Project 1 Deliverable 3

LING 575: Summarization
Team 2

Team 2

Ben Cote

Mohamed Elkamhawy

Karl Haraldsson

Alyssa Vecht

Josh Warzecha

Project 1 Overview

Multi-document summarization

- News articles
- Various categories and topics

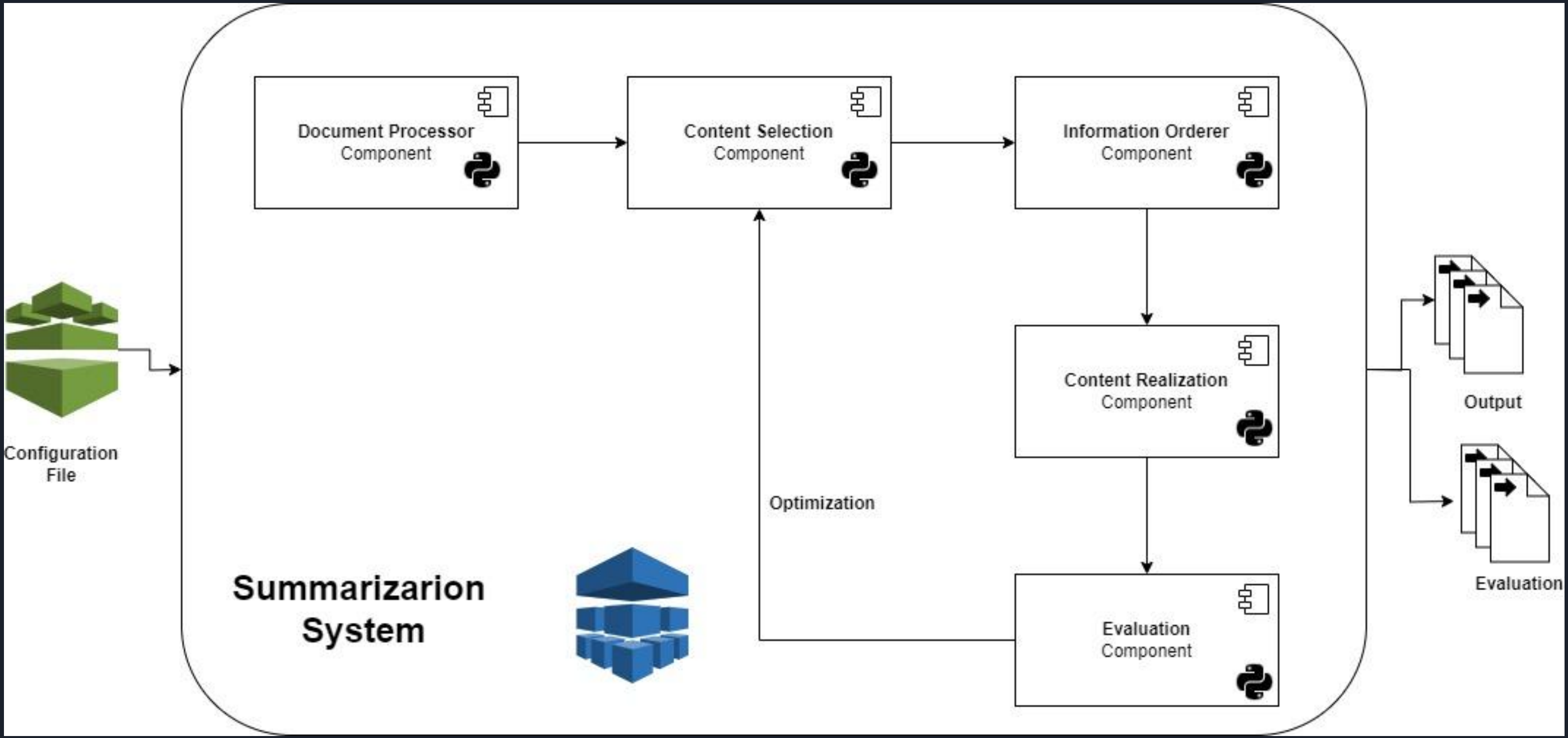
End-to-end system

- Document ingestion
- Content selection
- Information ordering
- Content realization

Evaluation of output summaries

D3: Initial System

System Overview



Content Selection Methods

Goal:

Identify the most salient sentences for inclusion in the summary.

TF-IDF

Identifies important phrases based on their frequency across documents relative to their frequency within a single document, prioritizing content that is both common across and unique to the document set.

TextRank

Utilizes a graph-based ranking model to evaluate the importance of sentences within the text, based on the strength of the relationships between and across other sentences.

Initial Information Ordering Method

Goal:

improve the coherence and readability of summaries by determining an optimal sequence for the selected content.

Implementation:

Inspired by the Traveling Salesperson Problem (TSP) to sequence sentences, using the MASI distance to measure similarity between sentences and applying a two-opt algorithm for optimization.

Sentences are treated as nodes and the goal is to find the shortest path that visits all sentences once, ensuring a logically coherent flow.

The method seeks to arrange sentences in a manner that maximizes topical continuity, enhancing the narrative structure of the summary.

Naive Content Realization Method

Goal:

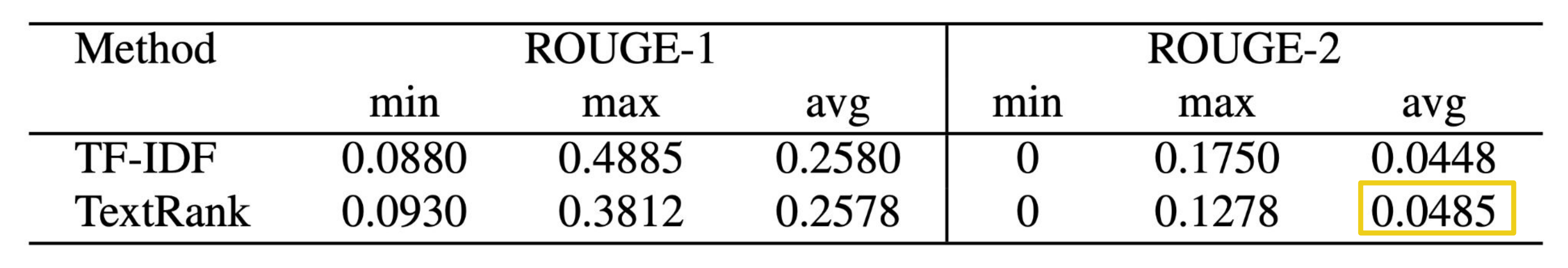
Compose selected and ordered content into a coherent summary that retains the integrity of the underlying sentences.

Implementation:

Utilizes a basic concatenation of sentences, prioritizing the evaluation of content selection and ordering strategies.

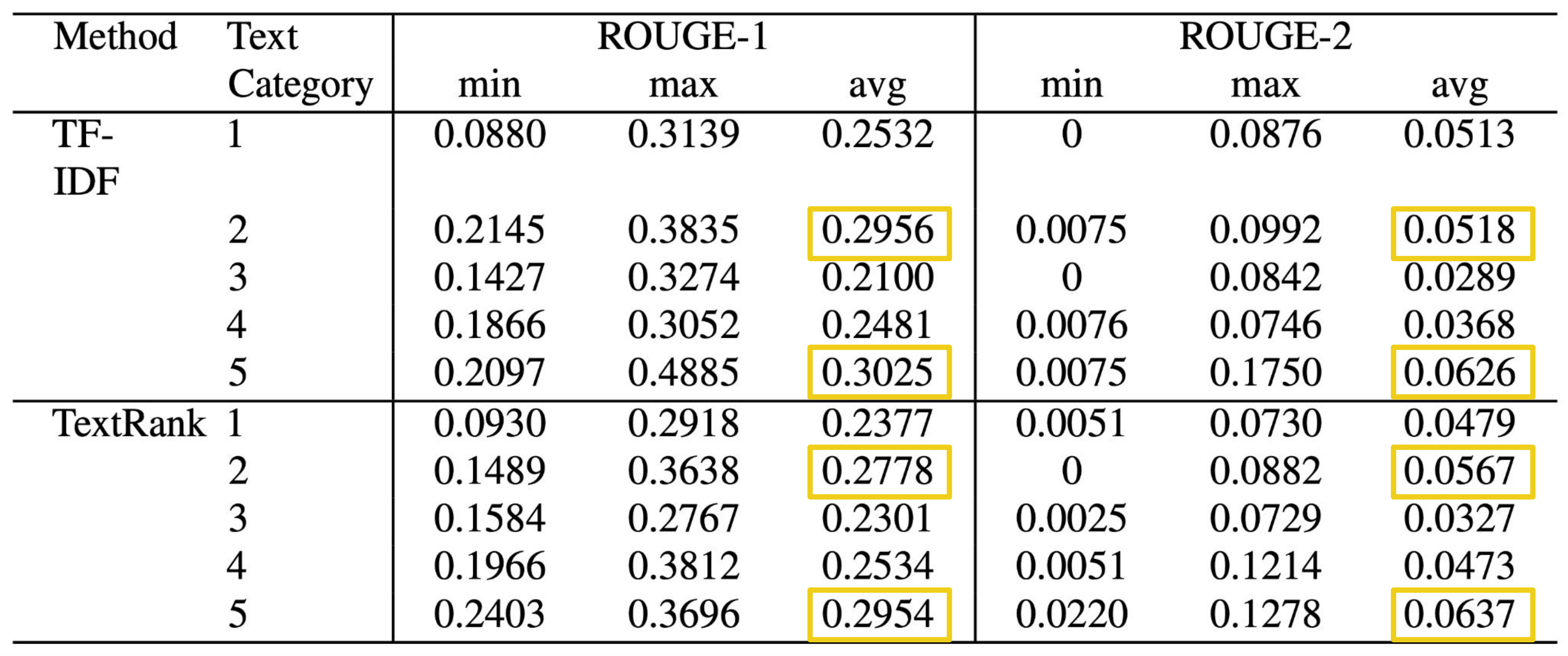
Truncates sentences that would force the summary to exceed 100 word tokens.

Issues & Successes



Method	ROUGE-1			ROUGE-2		
	min	max	avg	min	max	avg
TF-IDF	0.0880	0.4885	0.2580	0	0.1750	0.0448
TextRank	0.0930	0.3812	0.2578	0	0.1278	0.0485

Issues & Successes



Method	Text Category	ROUGE-1			ROUGE-2		
		min	max	avg	min	max	avg
TF-IDF	1	0.0880	0.3139	0.2532	0	0.0876	0.0513
	2	0.2145	0.3835	0.2956	0.0075	0.0992	0.0518
	3	0.1427	0.3274	0.2100	0	0.0842	0.0289
	4	0.1866	0.3052	0.2481	0.0076	0.0746	0.0368
	5	0.2097	0.4885	0.3025	0.0075	0.1750	0.0626
TextRank	1	0.0930	0.2918	0.2377	0.0051	0.0730	0.0479
	2	0.1489	0.3638	0.2778	0	0.0882	0.0567
	3	0.1584	0.2767	0.2301	0.0025	0.0729	0.0327
	4	0.1966	0.3812	0.2534	0.0051	0.1214	0.0473
	5	0.2403	0.3696	0.2954	0.0220	0.1278	0.0637

Next Steps

Initial Plans for Future Versions

Content Selection

- LDA/Topic modeling with Gensim
- Supervised model(s)
- Enhance TF-IDF through:
    - Term Weighting Adjustments
    - Incorporate Semantic Similarity
    - Graph-based Ranking

Information Ordering

- Naive implementations planned for initial system iteration.
- Literature review in progress to guide more refined implementations.

Content Realization

- Sentence compression methods