TLN Part 3 Lab 04 - Text Segmentation

Paletto Andrea, Tuninetti André

September 12, 2023

Contents

1 Introduction

The task of this exercise is implementing a Simple Text Segmentation Algorithm:

- Implement a simple algorithm for text segmentation.
- Use a test input of k paragraphs taken from different topics (e.g., Wikipedia pages).
- Can your system correctly identify the appropriate "cuts" in the text?

2 Methodology & Workflow

The implemented algorithm consists of two primary functions:

- `calculate_cohesion`
- `text_tiling_segmentation`

2.1 Function Overview

2.1.1 `calculate_cohesion`

This function calculates the cohesion between two sentences using TF-IDF vectors and cosine similarity. It returns the mean cosine similarity as a measure of cohesion between sentences.

2.1.2 `text_tiling_segmentation`

1. The function reads an input text file and tokenizes it into sentences using NLTK.
2. It initializes the segmentation process by creating an initial segment containing the first sentence.
3. The algorithm iterates through a specified number of `max_iterations`.
4. For each segment:
    - It compares the current segment with subsequent sentences to calculate the cohesion value using the `calculate_cohesion` function.
    - If the cohesion value is above a specified `cohesion_threshold` or if it's very close to the threshold (within a small range), the sentence is added to the segment.
    - `cohesion_threshold` starts from a very small value, then it gets the value of the last cohesion score. When it drops, it means the topic is changing, so a new segment is created.
    - If the cohesion value is below the threshold, a new segment is created with the current sentence.
    - The process continues, updating and creating segments.
5. After the iteration, the algorithm calculate the coherence score of the last sentence of each segment with the remaining sentences. The sentence with the highest score will be added to the corresponding segment.
6. Duplicate segments are removed.
7. The function returns a list of unique segments, history of cohesion threshold and breakpoints.

2.2 Workflow

1. Read and preprocess the input text produced using chatgpt.
2. Tokenize the text into sentences using NLTK.
3. Initialize the segmentation process with the first sentence.
4. Iterate through a specified number of `max_iterations`.
5. For each iteration, compare and update segments based on cohesion.
6. After iterations, consolidate segments and remove duplicates.
7. plot the cohesion values and breakpoints.

3 Conclusion

3.1 Input:

The input consists of a lengthy text that covers various topics, including the Renaissance, sustainable development, artificial intelligence, space exploration, social media, and climate change.

3.2 Output:

The code has successfully segmented the input text into distinct thematic topics, as indicated by the "Topic" labels and the corresponding text segments. Each segment appears to capture a coherent discussion of a specific subject matter. Here's a summary of the segmented topics:

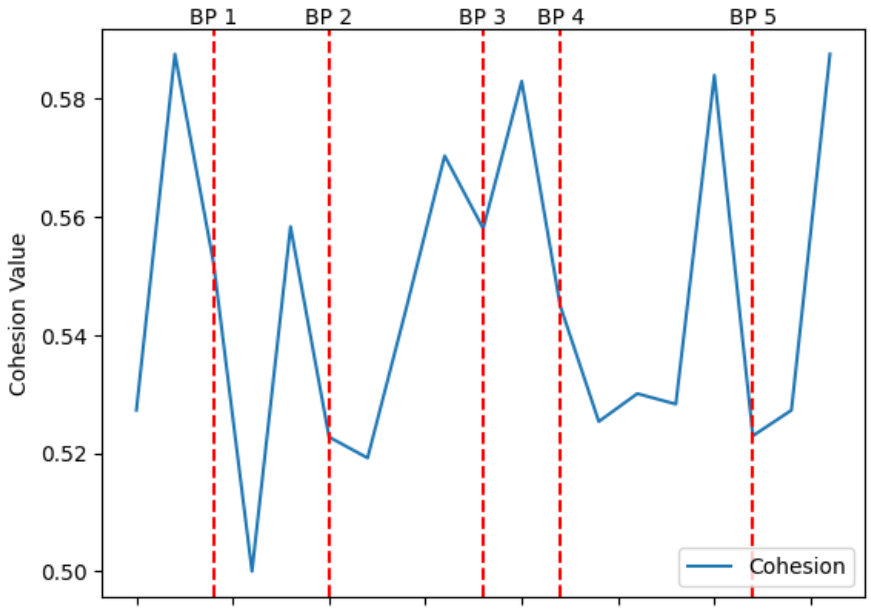

Figure 1: plot with coherence score and break points

1. **Topic 1 - The Renaissance:** This segment discusses the Renaissance period, its cultural significance, and its impact on art, literature, architecture, and science.
2. **Topic 2 - Sustainable Development Goals (SDGs):** This segment focuses on the United Nations' 2030 Agenda for Sustainable Development and outlines the 17 SDGs, addressing global challenges such as poverty, inequality, and climate change.
3. **Topic 3 - Artificial Intelligence (AI):** This segment delves into AI, its various components, and its applications across different industries, highlighting its transformative potential.
4. **Topic 4 - Space Exploration:** This segment covers humanity's endeavors in space exploration, including moon landings, planetary exploration, and the role of space agencies and private companies in pushing technological boundaries.
5. **Topic 5 - Social Media:** This segment explores the impact of social media on communication, information dissemination, and cultural trends, while also acknowledging the challenges and ethical considerations associated with its use.
6. **Topic 6 - Climate Change:** The final segment addresses the topic of climate change, discussing its causes, consequences, and the importance of mitigation and adaptation strategies.

The algorithm successfully segments the dataset into meaningful topics, capturing distinct discussions. The resulting segments are internally cohesive and align with the original content of the dataset.