Methods for detecting sentence level redundancy:

1. Exact match

Preprocess the sentence: lowercase, trim the sentence to remove space.

Normalize numbers: whenever see a number, turn it into a 0 or a special token.

- Total sentences: 59816

- Repeated sentences: 6792 (11.35%)

- Total words: 905795

- Repeated words: 63276 (6.99%)

2. N-grams match

For each sentence, extract the whole set of 6-10 N-grams. If two sentence share one of these N-grams, it’s considered repetitive.

- Total sentences: 59816

- Repeated sentences: 8964 (14.99%)

- Total words: 905795

- Repeated words: 206212 (22.77%)

Problem: some long sentences are defined as repetitive. The longer the sentence, the more likely it has some shared N-grams with other sentences.

3. Levenshtein Distance

The Levenshtein distance between two strings is the minimum number of single-character edits (insertions, deletions or substitutions) required to change one word into the other.

Levenshtein ratio:

A black text on a white background

AI-generated content may be incorrect.

Threshold for ratio: 0.85

- Total sentences: 59816

- Repeated sentences: 10693 (17.88%)

- Total words: 905795

- Repeated words: 120173 (13.27%)

Report some example sentences and the metric value. Report sentence that are 0.85, 0.9, and 0.95 accurate.

4. Cosine similarity

Use all-mpnet-based-v2 to vectorize the words. Calculate the cosine similarity of each sentence with all sentences in the previous note. If similarity > 0.85 then define as redundant.

- Total sentences: 59816

- Repeated sentences: 12649 (21.15%)

- Total words: 905795

- Repeated words: 149325 (16.49%)

5. Longest common sequence before

Average number of words removed per note: 219.55

Average % of words removed per note: 8.40%

To do:

1. 10 examples, show the original text and cleaned text. Try different threshold.

2. Show examples of n-grams, and other methods.