What is ROUGE-N?

ROUGE-N (Recall-Oriented Understudy for Gisting Evaluation) is a fundamental metric for evaluating how well a machine-generated summary matches human-written reference summaries. It works by comparing the overlap of word sequences (n-grams) between the generated and reference texts.

How It Works:

1- N-gram Matching: Counts how many word sequences of length N (single words, word pairs, etc.) appear in both texts

2- Three Key Scores:

- •Recall: What percentage of reference n-grams appear in the generated text?
- •Precision: What percentage of generated n-grams appear in the reference?
- •F1 Score: The harmonic mean balancing recall and precision

Practical Example:

- Human Reference Summary:

"The quick brown fox jumps over the lazy dog."

- Machine-Generated Summary:

"A fast brown fox leaps over a sleepy dog."

ROUGE-1 (Single Words) Analysis:

- •Matching words: brown, fox, over, dog (4 words)
- •Reference has 8 unique words → Recall = 4/8 = 50%
- •Generated has 8 unique words → Precision = 4/8 = 50%
- •F1 Score = 50%

ROUGE-2 (Word Pairs) Analysis:

- •Matching word pairs: "brown fox", "over the" (2 pairs)

- Reference has 7 bigrams → Recall = 2/7 ≈ 29%
- Generated has 7 bigrams → Precision = 2/7 ≈ 29%
- F1 Score ≈ 29%

Key Insights from This Example:

1. The system captured some core elements (fox, dog, jumping action)
2. Missed exact phrasing ("quick" vs "fast", "jumps" vs "leaps")
3. Shows ROUGE's limitation with synonyms - similar meaning but different words

When to Use ROUGE-N:

- Best for evaluating content coverage in summarization
- Most commonly used: ROUGE-1 (words) and ROUGE-2 (word pairs)
- Standard in research papers and competitions

Important Limitations:

- Doesn't understand meaning or synonyms
- Favors longer summaries (higher recall)
- Doesn't evaluate grammar or coherence

Practical Advice:

- Always report both ROUGE-1 and ROUGE-2
- Include recall, precision and F1 scores
- Combine with other metrics for complete evaluation