this assignment. For evaluation, each system summary is compared against four human summaries on the same topic. The last character of the file name indicates the human annotator's code: (`A` to `H`).

3.1.1 ROUGE Toolkit

Your task is to generate evaluation scores for each text summarization system. Specifically, you will report ROUGE-1, ROUGE-2, and ROUGE-L scores produced by the toolkit. ROUGE was originally developed by Chin-Yew Lin, a researcher at Microsoft Research. The initial implementation was in Perl, but working with its Python wrapper is often more convenient due to its simplicity.

`https://pypi.org/project/pyrouge/`
`https://aclanthology.org/W04-1013.pdf`

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In Proceedings of the ACL Workshop on Text Summarization Branches Out. 2004.

ROUGE is an evaluation tool that measures n-gram overlap between system and human summaries. There are multiple variants, with ROUGE-1, ROUGE-2, and ROUGE-L being the most commonly used metrics. They measure the shared unigrams (single words), bigrams (two consecutive words), and longest common subsequence (LCS) between system and human summaries, respectively. Higher scores indicate better system performance. Traditionally, ROUGE-2 has been believed to correlate well with human judgments in news document summarization evaluation. Currently, it is a common practice to report all three metrics together.

You should be able to download the ROUGE toolkit from the above link and run it without change. Some of the ROUGE options (`-e RELEASE-1.5.5/data -n 4 -m -2 4 -u -c 95 -r 1000 -f A -p 0.5 -t 0 -a -l 100`) and explanations are provided below in case you are interested.

```
"-n 4" compute Rouge-n up to max-ngram
"-e" specify the data folder comes with ROUGE
"-m" use stemming
"-2 -4 -u" use unigram and skip-bigram with distance up to 4 (aka ROUGE-SU4)
"-l 100" use first 100 words of summary for evaluation
"-c 95" confidence level
```

Please: (1) Submit a report titled "report_firstname_lastname.pdf". In the report, include the ROUGE-1, ROUGE-2, and ROUGE-L scores for each of the five summarization systems using the table provided below. Also, describe your experimental setup, covering aspects such as programming language, preprocessing steps, running time, etc. (2) Provide the source code of your implementation in a zipped file.

	ROUGE-1			**ROUGE-2**			**ROUGE-L**		
System	P (%)	R (%)	F (%)	P (%)	R (%)	F (%)	P (%)	R (%)	F (%)
CENTROID	xx.xx	xx.xx	xx.xx	xx.xx	xx.xx	xx.xx	xx.xx	xx.xx	xx.xx
DPP	xx.xx	xx.xx	xx.xx	xx.xx	xx.xx	xx.xx	xx.xx	xx.xx	xx.xx
ICSISUMM	xx.xx	xx.xx	xx.xx	xx.xx	xx.xx	xx.xx	xx.xx	xx.xx	xx.xx
LEXRANK	xx.xx	xx.xx	xx.xx	xx.xx	xx.xx	xx.xx	xx.xx	xx.xx	xx.xx
SUBMODULAR	xx.xx	xx.xx	xx.xx	xx.xx	xx.xx	xx.xx	xx.xx	xx.xx	xx.xx

Table 3.1: Summarization results evaluated by ROUGE (%).